Rethinking LDA: Moment Matching for Discrete ICA

Anastasia Podosinnikova, Francis Bach, Simon Lacoste-Julien

Rethinking LDA: Moment Matching for Discrete ICA

Anastasia Podosinnikova **Francis Bach** **Simon Lacoste-Julien**
INRIA - École normale supérieure Paris

Abstract

We consider moment matching techniques for estimation in latent Dirichlet allocation (LDA). By drawing explicit links between LDA and discrete versions of independent component analysis (ICA), we first derive a new set of cumulant-based tensors, with an improved sample complexity. Moreover, we reuse standard ICA techniques such as joint diagonalization of tensors to improve over existing methods based on the tensor power method. In an extensive set of experiments on both synthetic and real datasets, we show that our new combination of tensors and orthogonal joint diagonalization techniques outperforms existing moment matching methods.

1 Introduction

Topic models have emerged as flexible and important tools for the modelisation of text corpora. While early work has focused on graphical-model approximate inference techniques such as variational inference [1] or Gibbs sampling [2], tensor-based moment matching techniques have recently emerged as strong competitors due to their computational speed and theoretical guarantees [3, 4]. In this paper, we draw explicit links with the independent component analysis (ICA) literature (e.g., [5] and references therein) by showing a strong relationship between latent Dirichlet allocation (LDA) [1] and ICA [6, 7, 8]. We can then reuse standard ICA techniques and results, and derive new tensors with better sample complexity and new algorithms based on joint diagonalization.

2 Is LDA discrete PCA or discrete ICA?

Notation. Following the text modeling terminology, we define a corpus $X = \{x_1, \dots, x_N\}$ as a collection of N documents. Each document is a collection $\{w_{n1}, \dots, w_{nL_n}\}$ of L_n tokens. It is convenient to represent the ℓ-th token of the n-th document as a 1-of-M encoding with an indicator vector $w_{n\ell} \in \{0,1\}^M$ with only one non-zero, where M is the vocabulary size, and each document as the count vector $x_n := \sum_\ell w_{n\ell} \in \mathbb{R}^M$. In such representation, the length L_n of the n-th document is $L_n = \sum_m x_{nm}$. We will always use the index $k \in \{1, \dots, K\}$ to refer to topics,the index $n \in \{1, \dots, N\}$ to refer to documents, the index $m \in \{1, \dots, M\}$ to refer to words from the vocabulary, and the index $\ell \in \{1, \dots, L_n\}$ to refer to tokens of the n-th document. The plate diagrams of the models from this section are presented in Appendix A.

Latent Dirichlet allocation [1] is a generative probabilistic model for discrete data such as text corpora. In accordance to this model, the n-th document is modeled as an *admixture* over the vocabulary of M words with K latent topics. Specifically, the latent variable θ_n, which is sampled from the Dirichlet distribution, represents the topic mixture proportion over K topics for the n-th document. Given θ_n, the topic choice $z_{n\ell}|\theta_n$ for the ℓ-th token is sampled from the multinomial distribution with the probability vector θ_n. The token $w_{n\ell}|z_{n\ell}, \theta_n$ is then sampled from the multinomial distribution with the probability vector $d_{z_{n\ell}}$, or d_k if k is the index of the non-zero element in $z_{n\ell}$. This vector d_k is the k-th topic, that is a vector of probabilities over the words from the vocabulary subject to the simplex constraint, i.e., $d_k \in \Delta_M$, where $\Delta_M := \{d \in \mathbb{R}^M \,:\, d \succeq 0,\ \sum_m d_m = 1\}$. This generative process of a document (the index n is omitted for simplicity) can be summarized as

$$\begin{aligned} \theta &\sim \text{Dirichlet}(c), \\ z_\ell|\theta &\sim \text{Multinomial}(1, \theta), \\ w_\ell|z_\ell, \theta &\sim \text{Multinomial}(1, d_{z_\ell}). \end{aligned} \tag{1}$$

One can think of the latent variables z_ℓ as auxiliary variables which were introduced for convenience of inference, but can in fact be marginalized out [9], which leads to the following model

$$\begin{aligned} \theta &\sim \text{Dirichlet}(c), \\ x|\theta &\sim \text{Multinomial}(L, D\theta), \end{aligned} \qquad \text{LDA model} \tag{2}$$

where $D \in \mathbb{R}^{M\times K}$ is the topic matrix with the k-th column equal to the k-th topic d_k, and $c \in \mathbb{R}^K_{++}$ is the vector of parameters for the Dirichlet distribution. While a document is represented as a set of tokens w_ℓ in the formulation (1), the formulation (2) instead compactly represents a document as the count vector x. Although the two representations are equivalent, we focus on the second one in this paper and therefore refer to it as the LDA model.

Importantly, the LDA model does not model the length of documents. Indeed, although the original paper [1] proposes to model the document length as $L|\lambda \sim \text{Poisson}(\lambda)$, this is never used in practice and, in particular, the parameter λ is not learned. Therefore, in the way that the LDA model is typically used, it does not provide a complete generative process of a document as there is no rule to sample $L|\lambda$. In this paper, this fact is important, as we need to model the document length in order to make the link with discrete ICA.

Discrete PCA. The LDA model (2) can be seen as a discretization of principal component analysis (PCA) via replacement of the normal likelihood with the multinomial one and adjusting the prior [9] in the following probabilistic PCA model [10, 11]: $\theta \sim \text{Normal}(0, I_K)$ and $x|\theta \sim \text{Normal}(D\theta, \sigma^2 I_M)$, where $D \in \mathbb{R}^{M\times K}$ is a transformation matrix and σ is a parameter.

Discrete ICA (DICA). Interestingly, a small extension of the LDA model allows its interpretation as a discrete independent component analysis model. The extension naturally arises when the document length for the LDA model is modeled as a random variable from the gamma-Poisson mixture (which is equivalent to a negative binomial random variable), i.e., $L|\lambda \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(c_0, b)$, where $c_0 := \sum_k c_k$ is the shape parameter and $b > 0$ is the rate parameter. The LDA model (2) with such document length is equivalent (see Appendix B.1) to

$$\begin{aligned} \alpha_k &\sim \text{Gamma}(c_k, b), \\ x_m|\alpha &\sim \text{Poisson}([D\alpha]_m), \end{aligned} \qquad \text{GP model} \tag{3}$$

where all $\alpha_1, \alpha_2, \ldots, \alpha_K$ are mutually independent, the parameters c_k coincide with the ones of the LDA model in (2), and the free parameter b can be seen (see Appendix B.2) as a scaling parameter for the document length when c_0 is already prescribed.

This model was introduced by Canny [12] and later named as a discrete ICA model [13]. It is more natural, however, to name model (3) as the gamma-Poisson (GP) model and the model

$$\begin{aligned} \alpha_1, \ldots, \alpha_K &\sim \text{mutually independent}, \\ x_m|\alpha &\sim \text{Poisson}([D\alpha]_m) \end{aligned} \qquad \text{DICA model} \tag{4}$$

as the discrete ICA (DICA) model. The only difference between (4) and the standard ICA model [6, 7, 8] (without additive noise) is the presence of the Poisson noise which enforces discrete, instead of continuous, values of x_m. Note also that (a) the discrete ICA model is a *semi-parametric* model that can adapt to any distribution on the topic intensities α_k and that (b) the GP model (3) is a particular case of both the LDA model (2) and the DICA model (4).

Thanks to this close connection between LDA and ICA, we can reuse standard ICA techniques to derive new efficient algorithms for topic modeling.

3 Moment matching for topic modeling

The method of moments estimates latent parameters of a probabilistic model by matching theoretical expressions of its moments with their sample estimates. Recently [3, 4], the method of moments was applied to different latent variable models including LDA, resulting in computationally fast

learning algorithms with theoretical guarantees. For LDA, they (a) construct *LDA moments* with a particular diagonal structure and (b) develop algorithms for estimating the parameters of the model by exploiting this diagonal structure. In this paper, we introduce the novel *GP/DICA cumulants* with a similar to the LDA moments structure. This structure allows to reapply the algorithms of [3, 4] for the estimation of the model parameters, with the same theoretical guarantees. We also consider another algorithm applicable to both the LDA moments and the GP/DICA cumulants.

3.1 Cumulants of the GP and DICA models

In this section, we derive and analyze the novel cumulants of the DICA model. As the GP model is a particular case of the DICA model, all results of this section extend to the GP model.

The first three *cumulant tensors* for the random vector x can be defined as follows

$$\operatorname{cum}(x) := \mathbb{E}(x), \tag{5}$$

$$\operatorname{cum}(x, x) := \operatorname{cov}(x, x) = \mathbb{E}\left[(x - \mathbb{E}(x))(x - \mathbb{E}(x))^\top\right], \tag{6}$$

$$\operatorname{cum}(x, x, x) := \mathbb{E}\left[(x - \mathbb{E}(x)) \otimes (x - \mathbb{E}(x)) \otimes (x - \mathbb{E}(x))\right], \tag{7}$$

where $\otimes$ denotes the tensor product (see some properties of cumulants in Appendix C.1). The essential property of the cumulants (which does not hold for the moments) that we use in this paper is that the cumulant tensor for a random vector with *independent* components is *diagonal*.

