Beyond CCA: Moment Matching for Multi-View Models

Anastasia Podosinnikova ANASTASIA.PODOSINNIKOVA@INRIA.FR
Francis Bach FRANCIS.BACH@INRIA.FR
Simon Lacoste-Julien FIRSTNAME.LASTNAME@INRIA.FR
INRIA - École normale supérieure, Paris

Abstract

We introduce three novel semi-parametric extensions of probabilistic canonical correlation analysis with identifiability guarantees. We consider moment matching techniques for estimation in these models. For that, by drawing explicit links between the new models and a discrete version of independent component analysis (DICA), we first extend the DICA cumulant tensors to the new discrete version of CCA. By further using a close connection with independent component analysis, we introduce generalized covariance matrices, which can replace the cumulant tensors in the moment matching framework, and, therefore, improve sample complexity and simplify derivations and algorithms significantly. As the tensor power method or orthogonal joint diagonalization are not applicable in the new setting, we use non-orthogonal joint diagonalization techniques for matching the cumulants. We demonstrate performance of the proposed models and estimation techniques on experiments with both synthetic and real datasets.

1. Introduction

Canonical correlation analysis (CCA), originally introduced by Hotelling (1936), is a common statistical tool for the analysis of multi-view data. Examples of such data include, for instance, representation of some text in two languages (e.g., Vinokourov et al., 2002) or images aligned with text data (e.g., Hardoon et al., 2004; Gong et al., 2014). Given two multidimensional variables (or datasets), CCA finds two linear transformations (factor loading matrices) that mutually maximize the correlations between the transformed variables (or datasets). Together with its kernelized version (see, e.g., Shawe-Taylor & Cristianini, 2004; Bach & Jordan, 2003), CCA has a wide range of applications (see, e.g., Hardoon et al. (2004) for an overview).

Bach & Jordan (2005) provide a probabilistic interpretation of CCA: they show that the maximum likelihood estimators of a particular Gaussian graphical model, which we refer to as Gaussian CCA, is equivalent to the classical CCA by Hotelling (1936). The key idea of Gaussian CCA is to allow some of the covariance in the two observed variables to be explained by a linear transformation of common independent sources, while the rest of the covariance of each view is explained by their own (unstructured) noises. Importantly, the dimension of the common sources is often significantly smaller than the dimensions of the observations and, potentially, than the dimensions of the noise. Examples of applications and extensions of Gaussian CCA are the works by Socher & Fei-Fei (2010), for mapping visual and textual features to the same latent space, and Haghighi et al. (2008), for machine translation applications.

Gaussian CCA is subject to some well-known unidentifiability issues, in the same way as the closely related factor analysis model (FA; Bartholomew, 1987; Basilevsky, 1994) and its special case, the probabilistic principal component analysis model (PPCA; Tipping & Bishop, 1999; Roweis, 1998). Indeed, as FA and PPCA are identifiable only up to multiplication by any rotation matrix, Gaussian CCA is only identifiable up to multiplication by any invertible matrix. Although this unidentifiability does not affect the predictive performance of the model, it does affect the factor loading matrices and hence the interpretability of the latent factors. In FA and PPCA, one can enforce additional constraints to recover unique factor loading matrices (see, e.g., Murphy, 2012). A notable identifiable version of FA is independent component analysis (ICA; Jutten, 1987; Jutten & Hérault, 1991; Comon & Jutten, 2010). One of our goals is to introduce identifiable versions of CCA.

The main contributions of this paper are as follows. We first introduce for the first time, to the best of our knowledge, three new formulations of CCA: *discrete, non-Gaussian, and mixed* (see Section 2.1). We then provide *identifiability guarantees* for the new models (see Section 2.2). Then,

Proceedings of the 33^{rd} *International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

in order to use a moment matching framework for estimation, we first derive a *new set of cumulant tensors* for the discrete version of CCA (Section 3.1). We further replace these tensors with their approximations by *generalized covariance matrices* for all three new models (Section 3.2). Finally, as opposed to standard approaches, we use a particular type of *non-orthogonal joint diagonalization algorithms* for extracting the model parameters from the cumulant tensors or their approximations (Section 4).

Models. The new CCA models are adapted to applications where one or both of the data-views are either counts, like in the bag-of-words representation for text, or continuous data, for instance, any continuous representation of images. A key feature of CCA compared to joint PCA is the focus on modeling the common variations of the two views, as opposed to modeling all variations (including joint and marginal ones).

Moment matching. Regarding parameter estimation, we use the method of moments, also known as "spectral methods." It recently regained popularity as an alternative to other estimation methods for graphical models, such as approximate variational inference or MCMC sampling. Estimation of a wide range of models is possible within the moment matching framework: ICA (e.g., Cardoso & Comon, 1996; Comon & Jutten, 2010), mixtures of Gaussians (e.g., Arora & Kannan, 2005; Hsu & Kakade, 2013), latent Dirichlet allocation and topic models (Arora et al., 2012; 2013; Anandkumar et al., 2012; Podosinnikova et al., 2015), supervised topic models (Wang & Zhu, 2014), Indian buffet process inference (Tung & Smola, 2014), stochastic languages (Balle et al., 2014), mixture of hidden Markov models (Sübakan et al., 2014), neural networks (see, e.g., Anandkumar & Sedghi, 2015; Janzamin et al., 2016), and other models (see, e.g., Anandkumar et al., 2014, and references therein).

Moment matching algorithms for estimation in graphical models mostly consist of two main steps: (a) construction of moments or cumulants with a particular diagonal structure and (b) joint diagonalization of the sample estimates of the moments or cumulants to estimate the parameters.

Cumulants and generalized covariance matrices. By using the close connection between ICA and CCA, we first derive in Section 3.1 the cumulant tensors for the discrete version of CCA from the cumulant tensors of a discrete version of ICA (DICA) proposed by Podosinnikova et al. (2015). Extending the ideas from the ICA literature (Yeredor, 2000; Todros & Hero, 2013), we further generalize in Section 3.2 cumulants as the derivatives of the cumulant generating function. This allows us to replace cumulant tensors with "generalized covariance matrices", while preserving the rest of the framework. As a consequence of working with the second-order information only, the derivations and algorithms get significantly simplified and the sample complexity potentially improves.

Non-orthogonal joint diagonalization. When estimating model parameters, both CCA cumulant tensors and generalized covariance matrices for CCA lead to non-symmetric approximate joint diagonalization problems. Therefore, the workhorses of the method of moments in similar context — orthogonal diagonalization algorithms, such as the tensor power method (Anandkumar et al., 2014), and orthogonal joint diagonalization (Bunse-Gerstner et al., 1993; Cardoso & Souloumiac, 1996) — are not applicable. As an alternative, we use a particular type of non-orthogonal Jacobi-like joint diagonalization algorithms (see Section 4). Importantly, the joint diagonalization problem we deal with in this paper is conceptually different from the one considered, e.g., by Kuleshov et al. (2015) (and references therein) and, therefore, the respective algorithms are not applicable here.

