Beyond CCA: Moment Matching for Multi-View Models

Anastasia Podosinnikova, Francis Bach, Simon Lacoste-Julien

Beyond CCA: Moment Matching for Multi-View Models

Anastasia Podosinnikova Francis Bach Simon Lacoste-Julien
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., Cristianini & Shawe-Taylor, 2000; 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; Roweis, 1998). Indeed, as FA and PPCA are identifiable only up to multiplication by any orthogonal 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, 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 in order 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

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

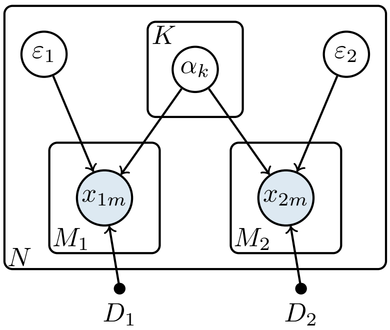

Figure 1: Graphical models for non-Gaussian (4), discrete (5), and mixed (6) CCA.

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).

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).

By analogy with Podosinnikova et al. (2015), we can 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. Depending on the context, the matrices D_1 and D_2 are called differently: topic matrices[1] in the topic learning context, factor loading or projection matrices in the FA and/or PPCA context, mixing matrices in the ICA context, or dictionaries in the dictionary learning context. In this paper, we will use the name *factor loading matrices* to refer to D_1 and D_2.

Relation between PCA and CCA. The important 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 in 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

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

[1] Note that Podosinnikova et al. (2015) show that DICA is closely connected (and under some conditions is equivalent) to latent Dirichlet allocation (Blei et al., 2003). Due to the close relation of DCCA and DICA, the former is thus closely related to the multi-view topic models (see, e.g., Blei & Jordan, 2003).

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

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

one can rewrite the model as $x = Dz$. If the noise sources β_1 and β_2 are assumed to 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 (as well as in the models (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.

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 A 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 $K \le \min(M_1, M_2)$, 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_1P is the factor loading matrix of the first view, than D_2P 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 the experimental 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:

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

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

$$\begin{aligned} S &:= \mathrm{cov}(x) - \mathrm{diag}\left[\mathbb{E}x\right], \\ [T]_{m_1m_2m_3} &:= \mathrm{cum}(x)_{m_1m_2m_3} + [\tau]_{m_1m_2m_3}, \end{aligned} \tag{10}$$

where the indices m_1, m_2, and m_3 take the values in $1, \ldots, M$, and

$$[\tau]_{m_1m_2m_3} = 2\delta_{m_1m_2m_3}\mathbb{E}x_{m_1} - \delta_{m_2m_3}\mathrm{cov}(x)_{m_1m_2} - \delta_{m_1m_3}\mathrm{cov}(x)_{m_1m_2} - \delta_{m_1m_2}\mathrm{cov}(x)_{m_1m_3}$$

with δ being the Kronecker delta. For completeness, we outline the derivation by Podosinnikova et al. (2015) below. Denoting $y := D\alpha$, one obtains by the law of total expectation that $\mathbb{E}(x) = \mathbb{E}(x|y) = \mathbb{E}(y)$ and by the law of total covariance:

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

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

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

which is called the *diagonal form* since the covariance $\mathrm{cov}(\alpha)$ and cumulant $\mathrm{cum}(\alpha)$ of the independent sources are diagonal. Note that $\times_i$ denotes the i-mode tensor-matrix product (see, e.g., Kolda & Bader, 2009). 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):

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

Let $y := D\alpha + \varepsilon$ and S and T are defined as in (10). Then a simple extension of the derivations from above gives $S = \mathrm{cov}(y)$ and $T = \mathrm{cum}(y)$. 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

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

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:

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

which is due to the independence $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$. Similarly, the cumulant $\mathrm{cum}(\varepsilon)$ 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 $\mathrm{cov}(\varepsilon)$ and cumulant $\mathrm{cum}(\varepsilon)$ 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*[2] as the cross-covariance matrix of x_1 and x_2:

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

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

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

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

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 $j = 1, 2$:

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

where the indices m_1, m_2, and $\tilde{m}_j$ take the values $m_1 \in 1, \dots, M_1$, $m_2 \in 1, \dots, M_2$, and $\tilde{m}_j \in 1, \dots, M_j$. 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[3] 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 $s = 1, 2, 3, \dots$, 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: $\kappa_s(x) = \nabla^s K_x(0)$, $s = 1, 2, 3, \dots$, 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:

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

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

Some properties of these statistics and their natural finite sample estimators are analyzed by Slapak & Yeredor (2012b).

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 B.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).

[3] We find the name "generalized covariance matrix" to be more meaningful than "charrelation" matrix as was proposed by previous authors (see, e.g. Slapak & Yeredor, 2012a,b).

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 $y = D\alpha$. Just to provide some intuition, we compute here one of these expectations (see Appendix B.2 for further derivations):

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

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 (42) in Appendix B.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 B.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*:

$$S_{12}(t) := \frac{\mathbb{E}(x_1x_2^\top e^{t^\top x})}{\mathbb{E}(e^{t^\top x})} - \frac{\mathbb{E}(x_1e^{t^\top x})}{\mathbb{E}(e^{t^\top x})}\frac{\mathbb{E}(x_2^\top e^{t^\top x})}{\mathbb{E}(e^{t^\top x})}, \tag{24}$$

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

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

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

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

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

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

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 B.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 E.

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 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 D.4).

Matrices for diagonalization. There are two ways to construct matrices for subsequent joint diagonalization: either with the CCA S-matrices (15) and T-cumulants (17) (only DCCA) or the generalized covariance matrices (24) (D/N/MCCA). Given a dataset, these matrices are estimated using natural finite sample estimators (see Appendices C.1 and C.2).

When dealing with S- and T-cumulants, the matrices are obtained via tensor projections. We define a projection $\mathcal{T}(v) \in \mathbb{R}^{M_1 \times M_2}$ of a third-order tensor $\mathcal{T} \in \mathbb{R}^{M_1 \times M_2 \times M_3}$ onto a vector $v \in \mathbb{R}^{M_3}$ as

$$[\mathcal{T}(v)]_{m_1 m_2} := \sum_{m_3=1}^{M_3} [\mathcal{T}]_{m_1 m_2 m_3} v_{m_3}. \tag{28}$$

Note that the projection $\mathcal{T}(v)$ is a matrix. Therefore, given $2P$ vectors $\{v_{11}, v_{21}, v_{12}, v_{22}, \dots, v_{1P}, v_{2P}\}$, one can construct $2P + 1$ matrices

$$\{S_{12}, \quad T_{121}(v_{1p}), \quad T_{122}(v_{2p}), \quad \text{for } p = 1, \dots, P\}, \tag{29}$$

which have the diagonal form (16) and (18). Importantly, the tensors are never constructed (see Anandkumar et al. (2012, 2014); Podosinnikova et al. (2015) and Appendix C.2).

Alternatively to (29), the set of matrices can be constructed by estimating the generalized S-covariance matrices at $P + 1$ processing points $0, t_1, \dots, t_P \in \mathbb{R}^{M_1 + M_2}$:

$$\{S_{12} = S_{12}(0), \quad S_{12}(t_1), \quad \dots, \quad S_{12}(t_P)\}, \tag{30}$$

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 B.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. 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

$$W_1 S_{12} W_2^\top = I_K, \tag{31}$$

where I_K is the K-dimensional identity matrix. Such matrices 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 (31). 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 D.1). 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). For simplicity, let us consider joint diagonalization of the generalized covariance matrices (30) (the same procedure holds for the S- and T-cumulants (29); see Appendix D.2). Given the whitening matrices W_1 and W_2, the transformation of the generalized covariance matrices (30) gives $P + 1$ matrices

$$\{W_1 S_{12} W_2^\top, \; W_1 S_{12}(t_p) W_2^\top, \quad p = 1, \dots, P\}, \tag{32}$$

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 (30) (see Appendices C.1 and C.2).

Due to the diagonal form (16) and (25)–(27), each matrix in (30) has the form[4] $(W_1 D_1) \operatorname{diag}(\cdot) (W_2 D_2)^\top$. 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 $V_1 = W_1 D_1$ and $V_2 = W_2 D_2$ are invertible. From (16) and (31), we get $V_1 \operatorname{cov}(\alpha) V_2^\top = I$ and hence $V_2 = \operatorname{diag}[\operatorname{var}(\alpha)^{-1}] V_1^{-1}$ (the covariance matrix of the sources is diagonal and we assume they are non-deterministic, i.e. $\operatorname{var}(\alpha) \neq 0$). Substituting this into $W_1 S_{12}(t) W_2^\top$ and using the diagonal form (25)–(27), we obtain that the matrices in (30) have the form $V_1 \operatorname{diag}(\cdot) V_1^{-1}$. Hence, we deal with the problem of the following type: Given P non-defective (a.k.a. diagonalizable) matrices $\mathcal{B} = \{B_1, \dots, B_P\}$, where each matrix $B_p \in \mathbb{R}^{K \times K}$, find and invertible matrix $Q \in \mathbb{R}^{K \times K}$ such that

$$Q \mathcal{B} Q^{-1} = \{Q B_1 Q^{-1}, \; \dots, \; Q B_P Q^{-1}\} \tag{33}$$

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

are (jointly) as diagonal as possible. This can be seen as a joint non-symmetric eigenvalue problem. This problem should not be confused with 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 Anandkumar et al., 2014) are applicable for the problem (33).