Let $y = D\alpha$; then for the Poisson random variable $x_m|y_m \sim \text{Poisson}(y_m)$, the expectation is $\mathbb{E}(x_m|y_m) = y_m$. Hence, by the law of total expectation and the linearity of expectation, the expectation in (5) has the following form

$$\mathbb{E}(x) = \mathbb{E}(\mathbb{E}(x|y)) = \mathbb{E}(y) = D\mathbb{E}(\alpha). \tag{8}$$

Further, the variance of the Poisson random variable x_m is $\operatorname{var}(x_m|y_m) = y_m$ and, as x_1, x_2, …, x_M are conditionally independent given y, then their covariance matrix is diagonal, i.e., $\operatorname{cov}(x, x|y) = \operatorname{diag}(y)$. Therefore, by the law of total covariance, the covariance in (6) has the form

$$\begin{aligned} \operatorname{cov}(x, x) &= \mathbb{E}\left[\operatorname{cov}(x, x|y)\right] + \operatorname{cov}\left[\mathbb{E}(x|y), \mathbb{E}(x|y)\right] \\ &= \operatorname{diag}\left[\mathbb{E}(y)\right] + \operatorname{cov}(y, y) = \operatorname{diag}\left[\mathbb{E}(x)\right] + D\operatorname{cov}(\alpha, \alpha)D^\top, \end{aligned} \tag{9}$$

where the last equality follows by the multilinearity property of cumulants (see Appendix C.1). Moving the first term from the RHS of (9) to the LHS, we define

$$S := \operatorname{cov}(x, x) - \operatorname{diag}\left[\mathbb{E}(x)\right]. \qquad \text{DICA S-cum.} \tag{10}$$

From (9) and by the independence of $\alpha_1, \ldots, \alpha_K$ (see Appendix C.3), S has the following diagonal structure

$$S = \sum\nolimits_k \operatorname{var}(\alpha_k) d_k d_k^\top = D\operatorname{diag}\left[\operatorname{var}(\alpha)\right] D^\top. \tag{11}$$

By analogy with the second order case, using the law of total cumulance, the multilinearity property of cumulants, and the independence of $\alpha_1, \ldots, \alpha_K$, we derive in Appendix C.2 the expression (24), similar to (9), for the third cumulant (7). Moving the terms in this expression, we define a tensor T with the following element

$$\begin{aligned} [T]_{m_1 m_2 m_3} &:= \operatorname{cum}(x_{m_1}, x_{m_2}, x_{m_3}) + 2\delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) \qquad \text{DICA T-cum.} \\ &- \delta(m_2, m_3)\operatorname{cov}(x_{m_1}, x_{m_2}) - \delta(m_1, m_3)\operatorname{cov}(x_{m_1}, x_{m_2}) - \delta(m_1, m_2)\operatorname{cov}(x_{m_1}, x_{m_3}), \end{aligned} \tag{12}$$

where δ is the Kronecker delta. By analogy with (11) (Appendix C.3), the diagonal structure of the tensor T:

$$T = \sum\nolimits_k \operatorname{cum}(\alpha_k, \alpha_k, \alpha_k) d_k \otimes d_k \otimes d_k. \tag{13}$$

In Appendix E.1, we recall (in our notation) the matrix S (39) and the tensor T (40) for the LDA model [3], which are analogues of the matrix S (10) and the tensor T (12) for the GP/DICA models. Slightly abusing terminology, we refer to the matrix S (39) and the tensor T (40) as the *LDA moments* and to the matrix S (10) and the tensor T (12) as the *GP/DICA cumulants*. The diagonal structure (41) & (42) of the LDA moments is similar to the diagonal structure (11) & (13) of the GP/DICA cumulants, though arising through a slightly different argument, as discussed at the end of

Appendix E.1. Importantly, due to this similarity, the algorithmic frameworks for both the GP/DICA cumulants and the LDA moments coincide.

The following sample complexity results apply to the sample estimates of the GP cumulants:[1]

Proposition 3.1. *Under the GP model, the expected error for the sample estimator* $\widehat{S}$ (29) *for the GP cumulant* S (10) *is:*

$$\mathbb{E}\left[\|\widehat{S} - S\|_F\right] \leq \sqrt{\mathbb{E}\left[\|\widehat{S} - S\|_F^2\right]} \leq O\left(\frac{1}{\sqrt{N}} \max\left[\Delta \bar{L}^2, \bar{c}_0 \bar{L}\right]\right), \tag{14}$$

where $\Delta := \max_k \|d_k\|_2^2$, $\bar{c}_0 := \min(1, c_0)$ *and* $\bar{L} := \mathbb{E}(L)$.

A high probability bound could be derived using concentration inequalities for Poisson random variables [14]; but the expectation already gives the right order of magnitude for the error (for example via Markov's inequality). The expression (29) for an unbiased finite sample estimate $\widehat{S}$ of S and the expression (30) for an unbiased finite sample estimate $\widehat{T}$ of T are defined[2] in Appendix C.4. A sketch of a proof for Proposition 3.1 can be found in Appendix D.

By following a similar analysis as in [15], we can rephrase the topic recovery error in term of the error on the GP cumulant. Importantly, the whitening transformation (introduced in Section 4) redivides the error on S (14) by $\bar{L}^2$, which is the scale of S (see Appendix D.5 for details). This means that the contribution from $\hat{S}$ to the recovery error will scale as $O(1/\sqrt{N} \max\{\Delta, \bar{c}_0/\bar{L}\})$, where both Δ and $\bar{c}_0/\bar{L}$ are smaller than 1 and can be very small. We do not present the exact expression for the expected squared error for the estimator of T, but due to a similar structure in the derivation, we expect the analogous bound of $\mathbb{E}[\|\widehat{T} - T\|_F] \leq 1/\sqrt{N} \max\{\Delta^{3/2}\bar{L}^3, \bar{c}_0^{3/2}\bar{L}^{3/2}\}$.

Current sample complexity results of the LDA moments [3] can be summarized as $O(1/\sqrt{N})$. However, the proof (which can be found in the supplementary material [15]) analyzes only the case when finite sample estimates of the LDA moments are constructed from *one* triple per document, i.e., $w_1 \otimes w_2 \otimes w_3$ only, and not from the U-statistics that average multiple (dependent) triples per document as in the practical expressions (43) and (44) (Appendix F.4). Moreover, one has to be careful when comparing upper bounds. Nevertheless, comparing the bound (14) with the current theoretical results for the LDA moments, we see that the GP/DICA cumulants sample complexity contains the ℓ_2-norm of the columns of the topic matrix D in the numerator, as opposed to the $O(1)$ coefficient for the LDA moments. This norm can be significantly smaller than 1 for vectors in the simplex (e.g., $\Delta = O(1/\|d_k\|_0)$ for sparse topics). This suggests that the GP/DICA cumulants may have better finite sample convergence properties than the LDA moments and our experimental results in Section 5.2 are indeed consistent with this statement.

The GP/DICA cumulants have a somewhat more intuitive derivation than the LDA moments as they are expressed via the count vectors x (which are the sufficient statistics for the model) and not the tokens w_ℓ's. Note also that the construction of the LDA moments depend on the unknown parameter c_0. Given that we are in an unsupervised setting and that moreover the evaluation of LDA is a difficult task [16], setting this parameter is non-trivial. In Appendix G.4, we observe experimentally that the LDA moments are somewhat sensitive to the choice of c_0.

4 Diagonalization algorithms

How is the diagonal structure (11) of S and (13) of T going to be helpful for the estimation of the model parameters? This question has already been thoroughly investigated in the signal processing (see, e.g., [17, 18, 19, 20, 21, 5] and references therein) and machine learning (see [3, 4] and references therein) literature. We review the approach in this section. Due to similar diagonal structure, the algorithms of this section apply to both the LDA moments and the GP/DICA cumulants.

For simplicity, let us rewrite the expressions (11) and (13) for S and T as follows

$$S = \sum\nolimits_k s_k d_k d_k^\top, \qquad T = \sum\nolimits_k t_k d_k \otimes d_k \otimes d_k, \tag{15}$$

[1] Note that the expected squared error for the DICA cumulants is similar, but the expressions are less compact and, in general, depend on the prior on α_k.

[2] For completeness, we also present the finite sample estimates $\widehat{S}$ (43) and $\widehat{T}$ (44) of S (39) and T (40) for the LDA moments (which are consistent with the ones suggested in [4]) in Appendix F.4.

where $s_k := \mathrm{var}(\alpha_k)$ and $t_k := \mathrm{cum}(\alpha_k, \alpha_k, \alpha_k)$. Introducing the rescaled topics $\widetilde{d}_k := \sqrt{s_k} d_k$, we can also rewrite $S = \widetilde{D}\widetilde{D}^\top$. Following the same assumption from [3] that the topic vectors are linearly independent ($\widetilde{D}$ is full rank), we can compute a whitening matrix $W \in \mathbb{R}^{K \times M}$ of S, i.e., a matrix such that $WSW^\top = I_K$ where I_K is the K-by-K identity matrix (see Appendix F.1 for more details). As a result, the vectors $z_k := W\widetilde{d}_k$ form an orthonormal set of vectors.

Further, let us define a projection $\mathcal{T}(v) \in \mathbb{R}^{K \times K}$ of a tensor $\mathcal{T} \in \mathbb{R}^{K \times K \times K}$ onto a vector $u \in \mathbb{R}^K$:

$$\mathcal{T}(u)_{k_1 k_2} := \sum\nolimits_{k_3} \mathcal{T}_{k_1 k_2 k_3} u_{k_3}. \tag{16}$$

Applying the multilinear transformation (see, e.g., [4] for the definition) with $W^\top$ to the tensor T from (15) and projecting the resulting tensor $\mathcal{T} := T(W^\top, W^\top, W^\top)$ onto some vector $u \in \mathbb{R}^K$, we obtain

$$\mathcal{T}(u) = \sum\nolimits_k \widetilde{t}_k \langle z_k, u \rangle z_k z_k^\top, \tag{17}$$

where $\widetilde{t}_k := t_k / s_k^{3/2}$ is due to the rescaling of topics and $\langle \cdot, \cdot \rangle$ stands for the inner product. As the vectors z_k are orthonormal, the pairs z_k and $\lambda_k := \widetilde{t}_k \langle z_k, u \rangle$ are the eigenpairs of the matrix $\mathcal{T}(u)$, which are uniquely defined if the eigenvalues λ_k are all different. If they are unique, we can recover the GP/DICA (as well as LDA) model parameters via $\widetilde{d}_k = W^\dagger z_k$ and $\widetilde{t}_k = \lambda_k / \langle z_k, u \rangle$.

This procedure was referred to as the spectral algorithm for LDA [3] and the fourth-order[3] blind identification algorithm for ICA [17, 18]. Indeed, one can expect that the finite sample estimates $\widehat{S}$ (29) and $\widehat{T}$ (30) possess approximately the diagonal structure (11) and (13) and, therefore, the reasoning from above can be applied, assuming that the effect of the sampling error is controlled.

This spectral algorithm, however, is known to be quite unstable in practice (see, e.g., [22]). To overcome this problem, other algorithms were proposed. For ICA, the most notable ones are probably the FastICA algorithm [20] and the JADE algorithm [21]. The FastICA algorithm, with appropriate choice of a contrast function, estimates iteratively the topics, making use of the orthonormal structure (17), and performs the deflation procedure at every step. The recently introduced tensor power method (TPM) for the LDA model [4] is close to the FastICA algorithm. Alternatively, the JADE algorithm modifies the spectral algorithm by performing *multiple* projections for (17) and then jointly diagonalizing the resulting matrices with an orthogonal matrix. The spectral algorithm is a special case of this orthogonal joint diagonalization algorithm when only one projection is chosen. Importantly, a fast implementation [23] of the orthogonal joint diagonalization algorithm from [24] was proposed, which is based on closed-form iterative Jacobi updates (see, e.g., [25] for the later).