2. Multi-view models

2.1. Extensions of Gaussian CCA

Gaussian CCA. Classical CCA (Hotelling, 1936) aims to find projections $D_1 \in \mathbb{R}^{M_1 \times K}$ and $D_2 \in \mathbb{R}^{M_2 \times K}$, of two observation vectors $x_1 \in \mathbb{R}^{M_1}$ and $x_2 \in \mathbb{R}^{M_2}$, each representing a data-view, such that the projected data, $D_1^\top x_1$ and $D_2^\top x_2$, are maximally correlated. Similarly to classical PCA, the solution boils down to solving a generalized SVD problem. The following probabilistic interpretation of CCA is well known (Browne, 1979; Bach & Jordan, 2005; Klami et al., 2013). Given that K sources are i.i.d. standard normal random variables, $\alpha \sim \mathcal{N}(0, I_K)$, the *Gaussian CCA* model is given by

$$\begin{aligned} x_1 \,|\, \alpha,\ \mu_1,\ \Psi_1 &\sim \mathcal{N}(D_1\alpha + \mu_1,\ \Psi_1), \\ x_2 \,|\, \alpha,\ \mu_2,\ \Psi_2 &\sim \mathcal{N}(D_2\alpha + \mu_2,\ \Psi_2), \end{aligned} \tag{1}$$

where the matrices $\Psi_1 \in \mathbb{R}^{M_1 \times M_1}$ and $\Psi_2 \in \mathbb{R}^{M_2 \times M_2}$ are positive semi-definite. Then, the maximum likelihood solution of (1) coincides (up to permutation, scaling, and multiplication by any invertible matrix) with the classical CCA solution. The model (1) is equivalent to

$$\begin{aligned} x_1 &= D_1\alpha + \varepsilon_1, \\ x_2 &= D_2\alpha + \varepsilon_2, \end{aligned} \tag{2}$$

where the noise vectors are normal random variables, i.e. $\varepsilon_1 \sim \mathcal{N}(\mu_1, \Psi_1)$ and $\varepsilon_2 \sim \mathcal{N}(\mu_2, \Psi_2)$, and the following independence assumptions are made:

$$\begin{aligned} &\alpha_1, \ldots, \alpha_K \ \text{ are mutually independent}, \\ &\alpha \perp\!\!\!\perp \varepsilon_1,\ \varepsilon_2 \quad \text{and} \quad \varepsilon_1 \perp\!\!\!\perp \varepsilon_2. \end{aligned} \tag{3}$$

The following three models are our novel semi-parametric extensions of Gaussian CCA (1)–(2).

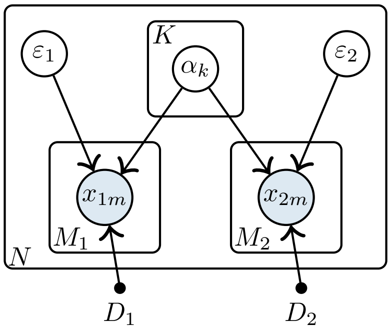

Figure 1. Graphical models for non-Gaussian (4), discrete (5), and mixed (6) CCA.

Multi-view models. The first new model follows by dropping the Gaussianity assumption on α, ε_1, and ε_2. In particular, the *non-Gaussian CCA* model is defined as

$$\begin{aligned} x_1 &= D_1\alpha + \varepsilon_1, \\ x_2 &= D_2\alpha + \varepsilon_2, \end{aligned} \tag{4}$$

where, as opposed to (2), no assumptions are made on the sources α and the noise ε_1 and ε_2 except for the independence assumption (3).

Similarly to Podosinnikova et al. (2015), we further "discretize" non-Gaussian CCA (4) by applying the Poisson distribution to each view (independently on each variable):

$$\begin{aligned} x_1 \,|\, \alpha,\ \varepsilon_1 &\sim \text{Poisson}(D_1\alpha + \varepsilon_1), \\ x_2 \,|\, \alpha,\ \varepsilon_2 &\sim \text{Poisson}(D_2\alpha + \varepsilon_2). \end{aligned} \tag{5}$$

We obtain the (non-Gaussian) *discrete CCA* (DCCA) model, which is adapted to count data (e.g., such as word counts in the bag-of-words model of text). In this case, the sources α, the noise ε_1 and ε_2, and the matrices D_1 and D_2 have non-negative components.

Finally, by combining non-Gaussian and discrete CCA, we also introduce the *mixed CCA* (MCCA) model:

$$\begin{aligned} x_1 &= D_1\alpha + \varepsilon_1, \\ x_2 \,|\, \alpha,\ \varepsilon_2 &\sim \text{Poisson}(D_2\alpha + \varepsilon_2), \end{aligned} \tag{6}$$

which is adapted to a combination of discrete and continuous data (e.g., such as images represented as continuous vectors aligned with text represented as counts). Note that no assumptions are made on distributions of the sources α except for independence (3).

The plate diagram for the models (4)–(6) is presented in Fig. 1. We call D_1 and D_2 *factor loading matrices* (see a comment on this naming convention in Appendix A.2).

Relation between PCA and CCA. The key difference between Gaussian CCA and the closely related FA/PPCA models is that the noise in each view of Gaussian CCA is not assumed to be isotropic unlike for FA/PPCA. In other words, the components of the noise are not assumed to be independent or, equivalently, the noise covariance matrix does not have to be diagonal and may exhibit a strong structure. In this paper, we never assume any diagonal structure of the covariance matrices of the noises of the models (4)–(6). The following example illustrates the mentioned relation. Assuming a linear structure for the noise, (non-) Gaussian CCA (NCCA) takes the form

$$\begin{aligned} x_1 &= D_1\alpha + F_1\beta_1, \\ x_2 &= D_2\alpha + F_2\beta_2, \end{aligned} \tag{7}$$

where $\varepsilon_1 = F_1\beta_1$ with $\beta_1 \in \mathbb{R}^{K_1}$ and $\varepsilon_2 = F_2\beta_2$ with $\beta_2 \in \mathbb{R}^{K_2}$. By stacking the vectors on the top of each other

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},\ D = \begin{pmatrix} D_1 & F_1 & 0 \\ D_2 & 0 & F_2 \end{pmatrix},\ z = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}, \tag{8}$$

we rewrite the model as $x = Dz$. Assuming that the noise sources β_1 and β_2 have mutually independent components, ICA is recovered. If the sources z are further assumed to be Gaussian, $x = Dz$ corresponds to PPCA. However, we do not assume the noise in Gaussian CCA (and in (4)–(6)) to have a very specific low dimensional structure.