To solve the problem (33), 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 E. 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., $B = W_1 S_{12}(t) W_2^\top$ 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).

Some important implementation details, including the choice of the processing points, are discussed in Appendix D.

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 $b_j = K_j c_j / L_n$ (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-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.

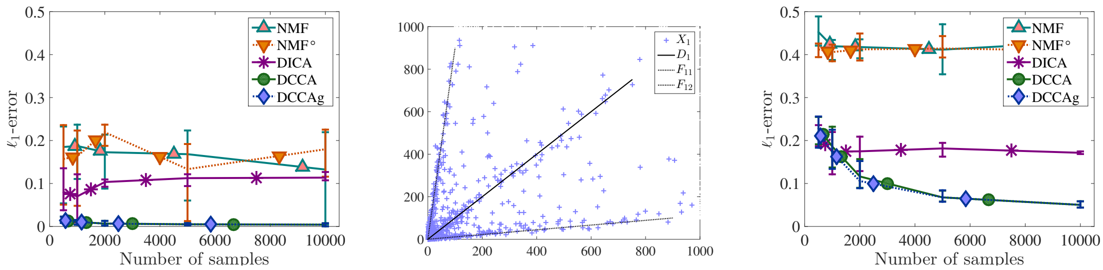

Figure 2: Synthetic experiment with discrete data. **Left (2D example)**: $M_1 = M_2 = K_1 = K_2 = 2$, $K = 1$, $c = c_1 = c_2 = 0.1$, and $L_s = L_n = 100$; **middle (2D data)**: the x_1-observations and factor loading matrices for the 2D example (F_{1j} denotes the j-th column of the noise factor matrix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

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 D.3) 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.

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 F.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. We first pre-process the dataset with the NLTK toolbox by Bird et al. (2009) (see details in Appendix F.3) to obtain $N = 12,932$ aligned documents with $M_1 = M_2 = 5,000$ English and French words. We then run DCCA with $K = 20$ to extract the aligned topics. Some of the extracted topics are presented in Table 1 and all the topics as well as the detailed description of this experiment are presented in Appendix F.3.

Running time. For the real experiment above, the runtime of DCCA algorithm is 24 seconds including 22 seconds for SVD at the whitening step. In general, the computational complexity of the D/N/MCCA algorithms is bounded by the time of SVD plus $O(RNK) + O(NK^2)$, where R is the largest number of non-zero components in the stacked vector $x = [x_1;\ x_2]$, plus the time of NOJD for the matrices of size K. In practice, DCCAg is faster than DCCA.

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 a new sets of moments (our generalized covariance matrices), which could prove useful in other settings.

Acknowledgments. 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.

Bird, S., Loper, E., and Klei, E. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

Blei, D.M. and Jordan, M.I. Modeling annotated data. In *Proc. SIGIR*, 2003.

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.

Cristianini, N. and Shawe-Taylor, J.R. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

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.

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.

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*.

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.

6 Appendix

The appendix is organized as follows.

- In Appendix A, we present the proof of Theorem 1 stating the *identifiability of the CCA models* (4)–(6).
- In Appendix B, we provide some details for the *generalized covariance matrices*: the form of the generalized covariance matrices of independent variables (Appendix B.1), the derivations of the diagonal form of the generalized covariance matrices of discrete ICA (Appendix B.3), the derivations of the diagonal form of the generalized covariance matrices of the CCA models (4)–(6) (Appendix B.4), and approximation of the T-cumulants with the generalized covariance matrix (Appendix B.5).
- In Appendix C, we provide expressions for natural *finite sample estimators of the generalized covariance matrices* and the T-cumulant tensors for the considered CCA models.
- In Appendix D, we discuss some rather technical *implementation details*: computation of whitening matrices (Appendix D.1), selection of the projection vectors for the T-cumulants and the processing points for the generalized covariance matrices (Appendix D.3), and the final estimation of the factor loading matrices (Appendix D.4).
- In Appendix E, we describe the *non-orthogonal joint diagonalization algorithms* used in this paper.
- In Appendix F, we present *some supplementary experiments*: a continuous analog of the synthetic experiment from Section 5 (Appendix F.1), an experiment to analyze the sensitivity of the DCCA algorithm with the generalized S-covariance matrices to the choice of the processing points (Appendix F.2), and a detailed description of the experiment with the real data from Section 5 (Appendices F.3 and F.4).

A Identifiability

In this section, we prove that the factor loading matrices D_1 and D_2 of the non-Gaussian CCA (4), discrete CCA (5), and mixed CCA (6) models are identifiable up to permutation and scaling if at most one source α_k is Gaussian. We provide a complete proof for the non-Gaussian CCA case and show that the other two cases can be proved by analogy.

A.1 Identifiability of non-Gaussian CCA (4)

The proof uses the notion of the second characteristic function (SCF) of a random variable $x \in \mathbb{R}^M$:

$$\phi_x(t) = \log \mathbb{E}(e^{it^\top x}),$$

for all $t \in \mathbb{R}^M$. The SCF completely defines the probability distribution of x (see, e.g., Jacod & Protter, 2004). Important difference between the SCF and the cumulant generating function (19) is that the former always exists.

The following property of the SCF is of central importance for the proof: if two random variables, z_1 and z_2, are independent, then $\phi_{A_1 z_1 + A_2 z_2}(t) = \phi_{z_1}(A_1^\top t) + \phi_{z_2}(A_2^\top t)$, where A_1 and A_2 are any matrices of compatible sizes.

We can now use our CCA model to derive an expression of $\phi_x(t)$. Indeed, defining a vector x by stacking the vectors x_1 and x_2, the SCF of x for any $t = [t_1;\ t_2]$, takes the form

$$\begin{aligned}
\phi_x(t) &= \log \mathbb{E}(e^{it_1^\top x_1 + it_2^\top x_2}) \\
&\overset{(a)}{=} \log \mathbb{E}(e^{i\alpha^\top (D_1^\top t_1 + D_2^\top t_2) + i\varepsilon_1^\top t_1 + i\varepsilon_2^\top t_2}) \\
&\overset{(b)}{=} \log \mathbb{E}(e^{i\alpha^\top (D_1^\top t_1 + D_2^\top t_2)}) \\
&\qquad\qquad + \log \mathbb{E}(e^{i\varepsilon_1^\top t_1}) + \log \mathbb{E}(e^{i\varepsilon_2^\top t_2}) \\
&= \phi_\alpha(D_1^\top t_1 + D_2^\top t_2) + \phi_{\varepsilon_1}(t_1) + \phi_{\varepsilon_2}(t_2),
\end{aligned}$$

where in (a) we substituted the definition (4) of x_1 and x_2 and in (b) we used the independence $\alpha \perp\!\!\!\perp \varepsilon_1 \perp\!\!\!\perp \varepsilon_2$. Therefore, the blockwise mixed derivatives of ϕ_x are equal to

$$\partial_1 \partial_2 \phi_x(t) = D_1 \phi_\alpha''(D_1^\top t_1 + D_2^\top t_2) D_2^\top, \tag{34}$$

where $\partial_1 \partial_2 \phi_x(t) := \nabla_{t_1} \nabla_{t_2} \phi_x(h(t_1, t_2)) \in \mathbb{R}^{M_1 \times M_2}$ and $\phi''_\alpha(u) := \nabla_u^2 \phi_\alpha(u)$, does not depend on the noise vectors ε_1 and ε_2.

For simplicity, we first prove the identifiability result when all components of the common sources are non-Gaussian. The high level idea of the proof is as follows. We assume two different representations of x_1 and x_2 and using (34) and the independence of the components of α and the noises, we first show that the two potential dictionaries are related by an orthogonal matrix (and not any invertible matrix), and then show that this implies that the two potential sets of independent components are (orthogonal) linear combinations of each other, which, for non-Gaussian components which are not reduced to point masses, imposes that this orthogonal transformation is the combination of a permutation matrix and marginal scaling—a standard result from the ICA literature (Comon, 1994, Theorem 11).