In practice, the orthogonal joint diagonalization (JD) algorithm is more robust than FastICA (see, e.g., [26, p. 30]) or the spectral algorithm. Moreover, although the application of the JD algorithm for the learning of topic models was mentioned in the literature [4, 27], it was never implemented in practice. In this paper, we apply the JD algorithm for the diagonalization of the GP/DICA cumulants as well as the LDA moments, which is described in Algorithm 1. Note that the choice of a projection vector $v_p \in \mathbb{R}^M$ obtained as $v_p = \widehat{W}^\top u_p$ for some vector $u_p \in \mathbb{R}^K$ is important and corresponds to the multilinear transformation of $\widehat{T}$ with $\widehat{W}^\top$ along the third mode. Importantly, in Algorithm 1, the joint diagonalization routine is performed over $(P+1)$ matrices of size $K \times K$, where the number of topics K is usually not too big. This makes the algorithm computationally fast (see Appendix G.1). The same is true for the spectral algorithm, but not for TPM.

In Section 5.1, we compare experimentally the performance of the spectral, JD, and TPM algorithms for the estimation of the parameters of the GP/DICA as well as LDA models. We are not aware of any experimental comparison of these algorithms in the LDA context. While already working on this manuscript, the JD algorithm was also independently analyzed by [27] in the context of tensor factorization for general latent variable models. However, [27] focused mostly on the comparison of approaches for tensor factorization and their stability properties, with brief experiments using a latent variable model related but not equivalent to LDA for community detection. -In contrast, we provide a detailed experimental comparison in the context of LDA in this paper, as well as propose a novel cumulant-based estimator. Due to the space restriction the estimation of the topic matrix D and the (gamma/Dirichlet) parameter c are moved to Appendix F.6.

[3] See Appendix C.5 for a discussion on the orders.

Algorithm 1 Joint diagonalization (JD) algorithm for GP/DICA cumulants (or LDA moments)

1: *Input:* $X \in \mathbb{R}^{M \times N}$, K, P (number of random projections); (and c_0 for LDA moments)
2: Compute sample estimate $\widehat{S} \in \mathbb{R}^{M \times M}$ ((29) for GP/DICA / (43) for LDA in Appendix F)
3: Estimate whitening matrix $\widehat{W} \in \mathbb{R}^{K \times M}$ of $\widehat{S}$ (see Appendix F.1)
option (a): Choose vectors $\{u_1, u_2, \ldots, u_P\} \subseteq \mathbb{R}^K$ uniformly at random from the unit ℓ_2-sphere and set $v_p = \widehat{W}^\top u_p \in \mathbb{R}^M$ for all $p = 1, \ldots, P$ ($P = 1$ yields the spectral algorithm)
option (b): Choose vectors $\{u_1, u_2, \ldots, u_P\} \subseteq \mathbb{R}^K$ as the canonical basis $e_1, e_2, \ldots, e_K$ of $\mathbb{R}^K$ and set $v_p = \widehat{W}^\top u_p \in \mathbb{R}^M$ for all $p = 1, \ldots, K$
4: For $\forall p$, compute $B_p = \widehat{W}\widehat{T}(v_p)\widehat{W}^\top \in \mathbb{R}^{K \times K}$ ((52) for GP/DICA / (54) for LDA; Appendix F)
5: Perform orthogonal joint diagonalization of matrices $\{\widehat{W}\widehat{S}\widehat{W}^\top = I_K,\ B_p,\ p = 1, \ldots, P\}$ (see [24] and [23]) to find an orthogonal matrix $V \in \mathbb{R}^{K \times K}$ and vectors $\{a_1, a_2, \ldots, a_P\} \subset \mathbb{R}^K$ such that
$$V\widehat{W}\widehat{S}\widehat{W}^\top V^\top = I_K, \text{ and } VB_pV^\top \approx \operatorname{diag}(a_p),\ p = 1, \ldots, P$$
6: Estimate joint diagonalization matrix $A = V\widehat{W}$ and values a_p, $p = 1, \ldots, P$
7: *Output:* Estimate of D and c as described in Appendix F.6

5 Experiments

In this section, (a) we compare experimentally the GP/DICA cumulants with the LDA moments and (b) the spectral algorithm [3], the tensor power method [4] (TPM), the joint diagonalization (JD) algorithm from Algorithm 1, and variational inference for LDA [1].

Real data: the associated press (AP) dataset, from D. Blei's web page,[4] with $N = 2,243$ documents and $M = 10,473$ vocabulary words and the average document length $\widehat{L} = 194$; the NIPS papers dataset[5] [28] of $2,483$ NIPS papers and $14,036$ words, and $\widehat{L} = 1,321$; the KOS dataset,[6] from the UCI Repository, with $3,430$ documents and $6,906$ words, and $\widehat{L} = 136$.

Semi-synthetic data are constructed by analogy with [29]: (1) the LDA parameters D and c are learned from the real datasets with variational inference and (2) toy data are sampled from a model of interest with the given parameters D and c. This provides the ground truth parameters D and c. For each setting, data are sampled 5 times and the results are averaged. We plot error bars that are the minimum and maximum values. For the AP data, $K \in \{10, 50\}$ topics are learned and, for the NIPS data, $K \in \{10, 90\}$ topics are learned. For larger K, the obtained topic matrix is ill-conditioned, which violates the identifiability condition for topic recovery using moment matching techniques [3]. All the documents with less than 3 tokens are resampled.

Sampling techniques. All the sampling models have the parameter c which is set to $c = c_0\bar{c}/\left\|\bar{c}\right\|_1$, where $\bar{c}$ is the learned c from the real dataset with variational LDA, and c_0 is a parameter that we can vary. The *GP* data are sampled from the gamma-Poisson model (3) with $b = c_0/\widehat{L}$ so that the expected document length is $\widehat{L}$ (see Appendix B.2). The *LDA-fix(L)* data are sampled from the LDA model (2) with the document length being fixed to a given L. The *LDA-fix2(γ,L_1,L_2)* data are sampled as follows: $(1-\gamma)$-portion of the documents are sampled from the *LDA-fix(L_1)* model with a given document length L_1 and γ-portion of the documents are sampled from the *LDA-fix(L_2)* model with a given document length L_2.

Evaluation. The evaluation of topic recovery for semi-synthetic data is performed with the ℓ_1-error between the recovered $\widehat{D}$ and true D topic matrices with the best permutation of columns: $\text{err}_{\ell_1}(\widehat{D}, D) := \min_{\pi \in \text{PERM}} \frac{1}{2K} \sum_k \left\|\widehat{d}_{\pi_k} - d_k\right\|_1 \in [0, 1]$. The minimization is over the possible permutations $\pi \in \text{PERM}$ of the columns of $\widehat{D}$ and can be efficiently obtained with the Hungarian algorithm for bipartite matching. For the evaluation of topic recovery in the real data case, we use an approximation of the log-likelihood for held out documents as the metric [16]. See Appendix G.6 for more details.

[4] http://www.cs.columbia.edu/~blei/lda-c
[5] http://ai.stanford.edu/~gal/data
[6] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

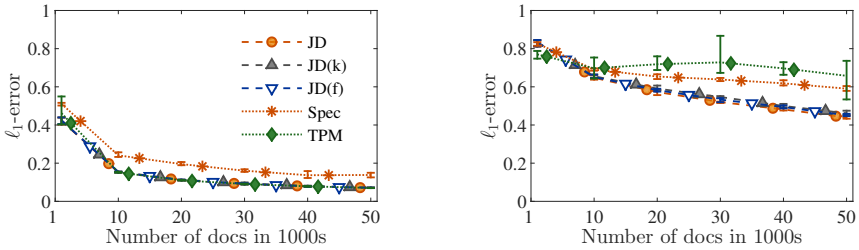


Figure 1: Comparison of the diagonalization algorithms. The topic matrix D and Dirichlet parameter c are learned for $K = 50$ from AP; c is scaled to sum up to 0.5 and b is set to fit the expected document length $\widehat{L} = 200$. The semi-synthetic dataset is sampled from *GP*; number of documents N varies from $1,000$ to $50,000$. **Left:** GP/DICA moments. **Right:** LDA moments. *Note*: a smaller value of the ℓ_1-error is better.

We use our Matlab implementation of the GP/DICA cumulants, the LDA moments, and the diagonalization algorithms. The datasets and the code for reproducing our experiments are available online.[7] In Appendix G.1, we discuss the complexity and implementation of the algorithms. We explain how we initialize the parameter c_0 for the LDA moments in Appendix G.3.

5.1 Comparison of the diagonalization algorithms

In Figure 1, we compare the diagonalization algorithms on the semi-synthetic AP dataset for $K = 50$ using the GP sampling. We compare the tensor power method (TPM) [4], the spectral algorithm (Spec), the orthogonal joint diagonalization algorithm (JD) described in Algorithm 1 with different options to choose the random projections: JD(k) takes $P = K$ vectors u_p sampled uniformly from the unit ℓ_2-sphere in $\mathbb{R}^K$ and selects $v_p = W^\top u_p$ (option (a) in Algorithm 1); JD selects the full basis $e_1, \dots, e_K$ in $\mathbb{R}^K$ and sets $v_p = W^\top e_p$ (as JADE [21]) (option (b) in Algorithm 1); $JD(f)$ chooses the full canonical basis of $\mathbb{R}^M$ as the projection vectors (computationally expensive).

Both the GP/DICA cumulants and LDA moments are well-specified in this setup. However, the LDA moments have a slower finite sample convergence and, hence, a larger estimation error for the same value N. As expected, the spectral algorithm is always slightly inferior to the joint diagonalization algorithms. With the GP/DICA cumulants, where the estimation error is low, all algorithms demonstrate good performance, which also fulfills our expectations. However, although TPM shows almost perfect performance in the case of the GP/DICA cumulants (left), it significantly deteriorates for the LDA moments (right), which can be explained by the larger estimation error of the LDA moments and lack of robustness of TPM. The running times are discussed in Appendix G.2. Overall, the orthogonal joint diagonalization algorithm with initialization of random projections as $W^\top$ multiplied with the canonical basis in $\mathbb{R}^K$ (JD) is both computationally efficient and fast.

5.2 Comparison of the GP/DICA cumulants and the LDA moments

In Figure 2, when sampling from the *GP* model (top, left), both the GP/DICA cumulants and LDA moments are well specified, which implies that the approximation error (i.e., the error w.r.t. the model (mis)fit) is low for both. The GP/DICA cumulants achieve low values of the estimation error already for $N = 10,000$ documents independently of the number of topics, while the convergence is slower for the LDA moments. When sampling from the *LDA-fix(200)* model (top, right), the GP/DICA cumulants are mis-specified and their approximation error is high, although the estimation error is low due to the faster finite sample convergence. One reason of poor performance of the GP/DICA cumulants, in this case, is the absence of variance in the document length. Indeed, if documents with two different lengths are mixed by sampling from the *LDA-fix2(0.5,20,200)* model (bottom, left), the GP/DICA cumulants performance improves. Moreover, the experiment with a changing fraction γ of documents (bottom, right) shows that a non-zero variance on the length improves the performance of the GP/DICA cumulants. As in practice real corpora usually have a non-zero variance for the document length, this bad scenario for the GP/DICA cumulants is not likely to happen.