Related work. Some extensions of Gaussian CCA were proposed in the literature: exponential family CCA (Virtanen, 2010; Klami et al., 2010) and Bayesian CCA (see, e.g., Klami et al., 2013, and references therein). Although exponential family CCA can also be discretized, it assumes in practice that the prior of the sources is a specific combination of Gaussians. Bayesian CCA models the factor loading matrices and the covariance matrix of Gaussian CCA. Sampling or approximate variational inference are used for estimation and inference in both models. Both models, however, lack our identifiability guarantees and are quite different from the models (4)–(6). Song et al. (2014) consider a multi-view framework to deal with non-parametric mixture components, while our approach is semi-parametric with an explicit linear structure (our loading matrices) and makes the explicit link with CCA. See also Ge & Zou (2016) for a related approach.

2.2. Identifiability

In this section, the identifiability of the factor loading matrices D_1 and D_2 is discussed. In general, for the type of models considered, the unidentifiability to permutation and scaling cannot be avoided. In practice, this unidentifiability is however easy to handle and, in the following, we only consider identifiability up to permutation and scaling.

ICA can be seen as an identifiable analog of FA/PPCA. Indeed, it is known that the mixing matrix D of ICA is identifiable if at most one source is Gaussian (Comon, 1994). The factor loading matrix of FA/PPCA is unidentifiable since it is defined only up to multiplication by any orthogonal rotation matrix.

Similarly, the factor loading matrices of Gaussian CCA (1), which can be seen as a multi-view extension of PPCA, are identifiable only up to multiplication by any invertible matrix (Bach & Jordan, 2005). We show the identifiability results for the new models (4)–(6): the factor loading matrices of these models are identifiable if at most one source is Gaussian (see Appendix B for a proof).

Theorem 1. *Assume that matrices $D_1 \in \mathbb{R}^{M_1 \times K}$ and $D_2 \in \mathbb{R}^{M_2 \times K}$, where $K \leq \min(M_1,\ M_2)$, have full rank. If the covariance matrices $\operatorname{cov}(x_1)$ and $\operatorname{cov}(x_2)$ exist and if at most one source α_k, for $k = 1, \dots, K$, is Gaussian and none of the sources are deterministic, then the models (4)–(6) are identifiable (up to scaling and joint permutation).*

Importantly, the permutation unidentifiability does not destroy the alignment in the factor loading matrices, that is, for some permutation matrix P, if $D_1 P$ is the factor loading matrix of the first view, than $D_2 P$ must be the factor loading matrix of the second view. This property is important for the interpretability of the factor loading matrices and, in particular, is used in our experiments in Section 5.

3. The cumulants and generalized covariances

In this section, we first derive the cumulant tensors for the discrete CCA model (Section 3.1) and then generalized covariance matrices (Section 3.2) for the models (4)–(6). We show that both cumulants and generalized covariances have a special diagonal form and, therefore, can be efficiently used within the moment matching framework (Section 4).

3.1. From discrete ICA to discrete CCA

In this section, we derive the DCCA cumulants as an extension of the cumulants of discrete independent component analysis (DICA; Podosinnikova et al., 2015).

Discrete ICA. Podosinnikova et al. (2015) consider the discrete ICA model (9), where $x \in \mathbb{R}^M$ has conditionally independent Poisson components with mean $D\alpha$ and $\alpha \in \mathbb{R}^K$ has independent non-negative components:

$$x \,|\, \alpha \sim \operatorname{Poisson}(D\alpha). \tag{9}$$

For estimating the factor loading matrix D, Podosinnikova et al. (2015) propose an algorithm based on the moment matching method with the cumulants of the DICA model. In particular, they define the DICA S-covariance matrix and T-cumulant tensor as

$$\begin{aligned} S &:= \operatorname{cov}(x) - \operatorname{diag}\left[\mathbb{E}x\right], \\ [T]_{m_1 m_2 m_3} &:= \operatorname{cum}(x)_{m_1 m_2 m_3} + [\tau]_{m_1 m_2 m_3}, \end{aligned} \tag{10}$$

where indices m_1, m_2, and m_3 take the values in $1, \dots, M$, and $[\tau]_{m_1 m_2 m_3} = 2\delta_{m_1 m_2 m_3}\mathbb{E}x_{m_1} - \delta_{m_2 m_3}\operatorname{cov}(x)_{m_1 m_2} - \delta_{m_1 m_3}\operatorname{cov}(x)_{m_1 m_2} - \delta_{m_1 m_2}\operatorname{cov}(x)_{m_1 m_3}$ with δ being the Kronecker delta. For completeness, we outline the derivation by Podosinnikova et al. (2015) below. Let $y := D\alpha$. By the law of total expectation $\mathbb{E}(x) = \mathbb{E}(x|y) = \mathbb{E}(y)$ and by the law of total covariance

$$\begin{aligned} \operatorname{cov}(x) &= \mathbb{E}[\operatorname{cov}(x|y)] + \operatorname{cov}[\mathbb{E}(x|y),\ \mathbb{E}(x|y)] \\ &= \operatorname{diag}[\mathbb{E}(y)] + \operatorname{cov}(y), \end{aligned}$$

since all the cumulants of a Poisson random variable with parameter y are equal to y. Therefore, $S = \operatorname{cov}(y)$. Similarly, by the law of total cumulance $T = \operatorname{cum}(y)$. Then, by the multilinearity property for cumulants, one obtains

$$\begin{aligned} S &= D \operatorname{cov}(\alpha) D^\top, \\ T &= \operatorname{cum}(\alpha) \times_1 D^\top \times_2 D^\top \times_3 D^\top, \end{aligned} \tag{11}$$

where $\times_i$ denotes the i-mode tensor-matrix product (see, e.g., Kolda & Bader, 2009). Since the covariance $\operatorname{cov}(\alpha)$ and cumulant $\operatorname{cum}(\alpha)$ of the independent sources are diagonal, (11) is called the *diagonal form*. This diagonal form is further used for estimation of D (see Section 4).

Noisy discrete ICA. The following noisy version (12) of the DICA model reveals the connection between DICA and DCCA. Noisy discrete ICA is obtained by adding non-negative noise ε, such that $\alpha \perp\!\!\!\perp \varepsilon$, to discrete ICA (9):

$$x \,|\, \alpha,\ \varepsilon \sim \operatorname{Poisson}\left(D\alpha + \varepsilon\right). \tag{12}$$

Let $y := D\alpha + \varepsilon$ and S and T are defined as in (10). Then a simple extension of the derivations from above gives $S = \operatorname{cov}(y)$ and $T = \operatorname{cum}(y)$. Since the covariance matrix (cumulant tensor) of the sum of two independent multivariate random variables, $D\alpha$ and ε, is equal to the sum of the covariance matrices (cumulant tensors) of these variables, the "perturbed" version of the diagonal form (11) follows

$$\begin{aligned} S &= D\operatorname{cov}(\alpha) D^\top + \operatorname{cov}(\varepsilon), \\ T &= \operatorname{cum}(\alpha) \times_1 D^\top \times_2 D^\top \times_3 D^\top + \operatorname{cum}(\varepsilon). \end{aligned} \tag{13}$$

DCCA cumulants. By analogy with (8), stacking the observations $x = [x_1;\ x_2]$, the factor loading matrices $D = [D_1;\ D_2]$, and the noise vectors $\varepsilon = [\varepsilon_1;\ \varepsilon_2]$ of discrete CCA (5) gives a noisy version of discrete ICA with a particular form of the covariance matrix of the noise:

$$\operatorname{cov}(\varepsilon) = \begin{pmatrix} \operatorname{cov}(\varepsilon_1) & 0 \\ 0 & \operatorname{cov}(\varepsilon_2) \end{pmatrix}, \tag{14}$$

which is due to the independence $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$. Similarly, the cumulant $\operatorname{cum}(\varepsilon)$ of the noise has only two diagonal blocks which are non-zero. Therefore, considering only those parts of the S-covariance matrix and T-cumulant tensor of noisy DICA that correspond to zero blocks of the covariance $\operatorname{cov}(\varepsilon)$ and cumulant $\operatorname{cum}(\varepsilon)$, gives immediately a matrix and tensor with a diagonal structure similar to the one in (11). Those blocks are the cross-covariance and cross-cumulants of x_1 and x_2.