Let us then assume that two equivalent representations of non-Gaussian CCA exist:

$$\begin{aligned} x_1 &= D_1 \alpha + \varepsilon_1 = E_1 \beta + \eta_1, \\ x_2 &= D_2 \alpha + \varepsilon_2 = E_2 \beta + \eta_2, \end{aligned} \tag{35}$$

where the other sources $\beta = (\beta_1, \dots, \beta_K)$ are also assumed mutually independent and non-degenerate. As a standard practice in the ICA literature and without loss of generality as the sources have non-degenerate components, one can assume that the sources have unit variances, i.e. $\mathrm{cov}(\alpha, \alpha) = I$ and $\mathrm{cov}(\beta, \beta) = I$, by respectively rescaling the columns of the factor loading matrices. Under this assumption, the two expressions of the cross-covariance matrix are

$$\mathrm{cov}(x_1, x_2) = D_1 D_2^\top = E_1 E_2^\top, \tag{36}$$

which, given that D_1, D_2 have full rank, implies that[5]

$$E_1 = D_1 Q, \quad E_2 = D_2 Q^{-\top}, \tag{37}$$

where $Q \in \mathbb{R}^{K \times K}$ is some invertible matrix. Substituting the representations (35) into the blockwise mixed derivatives of the SCF (34) and using the expressions (37) give

$$\begin{aligned} &D_1 \phi''_\alpha(D_1^\top t_1 + D_2^\top t_2) D_2^\top \\ &\qquad = D_1 Q \phi''_\beta(Q^\top D_1^\top t_1 + Q^{-1} D_2^\top t_2) Q^{-1} D_2^\top, \end{aligned}$$

for all $t_1 \in \mathbb{R}^{M_1}$ and $t_2 \in \mathbb{R}^{M_2}$. Since the matrices D_1 and D_2 have full rank, this can be rewritten as

$$\begin{aligned} &\phi''_\alpha(D_1^\top t_1 + D_2^\top t_2) \\ &\qquad = Q \phi''_\beta(Q^\top D_1^\top t_1 + Q^{-1} D_2^\top t_2) Q^{-1}, \end{aligned}$$

which holds for all $t_1 \in \mathbb{R}^{M_1}$ and $t_2 \in \mathbb{R}^{M_2}$. Moreover, still since D_1 and D_2 have full rank, we have, for any $u_1, u_2 \in \mathbb{R}^K$ the existence of $t_1 \in \mathbb{R}^{M_1}$ and $t_2 \in \mathbb{R}^{M_2}$, such that $u_1 = D_1^\top t_1$ and $u_2 = D_2^\top t_2$, that is,

$$\phi''_\alpha(u_1 + u_2) = Q \phi''_\beta(Q^\top u_1 + Q^{-1} u_2) Q^{-1}, \tag{38}$$

for all $u_1, u_2 \in \mathbb{R}^K$.

We will now prove two facts:

(F1) For any vector $v \in \mathbb{R}^K$, then $\phi''_\beta((Q^\top Q - I)v) = -I$, which will imply that $QQ^\top = I$ because of the non-Gaussian assumptions.

(F2) If $QQ^\top = I$, then $\phi''_\alpha(u) = \phi''_{Q\beta}(u)$ for any $u \in \mathbb{R}^K$, which will imply that Q is the composition of a permutation and a scaling. This will end the proof.

Proof of fact (F1). By letting $u_1 = Qv$ and $u_2 = -Qv$, we get:

$$\phi''_\alpha(0) = Q \phi''_\beta((Q^\top Q - I)v) Q^{-1}, \tag{39}$$

Since[6] $\phi''_\alpha(0) = -\mathrm{cov}(\alpha) = -I$, one gets

$$\phi''_\beta((Q^\top Q - I)v) = -I,$$

for any $v \in \mathbb{R}^K$.

[5]The fact that D_1, D_2 have full rank and that E_1, E_2 have K columns, combined with (36), implies that E_1, E_2 have also full rank.

[6] Note that $\nabla_u^2 \phi_\alpha(u) = -\frac{\mathbb{E}(\alpha\alpha^\top e^{iu^\top \alpha})}{\mathbb{E}(e^{iu^\top \alpha})} + \mathcal{E}_\alpha(u)\mathcal{E}_\alpha(u)^\top$, where $\mathcal{E}_\alpha(u) = \frac{\mathbb{E}(\alpha e^{iu^\top \alpha})}{\mathbb{E}(e^{iu^\top \alpha})}$.

Using the property that $\phi''_{A^\top\beta}(v) = A^\top \phi''_\beta(Av)A$ for any matrix A, and in particular with $A = Q^\top Q - I$, we have that $\phi''_{A^\top\beta}(v) = -A^\top A$, i.e. is constant.

If the second derivative of a function is constant, the function is quadratic. Therefore, $\phi_{A^\top\beta}(\cdot)$ is a quadratic function. Since the SCF completely defines the distribution of its variable (see,e.g., Jacod & Protter (2004)), $A^\top\beta$ must be Gaussian (the SCF of a Gaussian random variable is a quadratic function). Given Lemma 9 from Comon (1994) (i.e., Cramer's lemma: a linear combination of non-Gaussian random variables cannot be Gaussian unless the coefficients are all zero), this implies that $A = 0$, and hence $Q^\top Q = I$, i.e., Q is an orthogonal matrix.

Proof of fact (F2). Plugging $Q^\top = Q^{-1}$ into (38), with $u_1 = 0$ and $u_2 = u$, gives

$$\phi''_\alpha(u) = Q\phi''_\beta(Q^\top u)Q^\top = \phi''_{Q\beta}(u), \tag{40}$$

for any $u \in \mathbb{R}^K$. By integrating both sides of (40) and using $\phi_\alpha(0) = \phi_{Q\beta}(0) = 0$, we get that $\phi_\alpha(u) = \phi_{Q\beta}(u) + i\gamma^\top u$ for all $u \in \mathbb{R}^K$ for some constant vector γ. Using again that the SCF completely defines the distribution, it follows that $\alpha - \gamma$ and $Q\beta$ have the same distribution. Since both α and β have independent components, this is only possible when $Q = \Lambda P$, where P is a permutation matrix and Λ is some diagonal matrix (Comon, 1994, Theorem 11).

A.2 Case of a single Gaussian source

Without loss of generality, we assume that the potential Gaussian source is the first one for α and β. The first change is in the proof of fact (F1). We use the same argument up to the point where we conclude that $A^\top\beta$ is a Gaussian vector. As only β_1 can be Gaussian, Cramer's lemma implies that only the first row of A can have non-zero components, that is $A = Q^\top Q - I = e_1 f^\top$, where e_1 is the first basis vector and f any vector. Since $Q^\top Q$ is symmetric, we must have

$$Q^\top Q = I + a e_1 e_1^\top,$$

where a is a constant scalar different than -1 as $Q^\top Q$ is invertible. This implies that $Q^\top Q$ is an invertible diagonal matrix Λ, and hence $Q\Lambda^{-1/2}$ is an orthogonal matrix, which in turn implies that $Q^{-1} = \Lambda^{-1}Q^\top$.

Plugging this into (38) gives, for any u_1 and u_2:

$$\phi''_\alpha(u_1 + u_2) = Q\phi''_\beta(Q^\top u_1 + \Lambda^{-1}Q^\top u_2)\Lambda^{-1}Q^\top.$$

Given that diagonal matrices commute and that ϕ''_β is diagonal for independent sources (see Appendix B.1), this leads to

$$\phi''_\alpha(u_1 + u_2) = Q\Lambda^{-1/2}\phi''_\beta(Q^\top u_1 + \Lambda^{-1}Q^\top u_2)\Lambda^{-1/2}Q^\top.$$

For any given $v \in \mathbb{R}^K$, we are looking for u_1 and u_2 such that $Q^\top u_1 + \Lambda^{-1}Q^\top u_2 = \Lambda^{-1/2}Q^\top v$ and $u_1 + u_2 = v$, which is always possible by setting $Q^\top u_2 = (\Lambda^{-1/2} + I)^{-1}Q^\top v$ and $Q^\top u_1 = Q^\top v - Q^\top u_2$ by using the special structure of Λ. Thus, for any v,

$$\phi''_\alpha(v) = Q\Lambda^{-1/2}\phi''_\beta(\Lambda^{-1/2}Q^\top v)\Lambda^{-1/2}Q^\top = \phi''_{Q\Lambda^{-1/2}\beta}(v).$$

Integrating as previously, this implies that the characteristic function of α and $Q\Lambda^{-1/2}\beta$ differ only by a linear function $i\gamma^\top v$, and thus, that $\alpha - \gamma$ and $Q\Lambda^{-1/2}\beta$ have the same distribution. This in turn, from Comon (1994, Theorem 11), implies that $Q\Lambda^{-1/2}$ is a product of a scaling and a permutation, which ends the proof.

A.3 Identifiability of discrete CCA (5) and mixed CCA (6)

Given the discrete CCA model, the SCF $\phi_x(t)$ takes the form

$$\begin{aligned}\phi_x(t) = \phi_\alpha(D_1^\top(e^{it_1} - 1) + D_2^\top(e^{it_2} - 1))\\ + \phi_{\varepsilon_1}(e^{it_1} - 1) + \phi_{\varepsilon_2}(e^{it_2} - 1),\end{aligned}$$

where e^{it_j}, for $j = 1, 2$, denotes a vector with the m-th element equal to $e^{i[t_j]_m}$, and we used the arguments analogous with the non-Gaussian case. The rest of the proof extends with a correction that sometimes one has to replace D_j with $\mathrm{diag}[e^{it_j}]D_j$ and that $u_j = D_j^\top(e^{it_j} - 1)$ for $j = 1, 2$. For the mixed CCA case, only the part related to x_2 and D_2 changes in the same way as for the discrete CCA case.