[7] https://github.com/anastasia-podosinnikova/dica

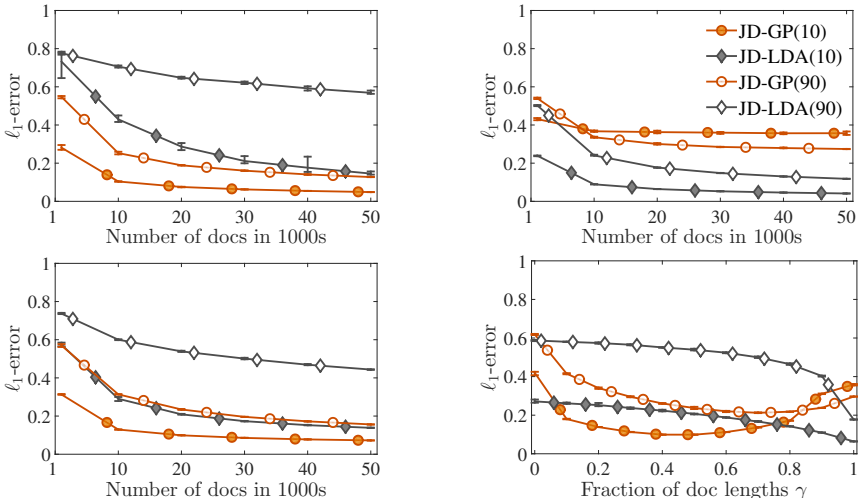


Figure 2: Comparison of the GP/DICA cumulants and LDA moments. Two topic matrices and parameters c_1 and c_2 are learned from the NIPS dataset for $K = 10$ and 90; c_1 and c_2 are scaled to sum up to $c_0 = 1$. Four corpora of different sizes N from $1,000$ to $50,000$: **top, left:** b is set to fit the expected document length $\widehat{L} = 1300$; sampling from the *GP* model; **top, right:** sampling from the *LDA-fix(200)* model; **bottom, left:** sampling from the *LDA-fix2(0.5,20,200)* model. **Bottom, right:** the number of documents here is fixed to $N = 20,000$; sampling from the *LDA-fix2(γ,20,200)* model varying the values of the fraction γ from 0 to 1 with the step 0.1. *Note*: a smaller value of the ℓ_1-error is better.

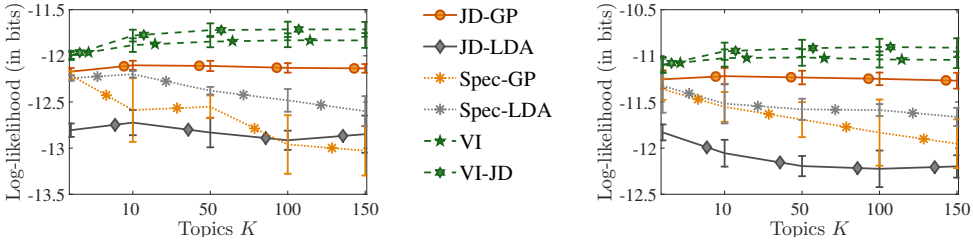


Figure 3: Experiments with real data. **Left:** the AP dataset. **Right:** the KOS dataset. *Note*: a higher value of the log-likelihood is better.

5.3 Real data experiments

In Figure 3, JD-GP, Spec-GP, JD-LDA, and Spec-LDA are compared with variational inference (VI) and with variational inference initialized with the output of JD-GP (VI-JD). We measure the held out log-likelihood per token (see Appendix G.7 for details on the experimental setup). The orthogonal joint diagonalization algorithm with the GP/DICA cumulants (JD-GP) demonstrates promising performance. In particular, the GP/DICA cumulants significantly outperform the LDA moments. Moreover, although variational inference performs better than the JD-GP algorithm, restarting variational inference with the output of the JD-GP algorithm systematically leads to better results. Similar behavior has already been observed (see, e.g., [30]).

6 Conclusion

In this paper, we have proposed a new set of tensors for a discrete ICA model related to LDA, where word counts are directly modelled. These moments make fewer assumptions regarding distributions, and are theoretically and empirically more robust than previously proposed tensors for LDA, both on synthetic and real data. Following the ICA literature, we showed that our joint diagonalization procedure is also more robust. Once the topic matrix has been estimated in a semi-parametric way where topic intensities are left unspecified, it would be interesting to learn the unknown distributions of the independent topic intensities.

Aknowledgements. This work was partially supported by the MSR-Inria Joint Center. The authors would like to thank Christophe Dupuy for helpful discussions.

References

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:903–1022, 2003.

[2] T. Griffiths. Gibbs sampling in the generative model of latent Dirichlet allocation. Technical report, Stanford University, 2002.

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

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

[5] P. Comon and C. Jutten, editors. *Handbook of blind sourse separation: independent component analysis and applications*. Academic Press, 2010.

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

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

[8] P. Comon. Independent component analysis, a new concept? *Signal Process.*, 36:287–314, 1994.

[9] W.L. Buntine. Variational extensions to EM and multinomial PCA. In *ECML*, 2002.

[10] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *J. R. Stat. Soc.*, 61:611–622, 1999.

[11] S. Roweis. EM algorithms for PCA and SPCA. In *NIPS*, 1998.

[12] J. Canny. GaP: a factor model for discrete data. In *SIGIR*, 2004.

[13] W.L. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *UAI*, 2004.

[14] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, 2013.

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

[16] H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *ICML*, 2009.

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

[18] J.-F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In *ICASSP*, 1990.

[19] J.-F. Cardoso and P. Comon. Independent component analysis, a survey of some algebraic methods. In *ISCAS*, 1996.

[20] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, 10(3):626–634, 1999.

[21] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. In *IEE Proceedings-F*, 1993.

[22] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Comput.*, 11:157–192, 1999.

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

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

[25] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer, 2nd edition, 2006.

[26] F.R. Bach and M.I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.

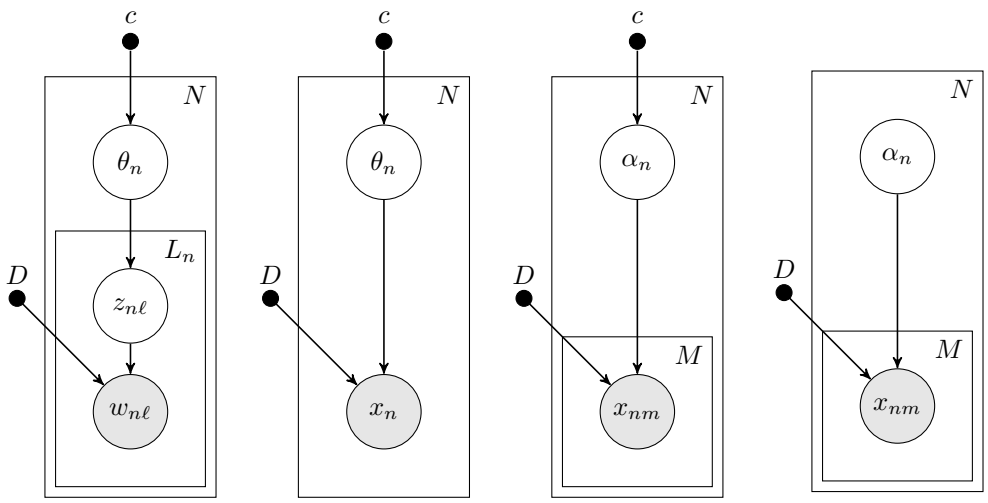
[27] V. Kuleshov, A.T. Chaganty, and P. Liang. Tensor factorization via matrix factorization. In *AISTATS*, 2015.

[28] A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. *J. Mach. Learn. Res.*, 8:2265–2295, 2007.

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

[30] S. Cohen and M. Collins. A provably correct learning algorithm for latent-variable PCFGs. In *ACL*, 2014.

A Appendix. Plate diagrams for the models from Section 2



(a) LDA (18) (b) LDA (19) (c) GP (20) (d) DICA (21)

Figure 4: Plate diagrams for the models from Section 2.

In Section 2, the index n, which stands for the n-th document, was omitted. For convenience, we recall the models. The LDA model in the tokens representation:

$$\begin{aligned} \theta_n &\sim \mathrm{Dirichlet}(c), \\ z_{n\ell}|\theta_n &\sim \mathrm{Multinomial}(1, \theta_n), \\ w_{n\ell}|z_{n\ell}, \theta_n &\sim \mathrm{Multinomial}(1, d_{z_{n\ell}}); \end{aligned} \tag{18}$$

the LDA model with the marginalized out latent variable z:

$$\begin{aligned} \theta_n &\sim \mathrm{Dirichlet}(c), \\ x_n|\theta_n &\sim \mathrm{Multinomial}(L_n, D\theta_n); \end{aligned} \tag{19}$$

the GP model:

$$\begin{aligned} \alpha_{nk} &\sim \mathrm{Gamma}(c_k, b), \\ x_{nm}|\alpha_n &\sim \mathrm{Poisson}([D\alpha_n]_m); \end{aligned} \tag{20}$$

and the DICA model:

$$\begin{aligned} \alpha_{n1}, \ldots, \alpha_{nK} &\sim \text{mutually independent}, \\ x_{nm}|\alpha_n &\sim \mathrm{Poisson}([D\alpha_n]_m). \end{aligned} \tag{21}$$

B Appendix. The GP model

B.1 The connection between the LDA and GP models

To show that the LDA model (2) with the additional assumption that the document length is modeled as a gamma-Poisson random variable is equivalent to the GP model (3), we show that:

- when modeling the document length L as a Poisson random variable with a parameter λ, the count vectors $x_1, x_2, \ldots, x_M$ are mutually independent Poisson random variables;
- the Gamma prior on λ reveals the connection $\alpha_k = \lambda\theta_k$ between the Dirichlet random variable θ and the mutually independent gamma random variables $\alpha_1, \alpha_2, \ldots, \alpha_K$.