We define the *S-covariance matrix of discrete CCA*[1] as the cross-covariance matrix of x_1 and x_2:

$$S_{12} := \operatorname{cov}(x_1, x_2). \tag{15}$$

[1] Note that $S_{21} := \operatorname{cov}(x_2, x_1)$ is just the transpose of S_{12}.

From (13) and (14), the matrix S_{12} has the following diagonal form

$$S_{12} = D_1 \mathrm{cov}(\alpha) D_2^\top. \tag{16}$$

Similarly, we define the *T-cumulant tensors of discrete CCA* ($T_{121} \in \mathbb{R}^{M_1 \times M_2 \times M_1}$ and $T_{122} \in \mathbb{R}^{M_1 \times M_2 \times M_2}$) through the cross-cumulants of x_1 and x_2, for $j = 1, 2$:

$$\begin{aligned}[T_{12j}]_{m_1 m_2 \tilde{m}_j} := & [\mathrm{cum}(x_1, x_2, x_j)]_{m_1 m_2 \tilde{m}_j} \\ & - \delta_{m_j \tilde{m}_j} [\mathrm{cov}(x_1, x_2)]_{m_1 m_2},\end{aligned} \tag{17}$$

where the indices m_1, m_2, and $\tilde{m}_j$ take the values $m_1 \in 1, \ldots, M_1$, $m_2 \in 1, \ldots, M_2$, and $\tilde{m}_j \in 1, \ldots, M_j$. From (11) and the mentioned block structure (14) of $\mathrm{cov}(\varepsilon)$, the DCCA T-cumulants have the diagonal form:

$$\begin{aligned} T_{121} &= \mathrm{cum}(\alpha) \times_1 D_1^\top \times_2 D_2^\top \times_3 D_1^\top, \\ T_{122} &= \mathrm{cum}(\alpha) \times_1 D_1^\top \times_2 D_2^\top \times_3 D_2^\top. \end{aligned} \tag{18}$$

In Section 4, we show how to estimate the factor loading matrices D_1 and D_2 using the diagonal form (16) and (18). Before that, in Section 3.2, we first derive the generalized covariance matrices of discrete ICA and the CCA models (4)–(6) as an extension of the ideas by Yeredor (2000); Todros & Hero (2013).

3.2. Generalized covariance matrices

In this section, we introduce the generalization of the S-covariance matrix for both DICA and the CCA models (4)–(6), which are obtained through the Hessian of the cumulant generating function. We show that (a) the generalized covariance matrices can be used for approximation of the T-cumulant tensors using generalized derivatives and (b) in the DICA case, these generalized covariance matrices have the diagonal form analogous to (11), and, in the CCA case, they have the diagonal form analogous to (16). Therefore, generalized covariance matrices can be seen as a substitute for the T-cumulant tensors in the moment matching framework. This (a) significantly simplifies derivations and the final expressions used for implementation of resulting algorithms and (b) potentially improves the sample complexity, since only the second-order information is used.

Generalized covariance matrices. The idea of generalized covariance matrices is inspired by the similar extension of the ICA cumulants by Yeredor (2000).

The cumulant generating function (CGF) of a multivariate random variable $x \in \mathbb{R}^M$ is defined as

$$K_x(t) = \log \mathbb{E}(e^{t^\top x}), \tag{19}$$

for $t \in \mathbb{R}^M$. The cumulants $\kappa_s(x)$, for $s = 1, 2, 3, \ldots$, are the coefficients of the Taylor series expansion of the CGF evaluated at zero. Therefore, the cumulants are the derivatives of the CGF evaluated at zero: $\kappa_s(x) = \nabla^s K_x(0)$, $s = 1, 2, 3, \ldots$, where $\nabla^s K_x(t)$ is the s-th order derivative of $K_x(t)$ with respect to t. Thus, the expectation of x is the gradient $\mathbb{E}(x) = \nabla K_x(0)$ and the covariance of x is the Hessian $\mathrm{cov}(x) = \nabla^2 K_x(0)$ of the CGF evaluated at zero.

The extension of cumulants then follows immediately: for $t \in \mathbb{R}^M$, we refer to the derivatives $\nabla^s K_x(t)$ of the CGF as the *generalized cumulants*. The respective parameter t is called a *processing point*. In particular, the gradient, $\nabla K_x(t)$, and Hessian, $\nabla^2 K_x(t)$, of the CGF are referred to as the *generalized expectation* and *generalized covariance matrix*, respectively:

$$\mathcal{E}_x(t) := \nabla K_x(t) = \frac{\mathbb{E}(x e^{t^\top x})}{\mathbb{E}(e^{t^\top x})}, \tag{20}$$

$$\mathcal{C}_x(t) := \nabla^2 K_x(t) = \frac{\mathbb{E}(x x^\top e^{t^\top x})}{\mathbb{E}(e^{t^\top x})} - \mathcal{E}_x(t) \mathcal{E}_x(t)^\top. \tag{21}$$

We now outline the key ideas of this section. When a multivariate random variable $\alpha \in \mathbb{R}^K$ has independent components, its CGF $K_\alpha(h) = \log \mathbb{E}(e^{h^\top \alpha})$, for some $h \in \mathbb{R}^K$, is equal to a sum of decoupled terms: $K_\alpha(h) = \sum_k \log \mathbb{E}(e^{h_k \alpha_k})$. Therefore, the Hessian $\nabla^2 K_\alpha(h)$ of the CGF $K_\alpha(h)$ is diagonal (see Appendix C.1). Like covariance matrices, these Hessians (a.k.a. generalized covariance matrices) are subject to the multilinearity property for linear transformations of a vector, hence the resulting diagonal structure of the form (11). This is essentially the previous ICA work (Yeredor, 2000; Todros & Hero, 2013). Below we generalize these ideas first to the discrete ICA case and then to the CCA models (4)–(6).