B The generalized expectation and covariance matrix

B.1 The generalized expectation and covariance matrix of the sources

The sources $\alpha = (\alpha_1, \ldots, \alpha_K)$ are mutually independent. Therefore, for some $h \in \mathbb{R}^K$, their CGF (19) $K_\alpha(h) = \log \mathbb{E}(e^{\alpha^\top h})$ takes the form

$$K_\alpha(h) = \sum\nolimits_k \log\left[\mathbb{E}(e^{\alpha_k h_k})\right].$$

Therefore, the k-th element of the generalized expectation (20) of α is (separable in α_k)

$$[\mathcal{E}_\alpha(h)]_k = \frac{\mathbb{E}(\alpha_k e^{\alpha_k h_k})}{\mathbb{E}(e^{\alpha_k h_k})} \tag{41}$$

and the generalized covariance (21) of α is diagonal due to the separability and its k-th diagonal element is

$$[\mathcal{C}_\alpha(h)]_{kk} = \frac{\mathbb{E}(\alpha_k^2 e^{\alpha_k h_k})}{\mathbb{E}(e^{\alpha_k h_k})} - [\mathcal{E}_\alpha(h)]_k^2. \tag{42}$$

B.2 Some expectations of a Poisson random variable

Let $x \in \mathbb{R}^M$ be a multivariate Poisson random variable with mean $y \in \mathbb{R}_+^M$. Then, for some $t \in \mathbb{R}^M$,

$$\begin{aligned}
\mathbb{E}(e^{t^\top x}) &= e^{y^\top(e^t - 1)}, \\
\mathbb{E}(x_m e^{t^\top x}) &= y_m e^{t_m} e^{y^\top(e^t-1)}, \\
\mathbb{E}(x_m^2 e^{t^\top x}) &= \left[y_m e^{t_m} + 1\right] y_m e^{t_m} e^{y^\top(e^t-1)}, \\
\mathbb{E}(x_m x_{m'} e^{t^\top x}) &= y_m e^{t_m} y_{m'} e^{t_{m'}} e^{y^\top(e^t-1)}, \quad m \neq m',
\end{aligned}$$

where e^t denotes an M-vector with the m-th element equal to e^{t_m}.

B.3 The generalized expectation and covariance matrix of discrete ICA

In this section, we use the expectations of a Poisson random variable presented in Appendix B.2.

Given the discrete ICA model (9), the generalized expectation (20) of $x \in \mathbb{R}^M$ takes the form

$$\begin{aligned}
\mathcal{E}_x(t) = \frac{\mathbb{E}(x e^{t^\top x})}{\mathbb{E}(e^{t^\top x})} &= \frac{\mathbb{E}\left[\mathbb{E}(x e^{t^\top x}|\alpha)\right]}{\mathbb{E}\left[\mathbb{E}(e^{t^\top x}|\alpha)\right]} \\
&= \operatorname{diag}[e^t] D \frac{\mathbb{E}(\alpha e^{\alpha^\top h(t)})}{\mathbb{E}(e^{\alpha^\top h(t)})} \\
&= \operatorname{diag}[e^t] D \mathcal{E}_\alpha(h(t)),
\end{aligned}$$

where $t \in \mathbb{R}^M$ is a parameter, $h(t) = D^\top(e^t - 1)$, and e^t denotes an M-vector with the m-th element equal to e^{t_m}. Note that in the last equation we used the definition (20) of the generalized expectation $\mathcal{E}_\alpha(\cdot)$.

Further, the generalized covariance (21) of x takes the form

$$\begin{aligned}
\mathcal{C}_x(t) &= \frac{\mathbb{E}(x x^\top e^{t^\top x})}{\mathbb{E}(e^{t^\top x})} - \mathcal{E}_x(t)\mathcal{E}_x(t)^\top \\
&= \frac{\mathbb{E}\left[\mathbb{E}(x x^\top e^{t^\top x}|\alpha)\right]}{\mathbb{E}\left[\mathbb{E}(e^{t^\top x}|\alpha)\right]} - \mathcal{E}_x(t)\mathcal{E}_x(t)^\top.
\end{aligned}$$

Plugging into this expression the expression for $\mathcal{E}_x(t)$ and

$$\begin{aligned}
\mathbb{E}(x x^\top e^{t^\top x}|\alpha) = \operatorname{diag}[e^t] D \mathbb{E}(\alpha\alpha^\top e^{\alpha^\top h(t)}) D^\top \operatorname{diag}[e^t] \\
+ \operatorname{diag}[e^t] \operatorname{diag}\left[D \mathbb{E}(\alpha e^{\alpha^\top h(t)})\right]
\end{aligned}$$

we get

$$\mathcal{C}_x(t) = \operatorname{diag}[\mathcal{E}_x(t)] + \operatorname{diag}[e^t] D \mathcal{C}_\alpha(h(t)) D^\top \operatorname{diag}[e^t],$$

where we used the definition (21) of the generalized covariance of α.

B.4 The generalized CCA S-covariance matrix

In this section we sketch the derivation of the diagonal form (27) of the generalized S-covariance matrix of mixed CCA (6). Expressions (25) and (26) can be obtained in a similar way.

Denoting $x = [x_1;\ x_2]$ and $t = [t_1;\ t_2]$ (i.e. stacking the vectors as in (8)), the CGF (19) of mixed CCA (6) can be written as

$$\begin{aligned} K_x(t) &= \log \mathbb{E}(e^{t_1^\top x_1 + t_2^\top x_2}) \\ &= \log \mathbb{E}\left[\mathbb{E}(e^{t_1^\top x_1 + t_2^\top x_2} \,|\, \alpha, \varepsilon_1, \varepsilon_2)\right] \\ &\overset{(a)}{=} \log \mathbb{E}\left[\mathbb{E}(e^{t_1^\top x_1} \,|\, \alpha, \varepsilon_1)\mathbb{E}(e^{t_2^\top x_2} \,|\, \alpha, \varepsilon_2)\right] \\ &\overset{(b)}{=} \log \mathbb{E}\left(e^{t_1^\top (D_1\alpha + \varepsilon_1)} e^{(D_2\alpha + \varepsilon_2)^\top (e^{t_2} - 1)}\right) \\ &\overset{(c)}{=} \log \mathbb{E}\left(e^{\alpha^\top h(t)}\right) + \log \mathbb{E}\left(e^{\varepsilon_2^\top (e^{t_2} - 1)}\right) + \log \mathbb{E}(e^{t_1^\top \varepsilon_1}), \end{aligned}$$

where $h(t) = (D_1^\top t_1 + D_2^\top (e^{t_2} - 1)$, in (a) we used the conditional independence of x_1 and x_2, in (b) we used the first expression from Appendix B.2, and in (c) we used the independence assumption (3).

The generalized CCA S-covariance matrix is defined as

$$S_{12}(t) := \nabla_{t_2} \nabla_{t_1} K_x(t).$$

Its gradient with respect to t_1 is

$$\nabla_{t_1} K_x(t) = \frac{D_1 \mathbb{E}(\alpha e^{\alpha^\top h(t)})}{\mathbb{E}(e^{\alpha^\top h(t)})} + \frac{\mathbb{E}(\varepsilon_1 e^{t_1^\top \varepsilon_1})}{\mathbb{E}(e^{t_1^\top \varepsilon_1})},$$

where the last term does not depend on t_2. Computing the gradient of this expression with respect to t_2 gives

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

where we substituted expression (42) for the generalized covariance of the independent sources.

B.5 Approximation of the T-cumulants with the generalized covariance matrix

Let $f_{mm'}(t) = [\mathcal{C}_x(t)]_{mm'}$ be a function $\mathbb{R} \to \mathbb{R}^M$ corresponding to the (m, m')-th element of the generalized covariance matrix. Then the following holds for its directional derivative at t_0 along the direction t:

$$\langle \nabla f_{mm'}(t_0), t \rangle = \lim_{\delta \to 0} \frac{f_{mm'}(t_0 + \delta t) - f_{mm'}(t_0)}{\delta},$$

where $\langle \cdot, \cdot \rangle$ stands for the inner product. Therefore, when using the fact that $\nabla f(t_0) = \nabla \mathcal{C}_x(t)$ is the generalized cumulant of x at t_0 and the definition of a projection of a tensor onto a vector (28), one obtains for $t_0 = 0$ the approximation of the cumulant $\mathrm{cum}(x)$ with the generalized covariance matrix $\mathcal{C}_x(t)$.