For completeness, we repeat the known result that if $L \sim \mathrm{Poisson}(\lambda)$ and $x|L \sim \mathrm{Multinomial}(L, D\theta)$ (which thus means that $L = \sum_m x_m$ with probability one), then x_1, x_2, $\ldots$, x_M are mutually independent Poisson random variables with parameters $\lambda\,[D\theta]_1$, $\lambda\,[D\theta]_2$, $\ldots$,

$\lambda [D\theta]_M$. Indeed, we consider the following joint probability mass function where x and L are assumed to be non-negative integers:

$$\begin{aligned}
p(x, L|\theta, \lambda) =& p(L|\lambda)p(x|L, \theta) \\
=& \mathbb{1}_{\{L=\sum_m x_m\}} \frac{\exp(-\lambda)\lambda^L}{\cancel{L!}} \frac{\cancel{L!}}{\prod_m x_m!} \prod_m [D\theta]_m^{x_m} \\
=& \mathbb{1}_{\{L=\sum_m x_m\}} \exp(-\lambda \sum_m [D\theta]_m) \lambda^{\sum_m x_m} \prod_m \frac{[D\theta]_m^{x_m}}{x_m!} \\
=& \mathbb{1}_{\{L=\sum_m x_m\}} \prod_m \frac{\exp(-\lambda [D\theta]_m)(\lambda [D\theta]_m)^{x_m}}{x_m!} \\
=& \mathbb{1}_{\{L=\sum_m x_m\}} \prod_m \text{Poisson}(x_m; \lambda [D\theta]_m),
\end{aligned}$$

where in the third equation we used the fact that

$$\sum_m [D\theta]_m = \sum_{m,k} D_{mk}\theta_k = \sum_k \theta_k \sum_m D_{mk} = 1.$$

We thus have $p(x, L|\theta, \lambda) = p(L|x) \prod_m p(x_m|\lambda[D\theta]_m)$ where $p(L|x)$ is simply the deterministic distribution $\mathbb{1}_{\{L=\sum_m x_m\}}$ and $p(x_m|\lambda[D\theta]_m)$ for $m = 1, \ldots, M$ are independent $\text{Poisson}(\lambda[D\theta]_m)$ distributions (and thus do not depend on L). Note that in the notation introduced in the paper, $D_{mk} = d_{km}$. Hence, by using the construction of the Dirichlet distribution from the normalization of independent gamma random variables, we can show that the LDA model with a gamma-Poisson prior over the length is equivalent to the following model (recall, that $c_0 = \sum_k c_k$):

$$\begin{aligned}
\lambda &\sim \text{Gamma}(c_0, b), \\
\theta &\sim \text{Dirichlet}(c), \\
x_m|\lambda, \theta &\sim \text{Poisson}([D(\lambda\theta)]_m).
\end{aligned} \tag{22}$$

More specifically, we complete the second part of the argument with the following properties. When $\alpha_1, \alpha_2, \ldots, \alpha_K$ are mutually independent gamma random variables, each $\alpha_k \sim \text{Gamma}(c_k, b)$, their sum is also a gamma random variable $\sum_k \alpha_k \sim \text{Gamma}(\sum_k c_k, b)$. The former is equivalent to λ. It is known (e.g., [32]) that a Dirichlet random variable can be sampled by first sampling independent gamma random variables (α_k) and then dividing each of them by their sum (λ): $\theta_k = \alpha_k / \sum_{k'} \alpha_{k'}$, and, in other direction, the variables $\alpha_k = \lambda\theta_k$ are mutually independent, giving back the GP model (3).

B.2 The expectation and the variance of the document length for the GP model

From the drivations in Appendix B.1, it follows that the document length of the GP model (3) is a gamma-Poisson random variable, i.e., $L|\lambda \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(c_0, b)$. Therefore, the following follows from the law of total expectation and the law of total variance

$$\begin{aligned}
\mathbb{E}(L) &= \mathbb{E}[\mathbb{E}(L|\lambda)] = \mathbb{E}(\lambda) = c_0/b \\
\text{var}(L) &= \text{var}[\mathbb{E}(L|\lambda)] + \mathbb{E}[\text{var}(L|\lambda)] = \text{var}(\lambda) + \mathbb{E}(\lambda) = c_0/b + c_0/b^2
\end{aligned}$$

The first expression shows that the parameter b controls the expected document length $\mathbb{E}(L)$ for a given parameter c_0: the smaller b, the larger $\mathbb{E}(L)$. On the other hand, if we allow c_0 to vary as well, only the ratio c_0/b is important for the document length. We can then interpret the role of c_0 as actually controlling the concentration of the distribution for the length L (through the variance). More specifically, we have that:

$$\frac{\text{var}(L)}{(\mathbb{E}(L))^2} = \frac{1}{\mathbb{E}(L)} + \frac{1}{c_0}. \tag{23}$$

For a fixed target document length $\mathbb{E}(L)$, we can increase the variance (and thus decrease the concentration) by using a smaller c_0.

C Appendix. The cumulants of the GP and DICA models

C.1 Cumulants

For a random vector $x \in \mathbb{R}^M$, the first three cumulant tensors[8] are

$$\begin{aligned}
\mathrm{cum}(x_m) &= \mathbb{E}(x_m), \\
\mathrm{cum}(x_{m_1}, x_{m_2}) &= \mathbb{E}\left[(x_{m_1} - \mathbb{E}(x_{m_1}))(x_{m_2} - \mathbb{E}(x_{m_2})\right] = \mathrm{cov}(x_{m_1}, x_{m_2}), \\
\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}) &= \mathbb{E}\left[(x_{m_1} - \mathbb{E}(x_{m_1}))(x_{m_2} - \mathbb{E}(x_{m_2}))(x_{m_3} - \mathbb{E}(x_{m_3}))\right].
\end{aligned}$$

Note that the 2nd and 3rd cumulants coincide with the 2nd and 3rd central moments (but not for higher orders). In the following, $\mathrm{cum}(x, x, x) \in \mathbb{R}^{M \times M \times M}$ denotes the third order tensor with elements $\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3})$. Some of the properties of cumulants are listed below (see [5, chap. 5]). The most important property that motivate us to use cumulants in this paper (and the ICA literature) is the **independence** property, which says that the cumulant tensor for a random vector with independent components is diagonal (this property *does not* hold for the (non-central) moment tensors of any order, and neither for the central moments of order 4 or more).

- **Independence.** If the elements of $x \in \mathbb{R}^M$ are independent, then their cross-cumulants are zero as soon as two indices are different, i.e., $\mathrm{cum}(x_{m_1}, x_{m_2}) = \delta(m_1, m_2)\mathbb{E}[(x_{m_1} - \mathbb{E}_{m_1}))^2]$ and $\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}) = \delta(m_1, m_2, m_3)\mathbb{E}[(x_{m_1} - \mathbb{E}(x_{m_1}))^3]$, where δ is the Kronecker delta.
- **Multilinearity.** If two random vectors $y \in \mathbb{R}^M$ and $\alpha \in \mathbb{R}^K$ are linearly dependent, i.e., $y = D\alpha$ for some $D \in \mathbb{R}^{M \times K}$, then

$$\begin{aligned}
\mathrm{cum}(y_m) &= \sum_k \mathrm{cum}(\alpha_k) D_{mk}, \\
\mathrm{cum}(y_{m_1}, y_{m_2}) &= \sum_{k_1, k_2} \mathrm{cum}(\alpha_{k_1}, \alpha_{k_2}) D_{m_1 k_1} D_{m_2 k_2}, \\
\mathrm{cum}(y_{m_1}, y_{m_2}, y_{m_3}) &= \sum_{k_1, k_2, k_3} \mathrm{cum}(\alpha_{k_1}, \alpha_{k_2}, \alpha_{k_3}) D_{m_1 k_1} D_{m_2 k_2} D_{m_3 k_3},
\end{aligned}$$

  which can also be denoted[9] by

$$\begin{aligned}
\mathbb{E}(y) &= D\mathbb{E}(\alpha), \\
\mathrm{cov}(y, y) &= D\mathrm{cov}(\alpha, \alpha)D^\top, \\
\mathrm{cum}(y, y, y) &= \mathrm{cum}(\alpha, \alpha, \alpha)(D^\top, D^\top, D^\top).
\end{aligned}$$

- **The law of total cumulance.** For two random vectors $x \in \mathbb{R}^M$ and $y \in \mathbb{R}^M$, it holds

$$\begin{aligned}
\mathrm{cum}(x_m) &= \mathbb{E}\left[\mathbb{E}(x_m|y)\right], \\
\mathrm{cum}(x_{m_1}, x_{m_2}) &= \mathbb{E}\left[\mathrm{cov}(x_{m_1}, x_{m_2}|y)\right] + \mathrm{cov}\left[\mathbb{E}(x_{m_1}|y), \mathbb{E}(x_{m_2}|y)\right], \\
\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}) &= \mathbb{E}\left[\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}|y)\right] + \mathrm{cum}\left[\mathbb{E}(x_{m_1}|y), \mathbb{E}(x_{m_2}|y), \mathbb{E}(x_{m_3}|y)\right] \\
&\quad + \mathrm{cov}\left[\mathbb{E}(x_{m_1}|y), \mathrm{cov}(x_{m_2}, x_{m_3}|y)\right] \\
&\quad + \mathrm{cov}\left[\mathbb{E}(x_{m_2}|y), \mathrm{cov}(x_{m_1}, x_{m_3}|y)\right] \\
&\quad + \mathrm{cov}\left[\mathbb{E}(x_{m_3}|y), \mathrm{cov}(x_{m_1}, x_{m_2}|y)\right].
\end{aligned}$$

  Note that the first expression is also well known as the law of total expectation or the tower property, while the second one is known as the law of total covariance.

[8] Strictly speaking, the (scalar) n-th cumulant κ_n of a random variable X is defined via the cumulant-generating function $g(t)$, which is the natural logarithm of the moment-generating function, i.e $g(t) := \log \mathbb{E}\left[e^{tX}\right]$. The cumulant κ_n is then obtained from a power series expansion of the cumulant-generating function, that is $g(t) = \sum_{n=1}^{\infty} \kappa_n t^n / n!$ [Wikipedia].

[9] In [4], given a tensor $T \in \mathbb{R}^{K \times K \times K}$, $T(D^\top, D^\top, D^\top)$ is referred to as the multilinear map. In [34], the same entity is denoted by $T \times_1 D^\top \times_2 D^\top \times_3 D^\top$, where $\times_n$ denotes the n-mode tensor-matrix product.

C.2 The third cumulant of the GP/DICA models

In this section, by analogy with Section 3.1, we derive the third GP/DICA cumulant.