Discrete ICA generalized covariance matrices. Like covariance matrices, generalized covariance matrices of a vector with independent components are diagonal: they satisfy the multilinearity property $\mathcal{C}_{D\alpha}(h) = D\, \mathcal{C}_\alpha(h) D^\top$, and are equal to covariance matrices when $h = 0$. Therefore, we can expect that the derivations of the diagonal form (11) of the S-covariance matrices extends to the generalized covariance matrices case. By analogy with (10), we define the *generalized S-covariance matrix* of DICA:

$$S(t) := \mathcal{C}_x(t) - \mathrm{diag}[\mathcal{E}_x(t)]. \tag{22}$$

To derive the analog of the diagonal form (11) for $S(t)$, we have to compute all the expectations in (20) and (21) for a Poisson random variable x with the parameter $y = D\alpha$. To illustrate the intuition, we compute here one of these expectations (see Appendix C.2 for further derivations):

$$\begin{aligned} \mathbb{E}(x x^\top e^{t^\top x}) &= \mathbb{E}[\mathbb{E}(x x^\top e^{t^\top x} \mid y)] \\ &= \mathrm{diag}[e^t]\, \mathbb{E}(y y^\top e^{y^\top (e^t - 1)}) \mathrm{diag}[e^t] \\ &= \left(\mathrm{diag}[e^t] D\right) \, \mathbb{E}(\alpha \alpha^\top e^{\alpha^\top h(t)}) \left(\mathrm{diag}[e^t] D\right)^\top, \end{aligned}$$

where $h(t) = D^\top (e^t - 1)$ and e^t denotes an M-vector with the m-th component equal to e^{t_m}. This gives

$$S(t) = \left(\mathrm{diag}[e^t] D\right) \, \mathcal{C}_\alpha\left(h(t)\right) \, \left(\mathrm{diag}[e^t] D\right)^\top, \tag{23}$$

which is a diagonal form similar (and equivalent for $t = 0$)

to (11) since the generalized covariance matrix $\mathcal{C}_\alpha(h)$ of independent sources is diagonal (see (40) in Appendix C.1). Therefore, the generalized S-covariance matrices, estimated at different processing points t, can be used as a substitute of the T-cumulant tensors in the moment matching framework. Interestingly enough, the T-cumulant tensor (10) can be approximated by the generalized covariance matrix via its directional derivative (see Appendix C.5).

CCA generalized covariance matrices. For the CCA models (4)–(6), straightforward generalizations of the ideas from Section 3.1 leads to the following definition of the *generalized CCA S-covariance matrix*:

$$S_{12}(t) := \frac{\mathbb{E}(x_1 x_2^\top e^{t^\top x})}{\mathbb{E}(e^{t^\top x})} - \frac{\mathbb{E}(x_1 e^{t^\top x})}{\mathbb{E}(e^{t^\top x})} \frac{\mathbb{E}(x_2^\top e^{t^\top x})}{\mathbb{E}(e^{t^\top x})}, \quad (24)$$

where the vectors x and t are obtained by vertically stacking x_1 & x_2 and t_1 & t_2 as in (8). In the discrete CCA case, $S_{12}(t)$ is essentially the upper-right block of the generalized S-covariance matrix $S(t)$ of DICA and has the form

$$S_{12}(t) = \left(\operatorname{diag}[e^{t_1}] D_1\right) \mathcal{C}_\alpha(h(t)) \left(\operatorname{diag}[e^{t_2}] D_2\right)^\top, \quad (25)$$

where $h(t) = D^\top (e^t - 1)$ and the matrix D is obtained by vertically stacking D_1 & D_2 by analogy with (8). For non-Gaussian CCA, the diagonal form is

$$S_{12}(t) = D_1 \mathcal{C}_\alpha\left(h(t)\right) D_2^\top, \quad (26)$$

where $h(t) = D_1^\top t_1 + D_2^\top t_2$. Finally, for mixed CCA,

$$S_{12}(t) = D_1 \mathcal{C}_\alpha\left(h(t)\right) \left(\operatorname{diag}[e^{t_2}] D_2\right)^\top, \quad (27)$$

where $h(t) = D_1^\top t_1 + D_2^\top (e^{t_2} - 1)$. Since the generalized covariance matrix of the sources $\mathcal{C}_\alpha(\cdot)$ is diagonal, expressions (25)–(27) have the desired diagonal form (see Appendix C.4 for detailed derivations).

4. Joint diagonalization algorithms

The standard algorithms such as TPM or orthogonal joint diagonalization cannot be used for the estimation of D_1 and D_2. Indeed, even after whitening, the matrices appearing in the diagonal form (16)&(18) or (25)–(27) are *not* orthogonal. As an alternative, we use Jacobi-like non-orthogonal diagonalization algorithms (Fu & Gao, 2006; Iferroudjene et al., 2009; Luciani & Albera, 2010). These algorithms are discussed in this section and in Appendix F.

The estimation of the factor loading matrices D_1 and D_2 of the CCA models (4)–(6) via non-orthogonal joint diagonalization algorithms consists of the following steps: (a) construction of a set of matrices, called *target matrices*, to be jointly diagonalized (using finite sample estimators), (b) a whitening step, (c) a non-orthogonal joint diagonalization step, and (d) the final estimation of the factor loading matrices (Appendix E.5).

Target matrices. There are two ways to construct target matrices: either with the CCA S-matrices (15) and T-cumulants (17) (only DCCA) or the generalized covariance matrices (24) (D/N/MCCA). These matrices are estimated with finite sample estimators (Appendices D.1 & D.2).

The (computationally efficient) construction of target matrices from S- and T-cumulants was discussed by Podosinnikova et al. (2015) and we recall it in Appendix E.1. Alternatively, the target matrices can be constructed by estimating the generalized S-covariance matrices at $P+1$ processing points $0, t_1, \dots, t_P \in \mathbb{R}^{M_1+M_2}$:

$$\{S_{12} = S_{12}(0),\ \ S_{12}(t_1),\ \ \dots,\ \ S_{12}(t_P)\}, \quad (28)$$

which also have the diagonal form (25)–(27). It is interesting to mention the connection between the T-cumulants and the generalized S-covariance matrices. The T-cumulant can be approximated via the directional derivative of the generalized covariance matrix (see Appendix C.5). However, in general, e.g., $S_{12}(t)$ with $t = [t_1; 0]$ is not exactly the same as $T_{121}(t_1)$ and the former can be non-zero even when the latter is zero. This is important since order-4 and higher statistics are used with the method of moments when there is a risk that an order-3 statistic is zero like for symmetric sources. In general, the use of higher-order statistics increases the sample complexity and makes the resulting expressions quite complicated. Therefore, replacing the T-cumulants with the generalized S-covariance matrices is potentially beneficial.

Whitening. The matrices $W_1 \in \mathbb{R}^{K \times M_1}$ and $W_2 \in \mathbb{R}^{K \times M_2}$ are called *whitening matrices* of S_{12} if

$$W_1 S_{12} W_2^\top = I_K, \quad (29)$$

where I_K is the K-dimensional identity matrix. W_1 and W_2 are only defined up to multiplication by any invertible matrix $Q \in \mathbb{R}^{K \times K}$, since any pair of matrices $\widetilde{W}_1 = QW_1$ and $\widetilde{W}_2 = Q^{-\top} W_2$ also satisfy (29). In fact, using higher-order information (i.e. the T-cumulants or the generalized covariances for $t \neq 0$) allows to solve this ambiguity.