Let us define $v_1 = W_1^\top u_1$ and $v_1 = W_2^\top u_2$ for some $u_1,\, u_2 \in \mathbb{R}^K$. Then, approximations for the T-cumulants (17) of discrete CCA take the following form: $W_1 T_{121}(v_1) W_2$ is approximated by the generalized S-covariances (24) $S_{12}(t)$ via the following expression

$$W_1 T_{121}(v_1) W_2 \approx \frac{W_1 S_{12}(\delta t_1) W_2^\top - W_1 S_{12}(0) W_2^\top}{\delta} - W_1 \mathrm{diag}(v_1) S_{12} W_2^\top,$$

where $t_1 = \begin{pmatrix} v_1 \\ 0 \end{pmatrix}$ and $W_1 T_{122}(v_2) W_2$ is approximated by the generalized S-covariances $S_{12}(t)$ via

$$W_1 T_{122}(v_2) W_2 \approx \frac{W_1 S_{12}(\delta t_2) W_2^\top - W_1 S_{12}(0) W_2^\top}{\delta} - W_1 S_{12} \mathrm{diag}(v_2) W_2^\top,$$

where $t_2 = \begin{pmatrix} 0 \\ v_2 \end{pmatrix}$ and δ are chosen to be small.

C Finite sample estimators

C.1 Finite sample estimators of the generalized expectation and covariance matrix

Following Yeredor (2000); Slapak & Yeredor (2012b), we use the most direct way of defining the finite sample estimators of the generalized expectation (20) and covariance matrix (21).

Given a finite sample $X = \{x_1, x_2, \ldots, x_N\}$, an estimator of the generalized expectation is

$$\widehat{\mathcal{E}}_x(t) = \frac{\sum_n x_n w_n}{\sum_n w_n}$$

where weights $w_n = e^{t^\top x_n}$ and an estimator of the generalized covariance is

$$\widehat{\mathcal{C}}_x(t) = \frac{\sum_n x_n x_n^\top w_n}{\sum_n w_n} - \widehat{\mathcal{E}}_x(t)\widehat{\mathcal{E}}_x(t)^\top.$$

Similarly, an estimator of the generalized S-covariance matrix is then

$$\widehat{\mathcal{C}}_{x_1,x_2}(t) = \frac{\sum_n x_{1n} x_{2n}^\top w_n}{\sum_n w_n} - \frac{\sum_n x_{1n} w_n}{\sum_n w_n} \frac{\sum_n x_{2n}^\top w_n}{\sum_n w_n},$$

where $x = [x_1;\ x_2]$ and $t = [t_1;\ t_2]$ for some $t_1 \in \mathbb{R}^{M_1}$ and $t_2 \in \mathbb{R}^{M_2}$.

Some properties of these estimators are analyzed by Slapak & Yeredor (2012b).

C.2 Finite sample estimators of the DCCA cumulants

In this section, we sketch the derivation of unbiased finite sample estimators for the CCA cumulants S_{12}, T_{121}, and T_{122}. Since the derivation is nearly identical to the derivation of the estimators for the DICA cumulants (see Appendix F.2 of Podosinnikova et al. (2015)), all details are omitted.

Given a finite sample $X_1 = \{x_{11}, x_{12}, \ldots, x_{1N}\}$ and $X_2 = \{x_{21}, x_{22}, \ldots, x_{2N}\}$, the finite sample estimator of the discrete CCA S-covariance (15), i.e., $S_{12} := \mathrm{cum}(x_1, x_2)$, takes the form

$$\widehat{S}_{12} = \eta_1 \left[X_1 X_2^\top - N \widehat{\mathbb{E}}(x_1) \widehat{\mathbb{E}}(x_2)^\top \right], \tag{43}$$

where $\widehat{\mathbb{E}}(x_1) = N^{-1} \sum_n x_{1n}$, $\widehat{\mathbb{E}}(x_2) = N^{-1} \sum_n x_{2n}$, and $\eta_1 = 1/(N-1)$.

Substitution of the finite sample estimators of the 2nd and 3rd cumulants (see, e.g., Appendix C.4 of Podosinnikova et al. (2015)) into the definition of the DCCA T-cumulants (17) leads to the following expressions

$$\begin{aligned}
\widehat{W}_1 \widehat{T}_{12j}(v_j) \widehat{W}_2^\top = {} & \eta_2 [(\widehat{W}_1 X_1) \mathrm{diag}(X_j^\top v_j)] \otimes (\widehat{W}_2 X_2) \\
& + \eta_2 \langle v_j, \widehat{\mathbb{E}}(x_j) \rangle 2N [\widehat{W}_1 \widehat{\mathbb{E}}(x_1)] \otimes [\widehat{W}_2 \widehat{\mathbb{E}}(x_2)] \\
& - \eta_2 \langle v_j, \widehat{\mathbb{E}}(x_j) \rangle (\widehat{W}_1 X_1) \otimes (\widehat{W}_2 X_2) \\
& - \eta_2 [(\widehat{W}_1 X_1)(X_j^\top v_j)] \otimes [\widehat{W}_2 \widehat{\mathbb{E}}(x_2)] \\
& - \eta_2 [\widehat{W}_1 \widehat{\mathbb{E}}(x_1)] \otimes [(\widehat{W}_2 X_2)(X_j^\top v_j)] \\
& - \eta_1 (\widehat{W}_1^{(j)} X_1) \otimes (\widehat{W}_2^{(j)} X_2) \\
& + \eta_1 N [\widehat{W}_1^{(j)} \widehat{\mathbb{E}}(x_1)] \otimes [\widehat{W}_2^{(j)} \widehat{\mathbb{E}}(x_2)],
\end{aligned}$$

where $\eta_2 = N/((N-1)(N-2))$ and $\widehat{W}_1^{(1)} = \widehat{W}_1 \mathrm{diag}(v_1)$, $\widehat{W}_2^{(1)} = \widehat{W}_2$, $\widehat{W}_1^{(2)} = \widehat{W}_1$, and $\widehat{W}_2^{(2)} = \widehat{W}_2 \mathrm{diag}(v_2)$.

In the expressions above, $\widehat{W}_1$ and $\widehat{W}_2$ denote whitening matrices of $\widehat{S}_{12}$, i.e. such that $\widehat{W}_1 \widehat{S}_{12} \widehat{W}_2^\top = I$.

D Implementation details

D.1 Computation of whitening matrices

One can compute such whitening matrices (31) via the singular value decomposition (SVD) of S_{12}. Let $S_{12} = U\Sigma V^\top$ be the SVD of S_{12}, then one can define $W_1 = U_{1:K}\Lambda$ and $W_2 = V_{1:K}\Lambda$, where $U_{1:K}$ and $V_{1:K}$ are the first K left- and right-singular vectors and $\Lambda = \mathrm{diag}(\sigma_1^{-1/2}, \dots, \sigma_K^{-1/2})$ and $\sigma_1, \dots, \sigma_K$ are the K largest singular values.

Although SVD is computed only once, the size of the matrix S_{12} can be significant even for storage. To avoid construction of this large matrix and speed up SVD, one can use randomized SVD techniques (Halko et al., 2011). Indeed, since the sample estimator $\widehat{S}_{12}$ has the form (43), one can reduce this matrix by sampling two Gaussian random matrices $\Omega_1 \in \mathbb{R}^{\tilde{K}\times M_1}$ and $\Omega_2 \in \mathbb{R}^{\tilde{K}\times M_2}$, where $\tilde{K}$ is slightly larger than K. Now, if U and V are the K largest singular vectors of the reduced matrix $\Omega_1\widehat{S}_{12}\Omega_2$, then $\Omega_1^\dagger U$ and $\Omega_2^\dagger V$ are approximately (and up to permutation and scaling of the columns) the K largest singular vectors of $\widehat{S}_{12}$.

D.2 Applying whitening transform to DCCA T-cumulants

Transformation of the T-cumulants (29) with whitening matrices W_1 and W_2 gives new tensors $\widehat{T}_{12j} \in \mathbb{R}^{K\times K\times K}$:

$$\widehat{T}_{12j} := T_{12j} \times_1 W_1^\top \times_2 W_2^\top \times_3 W_j^\top, \tag{44}$$

where $j = 1, 2$. Combining this transformation with the projection (28), one obtains $2P+1$ matrices

$$W_1 S_{12} W_2^\top,\ W_1 T_{12j}(W_j^\top u_{jp}) W_2^\top, \tag{45}$$

where $p = 1, \dots, P$ and $j = 1, 2$ and we used $v_{jp} = W_j^\top u_{jp}$ to take into account whitening along the third direction. By choosing $u_{jp} \in \mathbb{R}^K$ to be the canonical vectors of the R^K, the number of tensor projections is reduced from $M = M_1 + M_2$ to $2K$.

D.3 Choice of projection vectors or processing points

For the T-cumulants (29), we choose the K projection vectors as $v_{1p} = W_1^\top e_p$ and $v_{2p} = W_2^\top e_p$, where e_p is one of the columns of the K-identity matrix (i.e., a canonical vector). For the generalized S-covariances (30), we choose the processing points as $t_{1p} = \delta_1 v_{1p}$ and $t_{2p} = \delta_2 v_{2p}$, where δ_j, for $j = 1, 2$ are set to a small value such as 0.1 divided by $\sum_m \mathbb{E}(|x_{jm}|)/M_j$, for $j = 1, 2$.