As the third cumulant of a Poisson random variable x_m with parameter y_m is $\mathbb{E}((x_m - \mathbb{E}(x_m))^3|y_m) = y_m$, then by the independence property of cumulants from Section C.1, the cumulant of $x|y$ is diagonal:

$$\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}|y) = \delta(m_1, m_2, m_3)\, y_{m_1}.$$

Substituting the cumulant of $x|y$ into the law of total cumulance, we obtain

$$\begin{aligned}
\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}) &= \mathbb{E}\left[\mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}|y)\right] \\
&\quad + \mathrm{cum}\left[\mathbb{E}(x_{m_1}|y), \mathbb{E}(x_{m_2}|y), \mathbb{E}(x_{m_3}|y)\right] + \mathrm{cov}\left[\mathbb{E}(x_{m_1}|y), \mathrm{cov}(x_{m_2}, x_{m_3}|y)\right] \\
&\quad + \mathrm{cov}\left[\mathbb{E}(x_{m_2}|y), \mathrm{cov}(x_{m_1}, x_{m_3}|y)\right] + \mathrm{cov}\left[\mathbb{E}(x_{m_3}|y), \mathrm{cov}(x_{m_1}, x_{m_2}|y)\right] \\
&= \delta(m_1, m_2, m_3)\mathbb{E}(y_{m_1}) + \mathrm{cum}(y_{m_1}, y_{m_2}, y_{m_3}) \\
&\quad + \delta(m_2, m_3)\mathrm{cov}(y_{m_1}, y_{m_2}) + \delta(m_1, m_3)\mathrm{cov}(y_{m_1}, y_{m_2}) + \delta(m_1, m_2)\mathrm{cov}(y_{m_1}, y_{m_3}) \\
&= \delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) + \mathrm{cum}(y_{m_1}, y_{m_2}, y_{m_3}) \\
&\quad + \delta(m_2, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) - \delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) \\
&\quad + \delta(m_1, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) - \delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) \\
&\quad + \delta(m_1, m_2)\mathrm{cov}(x_{m_1}, x_{m_3}) - \delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) \\
&= \mathrm{cum}(y_{m_1}, y_{m_2}, y_{m_3}) - 2\delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) \\
&\quad + \delta(m_2, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) + \delta(m_1, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) + \delta(m_1, m_2)\mathrm{cov}(x_{m_1}, x_{m_3}) \\
&= \left[\mathrm{cum}(\alpha, \alpha, \alpha)(D^\top, D^\top, D^\top)\right]_{m_1 m_2 m_3} - 2\delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) \\
&\quad + \delta(m_2, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) + \delta(m_1, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) + \delta(m_1, m_2)\mathrm{cov}(x_{m_1}, x_{m_3}), \quad (24)
\end{aligned}$$

where, in the third equality, we used the previous result from (9) that $\mathrm{cov}(y, y) = \mathrm{cov}(x, x) - \mathrm{diag}(\mathbb{E}(x))$.

C.3 The diagonal structure of the GP/DICA cumulants

In this section, we provide detailed derivation of the diagonal structure (11) of the matrix S (10) and the diagonal structure (13) of the tensor T (12).

From the independence of $\alpha_1, \alpha_2, \ldots, \alpha_K$ and by the independence property of cumulants from Section C.1, it follows that $\mathrm{cov}(\alpha, \alpha)$ is a diagonal matrix and $\mathrm{cum}(\alpha, \alpha, \alpha)$ is a diagonal tensor, i.e., $\mathrm{cov}(\alpha_{k_1}, \alpha_{k_2}) = \delta(k_1, k_2)\mathrm{cov}(\alpha_{k_1}, \alpha_{k_2})$ and $\mathrm{cum}(\alpha_{k_1}, \alpha_{k_2}, \alpha_{k_3}) = \delta(k_1, k_2, k_3)\mathrm{cum}(\alpha_{k_1}, \alpha_{k_1}, \alpha_{k_1})$. Therefore, the following holds

$$\begin{aligned}
\mathrm{cov}(y_{m_1}, y_{m_2}) &= \sum_k \mathrm{cov}(\alpha_k, \alpha_k) D_{m_1 k} D_{m_2 k}, \\
\mathrm{cum}(y_{m_1}, y_{m_2}, y_{m_3}) &= \sum_k \mathrm{cum}(\alpha_k, \alpha_k, \alpha_k) D_{m_1 k} D_{m_2 k} D_{m_3 k},
\end{aligned}$$

which we can rewrite in a matrix/tensor form as

$$\begin{aligned}
\mathrm{cov}(y, y) &= \sum_k \mathrm{cov}(\alpha_k, \alpha_k) d_k d_k^\top, \\
\mathrm{cum}(y, y, y) &= \sum_k \mathrm{cum}(\alpha_k, \alpha_k, \alpha_k) d_k \otimes d_k \otimes d_k.
\end{aligned}$$

Moving $\mathrm{cov}(y, y)$ / $\mathrm{cum}(y, y, y)$ in the expression for $\mathrm{cov}(x, x)$ (9) / $\mathrm{cum}(x, x, x)$ (24) on one side of equality and all other terms on the other side, we define matrix $S \in \mathbb{R}^{M \times M}$ / tensor $T \in \mathbb{R}^{M \times M \times M}$ as follows

$$S := \mathrm{cov}(x, x) - \mathrm{diag}\left(\mathbb{E}(x)\right), \quad (25)$$

$$\begin{aligned}
T_{m_1 m_2 m_3} &:= \mathrm{cum}(x_{m_1}, x_{m_2}, x_{m_3}) + 2\delta(m_1, m_2, m_3)\mathbb{E}(x_{m_1}) \\
&\quad - \delta(m_2, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) \\
&\quad - \delta(m_1, m_3)\mathrm{cov}(x_{m_1}, x_{m_2}) \\
&\quad - \delta(m_1, m_2)\mathrm{cov}(x_{m_1}, x_{m_3}). \quad (26)
\end{aligned}$$

By construction, $S = \mathrm{cov}(y, y)$ and $T = \mathrm{cum}(y, y, y)$ and, therefore, it holds that

$$S = \sum_k \mathrm{cov}(\alpha_k, \alpha_k) d_k d_k^\top, \tag{27}$$

$$T = \sum_k \mathrm{cum}(\alpha_k, \alpha_k, \alpha_k) d_k \otimes d_k \otimes d_k. \tag{28}$$

This means that both the matrix S and the tensor T are sums of rank-1 matrices and tensors, respectively[10]. This structure of the matrix S and the tensor T is the basis for the algorithms considered in this paper.

C.4 Unbiased finite sample estimators for the GP/DICA cumulants

Given a sample $\{x_1, x_2, \ldots, x_N\}$, we obtain a finite sample estimate $\widehat{S}$ of S (10) / $\widehat{T}$ of T (12) for the GP/DICA cumulants:

$$\widehat{S} := \widehat{\mathrm{cov}}(x, x) - \mathrm{diag}\left(\widehat{\mathbb{E}}(x)\right), \tag{29}$$

$$\begin{aligned}\widehat{T}_{m_1 m_2 m_3} :=& \widehat{\mathrm{cum}}(x_{m_1}, x_{m_2}, x_{m_3}) + 2\delta(m_1, m_2, m_3)\widehat{\mathbb{E}}(x_{m_1}) \\ &- \delta(m_2, m_3)\widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) \\ &- \delta(m_1, m_3)\widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) \\ &- \delta(m_1, m_2)\widehat{\mathrm{cov}}(x_{m_1}, x_{m_3}),\end{aligned} \tag{30}$$

where unbiased estimators of the first three cumulants are

$$\begin{aligned}\widehat{\mathbb{E}}(x_{m_1}) =& \frac{1}{N}\sum_n x_{nm_1}, \\ \widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) =& \frac{1}{N-1}\sum_n z_{nm_1} z_{nm_2}, \\ \widehat{\mathrm{cum}}(x_{m_1}, x_{m_2}, x_{m_3}) =& \frac{N}{(N-1)(N-2)}\sum_n z_{nm_1} z_{nm_2} z_{nm_3},\end{aligned} \tag{31}$$

where the word vocabulary indexes are $m_1, m_2, m_3 = 1, 2, \ldots, M$ and the centered documents $z_{nm} := x_{nm} - \widehat{\mathbb{E}}(x_m)$. (The latter is introduced only for compact representation of (31) and is different from z in the LDA model.)

C.5 On the orders of cumulants

Note that the factorization of $S = \widetilde{D}\widetilde{D}^\top$ does not uniquely determine $\widetilde{D}$ as one can equivalently use $S = (\widetilde{D}U)(\widetilde{D}U)^\top$ with any orthogonal $K \times K$ matrix U. Therefore, one has to consider higher than the second order information. Moreover, in ICA the fourth-order tensors are used, because the third cumulant of the Gaussian distribution is zero, which is not the case in the DICA/LDA models, where the third order information is sufficient.

D Appendix. The sketch of the proof for Proposition 3.1

D.1 Expected squared error for the sample expectation

The sample expectation is $\widehat{\mathbb{E}}(x) = \frac{1}{N}\sum_n x_n$ is an unbiased estimator of the expectation and:

$$\begin{aligned}\mathbb{E}\left(\|\widehat{\mathbb{E}}(x) - \mathbb{E}(x)\|_2^2\right) &= \sum_m \mathbb{E}\left[\left(\widehat{\mathbb{E}}(x_m) - \mathbb{E}(x_m)\right)^2\right] \\ &= \frac{1}{N^2}\sum_m \left[\mathbb{E}\left(\sum_n \left(x_{nm} - \mathbb{E}(x_m)\right)^2\right) + \mathbb{E}\left(\sum_n \sum_{n \neq n'} \left(x_{nm} - \mathbb{E}(x_m)\right)\left(x_{n'm} - \mathbb{E}(x_m)\right)\right)\right] \\ &= \frac{1}{N}\sum_m \mathbb{E}\left[\left(x_m - \mathbb{E}(x_m)\right)^2\right] = \frac{1}{N}\sum_m \mathrm{var}(x_m).\end{aligned}$$

[10] For tensors, such decomposition is also known under the names CANDECOMP/PARAFAC or, simply, the CP decomposition (see, e.g., [34]).

Further, by the law of total variance:

$$\mathbb{E}\left(\|\widehat{\mathbb{E}}(x)-\mathbb{E}(x)\|_2^2\right)=\frac{1}{N}\sum_m\left[\mathbb{E}(\mathrm{var}(x_m|y))+\mathrm{var}(\mathbb{E}(x_m|y))\right]=\frac{1}{N}\sum_m\left[\mathbb{E}(y_m)+\mathrm{var}(y_m)\right]$$
$$=\frac{1}{N}\left[\sum_k\mathbb{E}(\alpha_k)+\sum_k\langle d_k,d_k\rangle\mathrm{var}(\alpha_k)\right],$$

using the fact that $\sum_m D_{mk}=1$ for any k.