The whitening matrices can be computed via SVD of S_{12} (see Appendix E.2). When M_1 and M_2 are too large, one can use a randomized SVD algorithm (see, e.g., Halko et al., 2011) to avoid the construction of the large matrix S_{12} and to decrease the computational time.

Non-orthogonal joint diagonalization (NOJD). Let us consider joint diagonalization of the generalized covariance matrices (28) (the same procedure holds for the S- and T-cumulants (43); see Appendix E.3). Given the whitening matrices W_1 and W_2, the transformation of the generalized covariance matrices (28) gives $P+1$ matrices

$$\{W_1 S_{12} W_2^\top,\ W_1 S_{12}(t_p) W_2^\top,\quad p = 1, \dots, P\}, \quad (30)$$

where each matrix is in $\mathbb{R}^{K \times K}$ and has reduced dimension since $K < M_1, M_2$. In practice, finite sample estimators are used to construct (28) (see Appendices D.1 and D.2).

Due to the diagonal form (16) and (25)–(27), each matrix in (28) has the form[2] $(W_1D_1)\,\mathrm{diag}(\cdot)\,(W_2D_2)^\top$. Both D_1 and D_2 are (full) K-rank matrices and W_1 and W_2 are K-rank by construction. Therefore, the square matrices $V_1 = W_1D_1$ and $V_2 = W_2D_2$ are invertible. From (16) and (29), we get $V_1\mathrm{cov}(\alpha)V_2^\top = I$ and hence $V_2 = \mathrm{diag}[\mathrm{var}(\alpha)^{-1}]V_1^{-1}$ (the covariance matrix of the sources is diagonal and we assume they are non-deterministic, i.e. $\mathrm{var}(\alpha) \neq 0$). Substituting this into $W_1S_{12}(t)W_2^\top$ and using the diagonal form (25)–(27), we obtain that the matrices in (28) have the form $V_1\mathrm{diag}(\cdot)V_1^{-1}$. Hence, we deal with the problem of the following type: Given P non-defective (a.k.a. diagonalizable) matrices $\mathcal{B} = \{B_1, \ldots, B_P\}$, where each matrix $B_p \in \mathbb{R}^{K\times K}$, find and invertible matrix $Q \in \mathbb{R}^{K\times K}$ such that

$$Q\mathcal{B}Q^{-1} = \{QB_1Q^{-1}, \ldots, QB_PQ^{-1}\} \qquad (31)$$

are (jointly) as diagonal as possible. This can be seen as a joint non-symmetric eigenvalue problem. This problem should not be confused with the classical joint diagonalization problem by congruence (JDC), where Q^{-1} is replaced by $Q^\top$, except when Q is an orthogonal matrix (Luciani & Albera, 2010). JDC is often used for ICA algorithms or moment matching based algorithms for graphical models when a whitening step is not desirable (see, e.g., Kuleshov et al. (2015) and references therein). However, neither JDC nor the orthogonal diagonalization-type algorithms (such as, e.g., the tensor power method by Anandkumar et al., 2014) are applicable for the problem (31).

To solve the problem (31), we use the Jacobi-like non-orthogonal joint diagonalization (NOJD) algorithms (e.g., Fu & Gao, 2006; Iferroudjene et al., 2009; Luciani & Albera, 2010). These algorithms are an extension of the orthogonal joint diagonalization algorithms based on Jacobi (=Givens) rotations (Golub & Van Loan, 1996; Bunse-Gerstner et al., 1993; Cardoso & Souloumiac, 1996). Due to the space constraint, the description of the NOJD algorithms is moved to Appendix F. Although these algorithms are quite stable in practice, we are not aware of any theoretical guarantees about their convergence or stability to perturbation.

Spectral algorithm. By analogy with the orthogonal case (Cardoso, 1989; Anandkumar et al., 2012), we can easily extend the idea of the spectral algorithm to the non-orthogonal one. Indeed, it amounts to performing whitening as before and constructing only one matrix with the diagonal structure, e.g., $B = W_1S_{12}(t)W_2^\top$ for some t. Then, the matrix Q is obtained as the matrix of the eigenvectors of B. The vector t can be, e.g., chosen as $t = Wu$, where $W = [W_1; W_2]$ and $u \in \mathbb{R}^K$ is a vector sampled uniformly at random.

This spectral algorithm and the NOJD algorithms are closely connected. In particular, when B has real eigenvectors, the spectral algorithm is equivalent to NOJD of B. Indeed, in such case, NOJD boils down to an algorithm for a non-symmetric eigenproblem (Eberlein, 1962; Ruhe, 1968). In practice, however, due to the presence of noise and finite sample errors, B may have complex eigenvectors. In such case, the spectral algorithm is different from NOJD. Importantly, the joint diagonalization type algorithms are known to be more stable in practice (see, e.g., Bach & Jordan, 2003; Podosinnikova et al., 2015).

While deriving precise theoretical guarantees is beyond the scope of this paper, the techniques outlined by Anandkumar et al. (2012) for the spectral algorithm for latent Dirichlet Allocation can potentially be extended. The main difference is obtaining the analogue of the SVD accuracy (Lemma C.3, Anandkumar et al., 2013) for the eigen decomposition. This kind of analysis can potentially be extended with the techniques outlined in (Chapter 4, Stewart & Sun, 1990). Nevertheless, with appropriate parametric assumptions on the sources, we expect that the above described extension of the spectral algorithm should lead to similar guarantee as the spectral algorithm of Anandkumar et al. (2012).

See Appendix E for some important implementation details, including the choice of the processing points.

5. Experiments

Synthetic data. We sample synthetic data to have ground truth information for comparison. We sample from linear DCCA which extends linear CCA (7) such that each view is $x_j \sim \mathrm{Poisson}(D_j\alpha + F_j\beta_j)$. The sources $\alpha \sim \mathrm{Gamma}(c, b)$ and the noise sources $\beta_j \sim \mathrm{Gamma}(c_j, b_j)$, for $j = 1, 2$, are sampled from the gamma distribution (where b is the rate parameter). Let $s_j \sim \mathrm{Poisson}(D_j\alpha)$ be the part of the sample due to the sources and $n_j \sim \mathrm{Poisson}(F_j\beta_j)$ be the part of the sample due to the noise (i.e., $x_j = s_j + n_j$). Then we define the expected sample length due to the sources and noise, respectively, as $L_{js} := \mathbb{E}[\sum_m s_{jm}]$ and $L_{jn} := \mathbb{E}[\sum_m n_{jm}]$. For sampling, the target values $L_s = L_{1s} = L_{2s}$ and $L_n = L_{1n} = L_{2n}$ are fixed and the parameters b and b_j are accordingly set to ensure these values: $b = Kc/L_s$ and $b_j = K_jc_j/L_n$ (see Appendix B.2 of Podosinnikova et al. (2015)). For the larger dimensional example (Fig. 2, right), each column of the matrices D_j and F_j, for $j = 1, 2$, is sampled from the symmetric Dirichlet distribution with the concentration parameter equal to 0.5. For the smaller 2D example (Fig. 2, left), they are fixed: $D_1 = D_2$ with $[D_1]_1 = [D_1]_2 = 0.5$ and $F_1 = F_2$ with $[F_1]_{11} = [F_1]_{22} = 0.9$ and $[F_1]_{12} = [F_1]_{21} = 0.1$. For each experiment, D_j and F_j, for $j = 1, 2$, are sampled once and, then, the x-

[2] Note that when the diagonal form has terms $\mathrm{diag}[e^t]$, we simply multiply the expression by $\mathrm{diag}[e^{-t}]$.