When projecting a tensor T_{12j} onto a vector, part of the information contained in this tensor gets lost. To preserve all information, one could project a tensor T_{12j} onto the canonical basis of $\mathbb{R}^{M_j}$ to obtain M_j matrices. However, this would be an expensive operation in terms of both memory and computational time. In practice, we use the fact, that the tensor T_{12j}, for $J = 1, 2$, is transformed with whitening matrices (44). Hence, the projection vector has to include multiplication by the whitening matrices. Since they reduce the dimension to K, choosing the canonical basis in $\mathbb{R}^K$ becomes sufficient. Hence, the choice $v_{1p} = W_1^\top e_p$ and $v_{2p} = W_2^\top e_p$, where e_p is one of the columns of the K-identity matrix.

Importantly, in practice, the tensors are never constructed (see Appendix C.2).

The choice of the processing points of the generalized covariance matrices has to be done carefully. Indeed, if the values of t_1 or t_2 are too large, the exponents blow up. Hence, it is reasonable to maintain the values of the processing points very small. Therefore, for $j = 1, 2$, we set $t_{jp} = \delta_j v_{jp}$ where δ_j is proportional to a parameter δ which is set to a small value ($\delta = 0.1$ by default), and the scale is determined by the inverse of the empirical average of the component of x_j, i.e.:

$$\delta_j := \delta \frac{N M_j}{\sum_{n=1}^N \sum_{m=1}^{M_j} [|X_j|]_{mn}}, \tag{46}$$

for $j = 1, 2$. See Appendix F.2 for an experimental comparison of different values of δ (the default value used in other experiments is $\delta = 0.1$).

D.4 Finalizing estimation of D_1 and D_2

The non-orthogonal joint diagonalization algorithm outputs an invertible matrix Q. If the estimated factor loading matrices are not supposed to be non-negative (continuous case of NCCA (4)), then

$$\begin{aligned} D_1 &= W_1^{\dagger} Q, \\ D_2 &= W_2^{\dagger} Q^{-1}, \end{aligned} \tag{47}$$

where $\dagger$ stands for the pseudo-inverse. For the spectral algorithm, where Q are eigenvectors of a non-symmetric matrix and are not guaranteed to be real, only real parts are kept after evaluating matrices D_1 and D_2 in accordance with (47).

If the matrices D_1 and/or D_2 have to be non-negative (the discrete case of DCCA (5) and MCCA (6)), they have to be further mapped. For that, we select the sign of each column such that the vector (column) has less negative than positive components, which is measured by the sum of squares of the components of each sign, (this is necessary since the scaling unidentifiability includes the scaling by -1) and then truncate all negative values at 0.

In practice, due to the scaling unidentifiability, each column of the obtained matrices D_1 and D_2 can be further normalized to have the unit ℓ_1-norm. This is applicable in all cases (D/M/NCCA).

E Jacobi-like joint diagonalization of non-symmetric matrices

Given N non-defective (a.k.a. diagonalizable) not necessary *normal*[7] matrices

$$\mathcal{A} = \{A_1, A_2, \ldots, A_N\},$$

where each matrix $A_n \in \mathbb{R}^{M \times M}$, find such matrix $Q \in \mathbb{R}^{M \times M}$ that matrices

$$Q^{-1}\mathcal{A}Q = \left\{Q^{-1}A_1Q, Q^{-1}A_2Q, \ldots, Q^{-1}A_NQ\right\}$$

are (jointly) as diagonal as possible. We refer to this problem as a non-orthogonal joint diagonalization (NOJD) problem.[8]

Algorithm 1 Non-orthogonal joint diagonalization (NOJD)

```
1:  Initialize: A^(0) ← A and Q^(0) ← I_M and iterations ℓ = 0
2:  for sweeps k = 1, 2, ... do
3:    for p = 1, ..., M − 1 do
4:      for q = p + 1, ..., M do
5:        Increase ℓ = ℓ + 1
6:        Find the (approx.) shear parameter y* defined in (54)
7:        Find the Jacobi angle θ* defined in (53)
8:        Update Q^(ℓ) ← Q^(ℓ−1) S*^(ℓ) U*^(ℓ)
9:        Update A^(ℓ) ← U*^(ℓ)⊤ S*^(ℓ)−1 A^(ℓ−1) S*^(ℓ) U*^(ℓ)
10:     end for
11:   end for
12: end for
13: Output: Q^(ℓ)
```

Algorithm. Non-orthogonal Jacobi-like joint diagonalization algorithms have the high level structure which is outlined in Alg. 1.

The algorithm iteratively constructs the sequence of matrices $\mathcal{A}^{(\ell)} = \left\{A_1^{(\ell)}, A_2^{(\ell)}, \ldots, A_N^{(\ell)}\right\}$, which is initialized with $\mathcal{A}^{(0)} = \mathcal{A}$. Each such iteration ℓ corresponds to a single update (Line (9) of Alg. (1)) of the matrices with the optimal shear $S_*^{(\ell)}$ and unitary $U_*^{(\ell)}$ transforms:

$$A_n^{(\ell)} = U_*^{(\ell)\top} S_*^{(\ell)-1} A_n^{(\ell-1)} S_*^{(\ell)} U_*^{(\ell)},$$

[7] A real matrix A is normal if $A^\top A = AA^\top$.

[8] An orthogonal joint diagonalization problem corresponds to the case where the matrices $A_1, A_2, \ldots, A_N$ are normal and, hence, diagonalizable by an orthogonal matrix Q.

where $S_*^{(\ell)} = S^{(\ell)}(y^*)$ and $U_*^{(\ell)} = U^{(\ell)}(\theta^*)$ for the chosen in accordance with some rules (see below) optimal shear parameter y^* and optimal Jacobi (=Givens) angle θ^*.

For the theoretical analysis purposes, the two transforms are considered separately:

$$\begin{aligned} A_n'^{(\ell)} &= S^{(\ell)-1}(y) A_n^{(\ell-1)} S^{(\ell)}(y), \\ A_n^{(\ell)} &= A_n''^{(\ell)} = U^{(\ell)\top}(\theta) A_n'^{(\ell)} U^{(\ell)}(\theta). \end{aligned} \tag{48}$$

Each such iteration ℓ is a combination of the iteration k and the pivots p and q (see Alg. 1). The iteration k is referred to as a *sweep*. Within each sweep k, $M(M-1)/2$ pivots $p < q$ are chosen in accordance with the lexicographical rule. The rule for the choice of pivots can affect convergence as was analyzed for the single matrix case (see, e.g., Ruhe, 1968; Eberlein, 1962), where more sophisticated rules were proposed for the algorithm to have a quadratic convergence phase. However, up to our best knowledge, no such analysis was done for the several matrices case. We assume the simple lexicographical rule all over the paper.

The *shear transform* is defined by the hyperbolic rotation matrix $S^{(\ell)} = S^{(\ell)}(y)$ which is equal to the identity matrix except for the following entries

$$\begin{pmatrix} S_{pp}^{(\ell)} & S_{pq}^{(\ell)} \\ S_{qp}^{(\ell)} & S_{qp}^{(\ell)} \end{pmatrix} = \begin{pmatrix} \cosh y & \sinh y \\ \sinh y & \cosh y \end{pmatrix}, \tag{49}$$

where the *shear parameter* $y \in \mathbb{R}$. The *unitary transform* is defined by the Jacobi (=Givens) rotation matrix $U^{(\ell)} = U^{(\ell)}(\theta)$ which is equal to the identity matrix except for the following entries

$$\begin{pmatrix} U_{pp}^{(\ell)} & U_{pq}^{(\ell)} \\ U_{qp}^{(\ell)} & U_{qp}^{(\ell)} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{50}$$

where the *Jacobi (=Givens)* angle $\theta \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$.

The following two objective functions are of the central importance for this type of algorithms: (a) the sum of squares of all the off-diagonal elements of the matrices[9] $\mathcal{A}''^{(\ell)}$ which are the transformed with the unitary transform $U^{(\ell)}$ matrices $\mathcal{A}'^{(\ell)}$:

$$\mathrm{Off}\left(\mathcal{A}''^{(\ell)}\right) = \sum_{n=1}^{N} \mathrm{Off}\left(U^{(\ell)\top} A_n'^{(\ell)} U^{(\ell)}\right) \tag{51}$$

and (b) the sum of the squared Frobenius norms of the matrices $\mathcal{A}'^{(\ell)}$ which are the transformed with the share transform $S^{(\ell)}$ matrices $\mathcal{A}^{(\ell-1)}$:

$$\left\|\mathcal{A}'^{(\ell)}\right\|_F^2 = \sum_{n=1}^{N} \left\|S^{(\ell)-1} A_n^{(\ell-1)} S^{(\ell)}\right\|_F^2. \tag{52}$$

We refer to (51) as the *diagonality measure* and to (52) as the *normality measure*.