D.2 Expected squared error for the sample covariance

The following finite sample estimator of the covariance $\mathrm{cov}(x,x)=\mathbb{E}(xx^\top)-\mathbb{E}(x)\mathbb{E}(x)^\top$

$$\widehat{\mathrm{cov}}(x,x)=\frac{1}{N-1}\sum_n x_nx_n^\top-\widehat{\mathbb{E}}(x)\widehat{\mathbb{E}}(x)^\top=\frac{1}{N-1}\sum_n\left(x_nx_n^\top-\frac{1}{N^2}\sum_{n'}\sum_{n''}x_{n'}x_{n''}^\top\right)$$
$$=\frac{1}{N}\sum_n\left(x_nx_n^\top-\frac{1}{N-1}x_n\sum_{n'\neq n}x_{n'}^\top\right) \tag{32}$$

is unbiased, i.e., $\mathbb{E}(\widehat{\mathrm{cov}}(x,x))=\mathrm{cov}(x,x)$. Its squared error is

$$\mathbb{E}\left(\|\widehat{\mathrm{cov}}(x,x)-\mathrm{cov}(x,x)\|_F^2\right)=\sum_{m,m'}\mathbb{E}\left[(\widehat{\mathrm{cov}}(x_m,x_{m'})-\mathbb{E}[\widehat{\mathrm{cov}}(x_m,x_{m'})])^2\right].$$

The m,m'-th element of the sum above is equal to

$$\frac{1}{N^2}\sum_{n,n'}\mathrm{cov}\left(x_{nm}x_{nm'}-\frac{1}{N-1}x_{nm}\sum_{n''\neq n}x_{n''m'}\quad,\quad x_{n'm}x_{n'm'}-\frac{1}{N-1}x_{n'm}\sum_{n'''\neq n'}x_{n'''m'}\right)$$
$$=\frac{1}{N^2}\sum_{n,n'}\mathrm{cov}\left(x_{nm}x_{nm'},x_{n'm}x_{n'm'}\right)-\frac{2}{N^2(N-1)}\sum_{n,n'}\mathrm{cov}\left(x_{nm}\sum_{n''\neq n}x_{n''m'},x_{n'm}x_{n'm'}\right)$$
$$+\frac{1}{N^2(N-1)^2}\sum_{n,n'}\mathrm{cov}\left(x_{nm}\sum_{n''\neq n}x_{n''m'},x_{n'm}\sum_{n'''\neq n'}x_{n'''m'}\right)$$
$$=\frac{1}{N^2}\sum_n\mathrm{cov}\left(x_{nm}x_{nm'},x_{nm}x_{nm'}\right)$$
$$-\frac{2}{N^2(N-1)}\left[\sum_n\sum_{n''\neq n}\mathrm{cov}\left(x_{nm}x_{n''m'},x_{nm}x_{nm'}\right)+\sum_n\sum_{n'\neq n}\mathrm{cov}\left(x_{nm}x_{n'm'},x_{n'm}x_{n'm'}\right)\right]$$
$$+\frac{1}{N^2(N-1)^2}\left[\sum_n\sum_{n''\neq n}\sum_{n'''\neq n}\mathrm{cov}\left(x_{nm}x_{n''m'},x_{nm}x_{n'''m'}\right)+\sum_{n'}\sum_{n\neq n'}\sum_{n''\neq n}\mathrm{cov}\left(x_{nm}x_{n''m'},x_{n'm}x_{nm'}\right)\right]$$
$$+\frac{1}{N^2(N-1)^2}\left[\sum_{n'}\sum_{n\neq n'}\sum_{n'''\neq n'}\mathrm{cov}\left(x_{nm}x_{n'm'},x_{n'm}x_{n'''m'}\right)+\sum_{n'}\sum_{n\neq n'}\sum_{n''\neq n}\mathrm{cov}\left(x_{nm}x_{n''m'},x_{n'm}x_{n''m'}\right)\right],$$

where we used mutual independence of the observations x_n in a sample $\{x_n\}_{n=1}^N$ to conclude that the covariance between the two expressions involving only independent variables is zero.

Further:

$$\mathbb{E}\left(\|\widehat{\mathrm{cov}}(x,x)-\mathrm{cov}(x,x)\|_F^2\right)=\frac{1}{N^2}\sum_{m,m'}N\left(\mathbb{E}(x_m^2x_{m'}^2)-[\mathbb{E}(x_mx_{m'})]^2\right)$$
$$-\frac{4}{N^2(N-1)}\sum_{m,m'}N(N-1)\left(\mathbb{E}(x_m^2x_{m'})\mathbb{E}(x_{m'})-\mathbb{E}(x_mx_{m'})\mathbb{E}(x_m)\mathbb{E}(x_{m'})\right)$$
$$+\frac{2}{N^2(N-1)^2}\sum_{m,m'}N(N-1)(N-2)\left(\mathbb{E}(x_m^2)\left[\mathbb{E}(x_{m'})\right]^2-\left[\mathbb{E}(x_m)\right]^2\left[\mathbb{E}(x_{m'})\right]^2\right)$$
$$+\frac{2}{N^2(N-1)^2}\sum_{m,m'}N(N-1)(N-2)\left(\mathbb{E}(x_mx_{m'})\mathbb{E}(x_m)\mathbb{E}(x_{m'})-\left[\mathbb{E}(x_m)\right]^2\left[\mathbb{E}(x_{m'})\right]^2\right)+O\left(\frac{1}{N^2}\right),$$

which after simplification gives

$$\mathbb{E}\left(\|\widehat{\mathrm{cov}}(x,x)-\mathrm{cov}(x,x)\|_F^2\right)=\frac{1}{N}\sum_{m,m'}\left[\mathrm{var}(x_mx_{m'})+2\left[\mathbb{E}(x_m)\right]^2\mathrm{var}(x_{m'})\right]$$
$$+\frac{1}{N}\sum_{m,m'}[2\mathbb{E}(x_m)\mathbb{E}(x_{m'})\mathrm{cov}(x_m,x_{m'})-4\mathbb{E}(x_m)\mathrm{cov}(x_mx_{m'},x_{m'})]+O\left(\frac{1}{N^2}\right),$$

where in the last equality, by symmetry, the summation indexes m and m' can be exchanged. As $x_m\sim\mathrm{Poisson}(y_m)$, by the law of total expectation and law of total covariance, it follows, for $m\neq m'$ (and using the auxiliary expressions from Section D.4):

$$\begin{aligned}
\mathrm{var}(x_mx_{m'})&=\mathbb{E}(x_m^2x_{m'}^2)-[\mathbb{E}[x_mx_{m'}]]^2=\mathbb{E}\left[\mathbb{E}(x_m^2x_{m'}^2|y)\right]-\left[\mathbb{E}\left[\mathbb{E}(x_mx_{m'}|y)\right]\right]^2\\
&=\mathbb{E}\left[y_m^2y_{m'}^2+y_m^2y_{m'}+y_my_{m'}^2+y_my_{m'}\right]-\left[\mathbb{E}(y_my_{m'})\right]^2,\\
\left[\mathbb{E}(x_m)\right]^2\mathrm{var}(x_{m'})&=\left[\mathbb{E}(y_m)\right]^2\mathbb{E}(y_{m'})+\left[\mathbb{E}(y_m)\right]^2\mathbb{E}(y_{m'}^2)-\left[\mathbb{E}(y_m)\right]^2\left[\mathbb{E}(y_{m'})\right]^2,\\
\mathbb{E}(x_m)\mathbb{E}(x_{m'})\mathrm{cov}(x_m,x_{m'})&=\mathbb{E}(y_my_{m'})\mathbb{E}(y_m)\mathbb{E}(y_{m'})-\left[\mathbb{E}(y_m)\right]^2\left[\mathbb{E}(y_{m'})\right]^2,\\
\mathbb{E}(x_m)\mathrm{cov}(x_mx_{m'},x_{m'})&=\mathbb{E}(y_m)\left[\mathbb{E}(y_my_{m'})+\mathbb{E}(y_my_{m'}^2)-\mathbb{E}(y_my_{m'})\mathbb{E}(y_{m'})\right].
\end{aligned}$$

Now, considering the $m=m'$ case, we have:

$$\begin{aligned}
\mathrm{var}(x_m^2)&=\mathbb{E}[\mathbb{E}(x_m^4|y)]-\left[\mathbb{E}[\mathbb{E}(x_m^2|y)]\right]^2\\
&=\mathbb{E}\left[y_m^4+6y_m^3+7y_m^2+y_m\right]-\left[\mathbb{E}\left[y_m^2+y_m\right]\right]^2,\\
\mathbb{E}(x_m)\mathbb{E}(x_m)\mathrm{cov}(x_m,x_m)&=\mathbb{E}(y_m)^2\left[\mathbb{E}(y_m^2)+\mathbb{E}(y_m)-\left[\mathbb{E}(y_m)\right]^2\right],\\
\mathbb{E}(x_m)\mathrm{cov}(x_m^2,x_m)&=\mathbb{E}(y_m)\left[\mathbb{E}(y_m^3)+3\mathbb{E}(y_m^2)+\mathbb{E}(y_m)-\mathbb{E}(y_m)\left[\mathbb{E}(y_m^2)+\mathbb{E}(y_m)\right]\right].
\end{aligned}$$

Substitution of $y_m=\sum_k D_{mk}\alpha_k$ gives the following

$$\begin{aligned}
\mathbb{E}\left(\|\widehat{\mathrm{cov}}(x,x)-\mathrm{cov}(x,x)\|_F^2\right)&=\frac{1}{N}\sum_{k,k',k'',k'''}\langle d_k,d_{k'}\rangle\langle d_{k''},d_{k'''}\rangle\mathcal{A}_{kk'k''k'''}\\
&+\frac{1}{N}\sum_{k,k',k''}\left[\langle d_k,d_{k'}\rangle\langle d_{k''},\vec{1}\rangle\mathcal{B}_{kk'k''}+\langle d_k\circ d_{k'},d_{k''}\rangle\mathcal{E}_{kk'k''}\right]\\
&+\frac{1}{N}\sum_{k,k'}\left[\langle d_k,\vec{1}\rangle\langle d_{k'},\vec{1}\rangle\mathbb{E}(\alpha_k\alpha_{k'})+\langle d_k,d_{k'}\rangle\mathcal{F}_{kk'}\right]\\
&+\sum_k\langle d_k,\vec{1}\rangle\mathbb{E}(\alpha_k)+O\left(\frac{1}{N^2}\right),
\end{aligned}$$

where $\vec{1}$ is the vector with all the elements equal to 1 and

$$\begin{aligned}
\mathcal{A}_{kk'k''k'''}&=\mathbb{E}(\alpha_k\alpha_{k'}\alpha_{k''}\alpha_{k'''})-\mathbb{E}(\alpha_k\alpha_{k''})\mathbb{E}(\alpha_{k'}\alpha_{k'''})+2\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'})\mathbb{E}(\alpha_{k''}\alpha_{k'''})\\
&-2\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'})\mathbb{E}(\alpha_{k''})\mathbb{E}(\alpha_{k'''})+2\mathbb{E}(\alpha_k\alpha_{k''})\mathbb{E}(\alpha_{k'})\mathbb{E}(\alpha_{k'''})-2\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'})\mathbb{E}(\alpha_{k''})\mathbb{E}(\alpha_{k'''})\\
&-4\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'}\alpha_{k''}\alpha_{k'''})+4\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'}\alpha_{k''})\mathbb{E}(\alpha_{k'''}),\\
\mathcal{B}_{kk'k''}&=2\mathbb{E}(\alpha_k\alpha_{k'}\alpha_{k''})+2\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'})\mathbb{E}(\alpha_{k''})-4\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'}\alpha_{k''}),\\
\mathcal{E}_{kk'k''}&=4\mathbb{E}(\alpha_k\alpha_{k'}\alpha_{k''})+6\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'})\mathbb{E}(\alpha_{k''})-10\mathbb{E}(\alpha_k\alpha_{k'})\mathbb{E}(\alpha_{k''}),\\
\mathcal{F}_{kk'}&=6\mathbb{E}(\alpha_k\alpha_{k'})-5\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'}),
\end{aligned}$$

where we used the expressions from Section D.4.