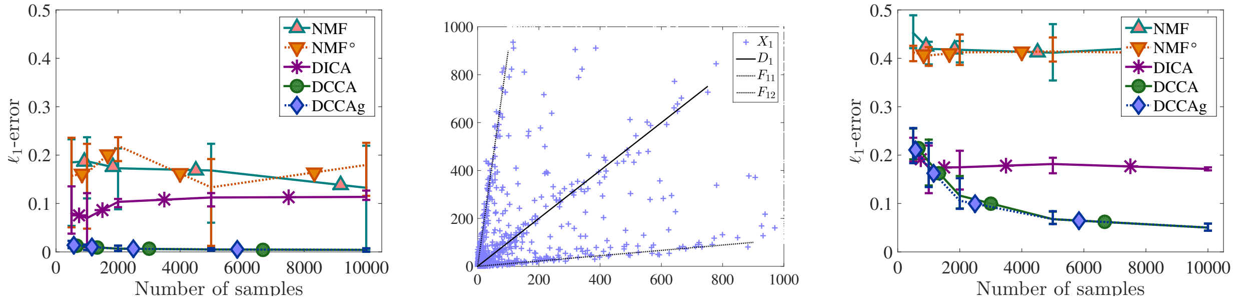

Figure 2. Synthetic experiment with discrete data. **Left (2D example)**: $M_1 = M_2 = K_1 = K_2 = 2$, $K = 1$, $c = c_1 = c_2 = 0.1$, and $L_s = L_n = 100$; **middle (2D data)**: the x_1-observations and factor loading matrices for the 2D example (F_{1j} denotes the j-th column of the noise factor matrix F_1); **right (20D example)**: $M_1 = M_2 = K_1 = K_2 = 20$, $K = 10$, $L_s = L_n = 1,000$, $c = 0.3$, and $c_1 = c_2 = 0.1$.

nato	otan	work	travail	board	commission	nisga	nisga
kosovo	kosovo	workers	négociations	wheat	blé	treaty	autochtones
forces	militaires	strike	travailleurs	farmers	agriculteurs	aboriginal	traité
military	guerre	legislation	grève	grain	administration	agreement	accord
war	international	union	emploi	producers	producteurs	right	droit
troops	pays	agreement	droit	amendment	grain	land	nations
country	réfugiés	labour	syndicat	market	conseil	reserve	britannique
world	situation	right	services	directors	ouest	national	indiennes
national	paix	services	accord	western	amendement	british	terre
peace	yougoslavie	negotiations	voix	election	comité	columbia	colombie

Table 1. Factor loadings (a.k.a. topics) extracted from the Hansard collection for $K = 20$ with DCCA.

observations are sampled for different sample sizes $N = \{500, 1,000, 2,000, 5,000, 10,000\}$, 5 times for each N.

Metric. The evaluation is performed on a matrix D obtained by stacking D_1 and D_2 vertically (see also the comment after Thm. 1). As in Podosinnikova et al. (2015), we use as evaluation metric the normalized ℓ_1-error between a recovered matrix $\widehat{D}$ and the true matrix D with the best permutation of columns $\mathrm{err}_1(\widehat{D}, D) := \min_{\pi \in \mathrm{PERM}} \frac{1}{2K} \sum_k \|\widehat{d}_{\pi_k} - d_k\|_1 \in [0, 1]$. The minimization is over the possible permutations $\pi \in \mathrm{PERM}$ of the columns of $\widehat{D}$ and can be efficiently obtained with the Hungarian algorithm for bipartite matching. The (normalized) ℓ_1-error takes the values in $[0,\ 1]$ and smaller values of this error indicate better performance of an algorithm.

Algorithms. We compare DCCA (implementation with the S- and T-cumulants) and DCCAg (implementation with the generalized S-covariance matrices and the processing points initialized as described in Appendix E.4) to DICA and the non-negative matrix factorization (NMF) algorithm with multiplicative updates for divergence (Lee & Seung, 2000). To run DICA or NMF, we use the stacking trick (8). DCCA is set to estimate K components. DICA is set to estimate either $K_0 = K + K_1 + K_2$ or $M = M_1 + M_2$ components (whichever is the smallest, since DICA cannot work in the over-complete case). NMF is always set to estimate K_0 components. For the evaluation of DICA/NMF, the K columns with the smallest ℓ_1-error are chosen. NMF° stands for NMF initialized with a matrix D of the form (8) with induced zeros; otherwise NMF is initialized with (uniformly) random non-negative matrices. The running times are discussed in Appendix G.5.

Synthetic experiment. We first perform an experiment with discrete synthetic data in 2D (Fig. 2) and then repeat the same experiment when the size of the problem is 10 times larger. In practice, we observed that for $K_0 < M$ all models work approximately equally well, except for NMF which breaks down in high dimensions. In the over-complete case as in Fig. 2, DCCA works better. A continuous analogue of this experiment is presented in Appendix G.1.

Real data (translation). Following Vinokourov et al. (2002), we illustrate the performance of DCCA by extracting bilingual topics from the Hansard collection (Vinokourov & Girolami, 2002) with aligned English and French proceedings of the 36-th Canadian Parliament. In Table 1, we present some of the topics extracted after running DCCA with $K = 20$ (see all the details in Appendices G.3 and G.4). The (Matlab/C++) code for reproducing the experiments of this paper is available at https://github.com/anastasia-podosinnikova/cca.

Conclusion

We have proposed the first identifiable versions of CCA, together with moment matching algorithms which allow the identification of the loading matrices in a semi-parametric framework, where no assumptions are made regarding the distribution of the source or the noise. We also introduce new sets of moments (our generalized covariance matrices), which could prove useful in other settings.

Acknowledgements

This work was partially supported by the MSR-Inria Joint Center.

References

Anandkumar, A. and Sedghi, H. Learning mixed membership community models in social tagging networks through tensor methods. *CoRR*, arXiv:1503.04567v2, 2015.

Anandkumar, A., Foster, D.P., Hsu, D., Kakade, S.M., and Liu, Y.-K. A spectral algorithm for latent Dirichlet allocation. In *Adv. NIPS*, 2012.