All the considered algorithms find the optimal Jacobi angle θ^* as the minimizer of the diagonality measure of the (unitary transformed) matrices $\mathcal{A}''^{(\ell)}$ (48):

$$\theta^* = \underset{\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]}{\arg\min}\ \mathrm{Off}\left(\mathcal{A}''^{(\ell)}\right), \tag{53}$$

which admits a unique closed form solution (Cardoso & Souloumiac, 1996). The optimal shear parameter y^* is found[10] as a minimizer of the normality measure of the (shear transformed) matrices $\mathcal{A}'^{(\ell)}$ (48):

$$y^* = \underset{y \in \mathbb{R}}{\arg\min} \left\|\mathcal{A}'^{(\ell)}\right\|_F^2. \tag{54}$$

All the considered algorithms (Fu & Gao, 2006; Iferroudjene et al., 2009; Luciani & Albera, 2010) solve this step only approximately. In particular, the sh-rt algorithm (Fu & Gao, 2006) approximates the equation for finding the nulls of the gradient of the objective; the JUST algorithm (Iferroudjene et al., 2009) replaces the normality measure with the diagonality measure and provides a closed form solution for the resulting problem; and the JDTM algorithm (Luciani & Albera, 2010) replaces the normality measure with the sum of only two squared elements $A'_{n,pq}$ and $A'_{n,qp}$ and provides a closed form solution for the resulting problem.

[9] In the JUST algorithm (Iferroudjene et al., 2009), this objective function is also considered for the (shear transformed) matrix $\mathcal{A}'^{(\ell)}$.

[10] The JUST algorithm is an exception here, since it minimizes the diagonality measure $\mathrm{Off}[\mathcal{A}'^{(\ell)}]$ of the (shear transformed) matrices $\mathcal{A}'^{(\ell)}$ with respect to y.

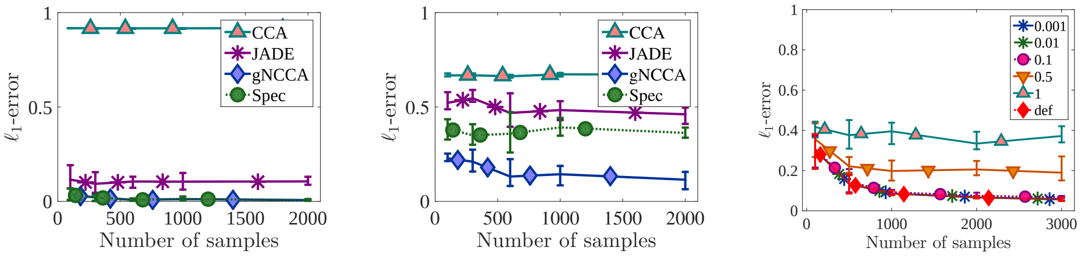

Figure 3: **(left and middle)** The continuous synthetic data experiment from Appendix F.1 with $M_1 = M_2 = 20$, $c = c_1 = c_2 = 0.1$ and $L_n = L_s = 1000$. The number of factors: **(left)** $K_1 = K_2 = K = 1$ and **(middle)** $K_1 = K_2 = K = 10$. **(right)**: An experimental analysis of the performance of DCCAg with generalized covariance matrices using different parameters δ_j for the processing points. The numbers in the legend correspond to the values of δ defining δ_j via (46) in Appendix D.3. The default value (def) is $\delta = 0.1$. The data is the discrete synthetic data as described in Section 5 with the parameters set as in Fig. 2 (right).

The three NOJD algorithms can have slightly different convergence properties, however, for the purposes of this paper their performance can hardly be distinguished. That is, the difference in the performance of the algorithms in terms of the ℓ_1-error of the factor loading matrices is hardly noticeable. For the experiments, we use the JDTM algorithm, the other two algorithms could be equally used. To the best of our knowledge, no theoretical analysis of the NOJD algorithms is available, except for the single matrix case when they boil down to the (non-symmetric) eigenproblem (Eberlein, 1962; Ruhe, 1968).

The following intuitively explains why the normality measure, i.e. the sum of the squared Frobenius norms, has to be minimized at the shear transform. As (Ruhe, 1968) mention, for every matrix A and non-singular Q:

$$\inf_Q \left\|Q^{-1}AQ\right\|_F^2 = \|\Lambda\|_F^2 ,$$

where Λ is the diagonal matrix containing the eigenvalues of A. Therefore, a diagonalized version of the matrix A must have the smallest Frobenius norm. Since the unitary transform does not change the Frobenius norm, it can only be minimized with the shear transform. Further, if a matrix is normal, i.e. $A^\top A = AA^\top$ with a symmetric matrix as a particular case, the upper triangular matrix in its Schur decomposition is zero (Golub & Van Loan, 1996, Chapter 7) and then the Schur vectors correspond to the (orthogonal in this case) eigenvectors of this matrix. Therefore, a normal non-defective matrix can be diagonalized by an orthogonal matrix, which preserves the Frobenius norm. Hence, the shear transform by minimizing the normality measure decreases the deviation from normality and then the unitary transform by minimizing the diagonality measure decreases the deviation from diagonality.

F Supplementary experiments

F.1 Continuous synthetic data

This experiment is essentially a continuous analogue to the synthetic experiment with the discrete data from Section 5.

Synthetic data. We sample synthetic data from the linear non-Gaussian CCA (NCCA) model (7) with each view $x_j = D_j\alpha + F_j\beta_j$. The (non-Gaussian) sources are $\alpha \sim z_\alpha \mathrm{Gamma}(c, b)$, where z_α is a Rademacher random variable (i.e., takes the values -1 or 1 with the equal probabilities). The noise sources are $\beta_j \sim z_{\beta_j}\mathrm{Gamma}(c_j, b_j)$, for $j = 1, 2$, where again z_{β_j} is a Rademacher random variable. Parameters of the gamma distribution are initialized by analogy with the discrete case (see Section 5). The elements of the matrices D_j and F_j, for $j = 1, 2$, are sampled i.i.d. for the uniform distribution in $[-1, 1]$. Each column of D_j and F_j, for $j = 1, 2$, is normalized to have the unit ℓ_1-norm.

Algorithms. We compare gNCCA (the implementation of NCCA with the generalized S-covariance matrices with the default values of the parameters δ_1 and δ_2 as described in Appendix D.3) the spectral algorithm for NCCA (also with the generalized S-covariance matrices) to the JADE algorithm[11] (Cardoso & Souloumiac, 1993) for independent component analysis (ICA) and to classical CCA.

Synthetic experiment. In Fig. 3 (left and middle), the results of the experiment for the different number of topics are presented. The error of the classical CCA is high due to the mentioned unidentifiability issues.

[11] The code is available at: http://perso.telecom-paristech.fr/ cardoso/Algo/Jade/jadeR.m

F.2 Sensitivity of the generalized covariance matrices to the choice of the processing points

In this section, we experimentally analyze the performance of the DCCAg algorithm based on the generalized S-covariance matrices vs. the parameters δ_1 and δ_2. We use the experimental setup of the synthetic discrete data from Section 5 with $K_1 = K_2 = K = 10$. The results are presented in Fig. 3 (right).

F.3 Real data experiment – translation topics

For the real data experiment, we estimate the factor loading matrices (topics, in the following) D_1 and D_2 of aligned proceedings of the 36-th Canadian Parliament in English and French languages. This Hansard collection can be found at http://www.isi.edu/natural-language/download/hansard/.

Although going into details of natural language processing (NLP) related problems is not the goal of this paper, we do minor pre-processing (see Appendix F.4) of this text data to improve the presentation of the estimated bilingual topics D_1 and D_2.

The 20 topics obtained with DCCA are presented in Tables 2–6. For each topic, we display the 20 most frequent words (ordered from top to bottom in the decreasing order). Most of the topic have quite clear interpretation. Moreover, we can often observe the pairs of words which are each others translations in the topics. Take, e.g.,

- the topic 10: the phrase “pension plan” can be translated as “régime de retraite”, the word “benefits” as “prestations”, and abbreviations “CPP” and “RPC” stand for “Canada Pension Plan” and “Régime de pensions du Canada”, respectively;
- the topic 3: “OTAN” is the French abbreviation for “NATO”, the word “war” is translated as “guerre”, and the word “peace” as “paix”;
- the topic 9: “Nisga” is the name of an Indigenous (or “aboriginal”) people in British Columbia, the word “aboriginal” translates to French as “autochtontes”, and, e.g., the word “right” can be translated as “droit”.

Note also that, e.g., in topics 10, although the French words “ans” and “années” are present in the French topic, their English translation “year” is not, since it was removed as one of the 15 most frequent words in English (see Appendix F.4).