Anandkumar, A., Foster, D.P., Hsu, D., Kakade, S.M., and Liu, Y.-K. A spectral algorithm for latent Dirichlet allocation. *CoRR*, abs:1204.6703v4, 2013.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S.M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15:2773–2832, 2014.

Arora, S. and Kannan, R. Learning mixtures of separated nonspherical Gaussians. *Ann. Appl. Probab.*, 15(1A): 69–92, 2005.

Arora, S., Ge, R., and Moitra, A. Learning topic models – Going beyond SVD. In *Proc. FOCS*, 2012.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. A practical algorithm for topic modeling with provable guarantees. In *Proc. ICML*, 2013.

Bach, F. and Jordan, M.I. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2003.

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

Balle, B., Hamilton, W.L., and Pineau, J. Method of moments for learning stochastic languages: Unified presentation and empirical comparison. In *Proc. ICML*, 2014.

Bartholomew, D.J. *Latent Variable Models and Factor Analysis*. Wiley, 1987.

Basilevsky, A. *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley, 1994.

Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

Browne, M.W. The maximum-likelihood solution in inter-battery factor analysis. *Br. J. Math. Stat. Psychol.*, 32(1): 75–86, 1979.

Bunse-Gerstner, A., Byers, R., and Mehrmann, V. Numerical methods for simultaneous diagonalization. *SIAM J. Matrix Anal. Appl.*, 14(4):927–949, 1993.

Cardoso, J.-F. Source separation using higher order moments. In *Proc. ICASSP*, 1989.

Cardoso, J.-F. and Comon, P. Independent component analysis, A survey of some algebraic methods. In *Proc. ISCAS*, 1996.

Cardoso, J.-F. and Souloumiac, A. Blind beamforming for non Gaussian signals. In *IEE Proc-F*, 1993.

Cardoso, J.-F. and Souloumiac, A. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17 (1):161–164, 1996.

Comon, P. Independent component analysis, A new concept? *Signal Process.*, 36(3):287–314, 1994.

Comon, P. and Jutten, C. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

Eberlein, P.J. A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix. *J. Soc. Indust. Appl. Math.*, 10(1):74–88, 1962.

Fu, T. and Gao, X. Simultaneous diagonalization with similarity transformation for non-defective matrices. In *Proc. ICASSP*, 2006.

Ge, R. and Zou, J. Rich component analysis. In *Proc. ICML*, 2016.

Golub, G.H. and Van Loan, C.F. *Matrix Computations*. John Hopkins University Press, 3rd edition, 1996.

Gong, Y., Ke, Q., Isard, M., and Lazebnik, S. A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. J. Comput. Vis.*, 106(2):210–233, 2014.

Haghighi, A., Liang, P., Kirkpatrick, T.B., and Klein, D. Learning bilingual lexicons from monolingual corpora. In *Proc. ACL*, 2008.

Halko, N., Martinsson, P.G., and Tropp, J.A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

Hardoon, D.R., Szedmak, S.R., and Shawe-Taylor, J.R. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12): 2639–2664, 2004.

Hotelling, H. Relations between two sets of variates. *Biometrica*, 28(3/4):321–377, 1936.

Hsu, D. and Kakade, S.M. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proc. ITCS*, 2013.

Iferroudjene, R., Abed Meraim, K., and Belouchrani, A. A new Jacobi-like method for joint diagonalization of arbitrary non-defective matrices. *Appl. Math. Comput.*, 211:363–373, 2009.

Janzamin, M., Sedghi, H., and Anandkumar, A. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *CoRR*, arXiv:1506.08473v3, 2016.

Jutten, C. *Calcul neuromimétique et traitement du signal: Analyse en composantes indépendantes*. PhD thesis, INP-USM Grenoble, 1987.

Jutten, C. and Hérault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetric architecture. *Signal Process.*, 24(1):1–10, 1991.

Klami, A., Virtanen, S., and Kaski, S. Bayesian exponential family projections for coupled data sources. In *Proc. UAI*, 2010.

Klami, A., Virtanen, S., and Kaski, S. Bayesian canonical correlation analysis. *J. Mach. Learn. Res.*, 14:965–1003, 2013.

Kolda, T.G. and Bader, B.W. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009.

Kuleshov, V., Chaganty, A.T., and Liang, P. Tensor factorization via matrix factorization. In *Proc. AISTATS*, 2015.

Lee, D.D. and Seung, H.S. Algorithms for non-negative matrix factorization. In *Adv. NIPS*, 2000.

Luciani, X. and Albera, L. Joint eigenvalue decomposition using polar matrix factorization. In *Proc. LVA ICA*, 2010.

Murphy, K.P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Podosinnikova, A., Bach, F., and Lacoste-Julien, S. Rethinking LDA: Moment matching for discrete ICA. In *Adv. NIPS*, 2015.

Roweis, S. EM algorithms for PCA and SPCA. In *Adv. NIPS*, 1998.

Ruhe, A. On the quadratic convergene of a generalization of the Jacobi method to arbitrary matrices. *BIT Numer. Math.*, 8(3):210–231, 1968.

Shawe-Taylor, J.R. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Slapak, A. and Yeredor, A. Charrelation matrix based ICA. In *Proc. LVA ICA*, 2012a.

Slapak, A. and Yeredor, A. Charrelation and charm: Generic statistics incorporating higher-order information. *IEEE Trans. Signal Process.*, 60(10):5089–5106, 2012b.

Socher, R. and Fei-Fei, L. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proc. CVPR*, 2010.

Song, L., Anandkumar, A., Dai, B., and Xie, B. Nonparametric estimation of multi-view latent variable models. In *Proc. ICML*, 2014.

Stewart, G.W. and Sun, J. *Matrix Perturbation Theory*. Academic Press, 1990.

Sübakan, Y.C., Traa, J., and Smaragdis, P. Spectral learning of mixture of hidden Markov models. In *Adv. NIPS*, 2014.

Tipping, M.E. and Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Series B*, 61(3):611–622, 1999.

Todros, K. and Hero, A.O. Measure transformed independent component analysis. *CoRR*, arXiv:1302.0730v2, 2013.

Tung, H.-Y. and Smola, A. Spectral methods for Indian buffet process inference. In *Adv. NIPS*, 2014.

Vinokourov, A. and Girolami, M. A probabilistic framework for the hierarchic organisation and classification of document collections. *J. Intell. Inf. Syst.*, 18(2/3):153–172, 2002.

Vinokourov, A., Shawe-Taylor, J.R., and Cristianini, N. Inferring a semantic representation of text via cross-language correlation analysis. In *Adv. NIPS*, 2002.

Virtanen, S. Bayesian exponential family projections. Master's thesis, Aalto University, 2010.

Wang, Y. and Zhu, J. Spectral methods for supervised topic models. In *Adv. NIPS*, 2014.

Yeredor, A. Blind source separation via the second characteristic function. *Signal Process.*, 80(5):897–902, 2000.