F.4 Data preprocessing

For the experiment, we use House Debate Training Set of the Hansard collection. To process this text data, we perform case conversion, stemming, and removal of some stop words. For stemming, the SnowballStemmer of the NLTK toolbox by Bird et al. (2009) was used for both English and French languages. Although this stemmer has particular problems (such as mapping several different forms of a word to a single stem in one language but not in the other), they are left beyond our consideration. Moreover, in addition to the standard stop words of the NLTK toolbox, we also removed the following words that we consider to be stop words for our task[12] (and their possible forms):

- from English: ask, become, believe, can, could, come, cost, cut, do, done, follow, get, give, go, know, let, like, listen, live, look, lost, make, may, met, move, must, need, put, say, see, show, take, think, talk, use, want, will, also, another, back, day, certain, certainly, even, final, finally, first, future, general, good, high, just, last, long, major, many, new, next, now, one, point, since, thing, time, today, way, well, without;
- from French (translations in brackets): demander (ask), doit (must), devenir (become), dit (speak, talk), devoir (have to), donner (give), ila (he has), met (put), parler (speak, talk), penser (think), pourrait (could), pouvoir (can), prendre (take), savoir (know), aller (go), voir (see), vouloir (want), actuellement, après (after), aujourd’hui (today), autres (other), bien (good), beaucoup (a lot), besoin (need), cas (case), cause, cela (it), certain, chose (thing), déjà (already), dernier (last), égal (equal), entre (between), façon (way), grand (big), jour (day), lorsque (when), neuf (new), passé (past), plus, point, présent, prêts (ready), prochain (next), quelque (some), suivant (next), unique.

[12]This list of words was obtained by looking at words that appear in the top-20 words of a large number of topics in a first experiment. Removing these words did not change much the content of the topics, but made them much more interpretable.

farmers	agriculteurs	division	no	nato	otan	tax	impôts
agriculture	programme	negatived	vote	kosovo	kosovo	budget	budget
program	agricole	paired	rejetée	forces	militaires	billion	enfants
farm	pays	declare	voix	military	guerre	families	économie
country	important	yeas	mise	war	international	income	années
support	problème	divided	pairs	troops	pays	country	dollars
industry	aide	nays	porte	country	réfugiés	debt	pays
trade	agriculture	vote	contre	world	situation	students	finances
provinces	années	order	déclaration	national	paix	children	familles
work	secteur	deputy	suppléant	peace	yougoslavie	money	fiscal
problem	provinces	thibeault	vice	international	milosevic	finance	milliards
issue	gens	mcclelland	lethbridge	conflict	forces	education	libéraux
us	économie	ms	poisson	milosevic	serbes	liberal	jeunes
tax	industrie	oversee	mme	debate	intervention	fund	gens
world	dollars	rise	plantes	support	troupes	care	important
help	mesure	past	harvey	action	humanitaire	poverty	revenu
federal	faut	army	perdront	refugees	nations	jobs	mesure
producers	situation	peterson	sciences	ground	conflit	benefits	argent
national	réformiste	heed	liberté	happen	ethnique	child	santé
business	accord	moral	prière	issue	monde	pay	payer

Table 2: Topics 1 to 4.

work	travail	justice	jeunes	business	entreprises	board	commission
workers	négociations	young	justice	small	petites	wheat	blé
strike	travailleurs	crime	victimes	loans	programme	farmers	agriculteurs
legislation	grève	offenders	systéme	program	banques	grain	administration
union	emploi	victims	crime	bank	finances	producers	producteurs
agreement	droit	system	mesure	money	important	amendment	grain
labour	syndicat	legislation	criminel	finance	économie	market	conseil
right	services	sentence	contrevenants	access	secteur	directors	ouest
services	accord	youth	peine	jobs	argent	western	amendement
negotiations	voix	criminal	ans	economy	emplois	election	comité
chairman	adopter	court	juge	industry	assurance	support	réformiste
public	réglement	issue	enfants	financial	financiére	party	propos
party	article	law	important	billion	appuyer	farm	important
employees	retour	community	gens	support	créer	agriculture	compte
collective	gens	right	tribunaux	ovid	choc	clause	prix
agreed	conseil	reform	droit	merger	accés	ottawa	no
board	collectivités	country	problème	information	milliards	us	dispositions
arbitration	postes	problem	réformiste	size	propos	vote	information
grain	grain	person	traité	korea	pme	cwb	mesure
order	trésor	support	faut	companies	obtenir	states	produits

Table 3: Topics 5 to 8.

After stemming and removing stop words, several files had different number of documents in each language and had to be removed too. The numbers of these files are: 16, 36, 49 55, 88, 103, 110, 114, 123, 155, 159, 204, 229, 240, 2-17, 2-35.

We also removed the 15 most frequent words from each language. These include:

- in English: Mr, govern, member, speaker, minist(er), Hon, Canadian, Canada, bill, hous(e), peopl(e), year, act, motion, question;
- in French: gouvern(er), président, loi, déput(é), ministr(e), canadien, Canada, projet, Monsieur, question, part(y), chambr(e), premi(er), motion, Hon.

Removing these words is not necessary, but improves the presentation of the learned topics significantly. Indeed, the most frequent words tend to appear in nearly every topic (often in pairs in both languages as translations of each other, e.g., "member" and "député" or "Canada" in both languages, which confirms one more time the correctness of our algorithm).

Finally, we select $M_1 = M_2 = 5,000$ words for each language to form matrices X_1 and X_2 each containing $N = 11,969$ documents in columns. As stemming removes the words endings, we map the stemmed words to the respective most frequent original words when showing off the topics in Tables 2-6.

Supplementary References

Jacod, J. and Protter, P. *Probability Essentials*. Springer, 2004.

nisga	nisga	pension	régime	newfoundland	terre	health	santé
treaty	autochtones	plan	pensions	amendment	droit	research	recherche
aboriginal	traité	fund	cotisations	school	modifications	care	fédéral
agreement	accord	benefits	prestations	education	provinces	federal	provinces
right	droit	public	retraite	right	école	provinces	soins
land	nations	investment	emploi	constitution	comité	budget	budget
reserve	britannique	money	assurance	provinces	éducation	billion	dollars
national	indiennes	contribution	investissement	committee	enseignement	social	systéme
british	terre	cpp	fonds	system	systéme	money	finances
columbia	colombie	retirement	années	reform	enfants	tax	transfert
indian	réserves	pay	ans	minority	vote	system	milliards
court	non	billion	argent	denominational	amendement	provincial	domaine
party	affaires	change	important	referendum	constitution	fund	sociale
law	négociations	liberal	administration	children	religieux	country	années
native	bande	legislation	dollars	quebec	référendum	quebec	maladie
non	réformiste	board	propos	parents	article	transfer	important
constitution	constitution	employment	milliards	students	réformiste	debt	programme
development	application	tax	gens	change	québec	liberal	libéraux
reform	user	rate	taux	party	constitutionnelle	services	environnement
legislation	gestion	amendment	rpc	labrador	confessionnelles	issue	assurance

Table 4: Topics 9 to 12.

party	pays	tax	agence	quebec	québec	court	pêches
country	politique	provinces	provinces	federal	québécois	right	droit
issue	important	agency	revenu	information	fédéral	fisheries	juge
us	comité	federal	impôts	provinces	provinces	decision	cours
debate	libéraux	revenue	fiscal	protection	protection	fish	gens
liberal	réformiste	taxpayers	fédéral	right	renseignements	issue	décision
committee	gens	equalization	contribuables	legislation	droit	law	important
work	débat	system	payer	provincial	personnel	work	pays
order	accord	services	taxe	person	privé	us	traité
support	démocratique	accountability	péréquation	law	protéger	party	conservateur
reform	québécois	amendment	argent	constitution	électronique	debate	région
election	réglement	billion	services	privacy	article	justice	problème
world	propos	money	fonction	country	commerce	problem	supréme
quebec	collégue	party	modifier	electronic	provinciaux	community	tribunaux
standing	parlementaire	provincial	article	court	bloc	supreme	faut
national	appuyer	public	ministére	bloc	vie	country	situation
interest	opposition	business	administration	students	application	area	victimes
important	élections	reform	déclaration	section	citoyens	case	appuyer
right	bloc	office	tps	clear	non	order	mesure
public	industrie	support	provinciaux	states	nationale	parliament	trouve

Table 5: Topics 13 to 16.

legislation	important	national	important	vote	voix	water	eau
issue	environnement	area	gens	yeas	no	trade	ressources
amendment	mesure	parks	environnement	division	adopter	resources	accord
committee	enfants	work	parcs	nays	vote	country	environnement
support	comité	country	pays	agreed	non	agreement	important
protection	propos	us	marine	deputy	contre	provinces	industrie
information	pays	development	mesure	paired	dépenses	industry	américains
industry	appuyer	support	propos	responsible	accord	protection	pays
concerned	protection	community	fédéral	treasury	conseil	export	provinces
right	article	federal	jeunes	divided	budget	environmental	exportations
important	droit	issue	appuyer	order	crédit	us	échange
change	accord	legislation	années	fiscal	trésor	freshwater	conservateur
world	gens	help	assurance	amount	oui	federal	responsabilité
law	amendement	liberal	gestion	pleased	mise	world	effet
families	adopter	world	conservateur	budget	propos	issue	quantité
work	industrie	responsible	accord	ms	porte	legislation	traité
children	non	concerned	région	infrastructure	lib	environment	commerce
order	société	committee	problème	board	pairs	responsible	unis
national	porte	problem	nationale	consent	veuillent	development	économie
states	no	important	québec	estimates	vice	culture	alena

Table 6: Topics 17 to 20.