D.3 Expected squared error of the estimator $\widehat{\mathbf{S}}$ for the GP/DICA cumulants

As the estimator $\widehat{S}$ (29) of S (10) is unbiased, its expected squared error is

$$\begin{aligned}\mathbb{E}\left[\|\widehat{S}-S\|_F^2\right] =& \mathbb{E}\left[\left\|(\widehat{\mathrm{cov}}(x,x)-\mathrm{cov}(x,x))+\left(\mathrm{diag}[\widehat{\mathbb{E}}(x)]-\mathrm{diag}\left[\mathbb{E}(x)\right]\right)\right\|_F^2\right] \\ =& \mathbb{E}\left[\|\widehat{\mathbb{E}}(x)-\mathbb{E}(x)\|_F^2\right]+\mathbb{E}\left[\|\widehat{\mathrm{cov}}(x,x)-\mathrm{cov}(x,x)\|_F^2\right] \\ &+2\sum_m \mathbb{E}\left[\left(\widehat{\mathbb{E}}(x_m)-\mathbb{E}(x_m)\right)\left(\widehat{\mathrm{cov}}(x_m,x_m)-\mathrm{cov}(x_m,x_m)\right)\right].\end{aligned} \tag{33}$$

As $\widehat{\mathbb{E}}(x_m)$ and $\widehat{\mathrm{cov}}(x_m,x_m)$ are unbiased, the m-th element of the last sum is equal to

$$\begin{aligned}&\mathrm{cov}\left[\widehat{\mathbb{E}}(x_m),\widehat{\mathrm{cov}}(x_m,x_m)\right] \\ &=\frac{1}{N^2}\sum_{n,n'}\mathrm{cov}\left[x_{nm},x_{n'm}^2\right]-\frac{1}{N^2(N-1)}\sum_{n,n',n''\neq n'}\mathrm{cov}\left[x_{nm},x_{n'm}x_{n''m}\right] \\ &=\frac{1}{N^2}\sum_{n}\mathrm{cov}\left[x_{nm},x_{nm}^2\right]-\frac{2}{N^2(N-1)}\sum_{n,n'\neq n}\mathrm{cov}\left[x_{nm},x_{n'm}x_{nm}\right]+O\left(\frac{1}{N^2}\right) \\ &=\frac{1}{N}\mathbb{E}(x_m^3)-\frac{2}{N}\left(\mathbb{E}(x_m^2)\mathbb{E}(x_m)-\left[\mathbb{E}(x_m)\right]^3\right)+O\left(\frac{1}{N^2}\right) \\ &\leq\frac{1}{N}\mathbb{E}(x_m^3)+\frac{2}{N}\left[\mathbb{E}(x_m)\right]^3+O\left(\frac{1}{N^2}\right)=\frac{1}{N}\left[\mathbb{E}(y_m^3)+3\mathbb{E}(y_m^2)+\mathbb{E}(y_m)+2\left[\mathbb{E}(y_m)\right]^3\right]+O\left(\frac{1}{N^2}\right),\end{aligned}$$

where we neglected the negative term $-\mathbb{E}(x_m^2)\mathbb{E}(x_m)$ for the inequality, and the last equality follows from the expressions in Section D.4. Further, the fact that $y_m=\sum_k D_{mk}\alpha_k$ gives

$$\begin{aligned}\sum_m \mathrm{cov}\left[\widehat{\mathbb{E}}(x_m),\widehat{\mathrm{cov}}(x_m,x_m)\right] =& \frac{1}{N}\sum_{k,k',k''}\langle d_k\circ d_{k'},d_{k''}\rangle\mathcal{C}_{kk'k''} \\ &+\frac{3}{N}\sum_{k,k'}\langle d_k,d_{k'}\rangle\mathbb{E}(\alpha_k\alpha_{k'})+\frac{1}{N}\sum_k\langle d_k,\vec{1}\rangle\mathbb{E}(\alpha_k)+O\left(\frac{1}{N^2}\right),\end{aligned}$$

where $\circ$ denotes the elementwise Hadamard product and

$$\mathcal{C}_{kk'k''}=\mathbb{E}(\alpha_k\alpha_{k'}\alpha_{k''})+2\mathbb{E}(\alpha_k)\mathbb{E}(\alpha_{k'})\mathbb{E}(\alpha_{k''}).$$

Plugging this and the expressions for $\mathbb{E}(\|\widehat{\mathbb{E}}(x)-\mathbb{E}(x)\|_F^2)$ and $\mathbb{E}(\|\widehat{\mathrm{cov}}(x,x)-\mathrm{cov}(x,x)\|_F^2)$ from Sections D.1 and D.2, respectively, into (33) gives

$$\begin{aligned}\mathbb{E}\left[\|\widehat{S}-S\|_F^2\right] =& \frac{1}{N}\left[\sum_k\langle d_k,d_k\rangle\mathrm{var}(\alpha_k)+\sum_k\mathbb{E}(\alpha_k)+\sum_{k,k',k'',k'''}\langle d_k,d_{k'}\rangle\langle d_{k''},d_{k'''}\rangle\mathcal{A}_{kk'k''k'''}\right]+O\left(\frac{1}{N^2}\right) \\ &+\frac{1}{N}\left[\sum_{k,k',k''}\left[\langle d_k,d_{k'}\rangle\mathcal{B}_{kk'k''}+2\langle d_k\circ d_{k'},d_{k''}\rangle\mathcal{C}_{kk'k''}\right]+\sum_{k,k'}\left(1+6\langle d_k,d_{k'}\rangle\right)\mathbb{E}(\alpha_k\alpha_{k'})+2\sum_k\mathbb{E}(\alpha_k)\right],\end{aligned}$$

where we used that, by the simplex constraint on the topics, $\langle d_k,\vec{1}\rangle=1$ for all k. To analyze this expression in more details, let us now consider the GP model, i.e., $\alpha_k\sim\mathrm{Gamma}(c_k,b)$:

$$\sum_{k,k',k'',k'''}\mathcal{A}_{kk'k''k'''}\leq\frac{30c_0^4+23c_0^3+14c_0^2+8c_0}{b^4},\quad\text{and}\quad\sum_{k,k',k''}\mathcal{B}_{kk'k''}\leq\frac{6c_0^3+10c_0^2+4c_0}{b^3},$$

$$\sum_{k,k',k''}\mathcal{C}_{kk'k''}\leq\frac{7c_0^3+6c_0^2+2c_0}{b^3},\quad\text{and}\quad\sum_{k,k',k''}\mathcal{E}_{kk'k''}\leq\frac{12c_0^3+10c_0^2+8c_0}{b^3},$$

$$\sum_{k,k'}\mathcal{F}_{kk'}\leq\frac{2c_0^2+c_0}{b^2}\quad\text{and}\quad\sum_{k,k'}\mathbb{E}(\alpha_k\alpha_{k'})\leq\frac{2c_0^2+c_0}{b^2},$$

where we used the expressions from Section D.4, which gives

$$\mathbb{E}\left[\|\widehat{S}-S\|_F^2\right] \leq \frac{\nu}{N}\left[\max_k \|d_k\|_2^2 \frac{c_0}{b^2} + \frac{c_0}{b} + \left(\max_{k,k'}\langle d_k, d_{k'}\rangle\right)^2 \max\left[\frac{c_0^4}{b^4}, \frac{c_0}{b^4}\right] + \max_{k,k'}\langle d_k, d_{k'}\rangle \max\left[\frac{c_0^3}{b^3}, \frac{c_0}{b^3}\right]\right]$$
$$+ \frac{\nu}{N}\left[\left(\max_{k,k',k''}\langle d_k \circ d_{k'}, d_{k''}\rangle\right)\max\left[\frac{c_0^3}{b^3}, \frac{c_0}{b^3}\right] + \left(1+\max_{k,k'}\langle d_k, d_{k'}\rangle\right)\max\left[\frac{c_0^2}{b^2}, \frac{c_0}{b^2}\right]\right] + O\left(\frac{1}{N^2}\right),$$

where $\nu \leq 30$ is a universal constant. As, by the Cauchy-Schwarz inequality, $\max_{k,k'}\langle d_k, d_{k'}\rangle \leq \max_k \|d_k\|_2^2 =: \Delta_1$ and $\max_{k,k',k''}\langle d_k \circ d_{k'}, d_{k''}\rangle \leq \max_k \|d_k\|_\infty \|d_k\|_2^2 \leq \max_k \|d_k\|_2^3 =: \Delta_2$ (note that for the topics in the simplex, $\Delta_2 \leq \Delta_1$ as well as $\Delta_1^2 \leq \Delta_1$), it follows that

$$\mathbb{E}\left[\|\widehat{S}-S\|_F^2\right] \leq \frac{\nu}{N}\left[\Delta_1\left(\frac{L^2}{\bar{c}_0}+\frac{L^3}{\bar{c}_0^2}\right) + L + \Delta_1^2\frac{L^4}{\bar{c}_0^3} + \frac{L^2}{\bar{c}_0^2} + \Delta_2\frac{L^3}{\bar{c}_0^2}\right] + O\left(\frac{1}{N^2}\right)$$
$$\leq \frac{2\nu}{N}\frac{1}{\bar{c}_0^3}\left[\Delta_1^2 L^4 + \bar{c}_0\Delta_1 L^3 + \bar{c}_0^2 L^2 + \bar{c}_0^3 L\right] + O\left(\frac{1}{N^2}\right),$$

where $\bar{c}_0 = \min(1, c_0) \leq 1$ and, from Section B.2, $c_0 = bL$ where L is the expected document length. The second term $\bar{c}_0\Delta_1 L^3$ cannot be dominant as the system $\bar{c}_0\Delta_1 L^3 > \bar{c}_0^2 L^2$ and $\bar{c}_0\Delta_1 L^3 > \Delta_1^2 L^4$ is infeasible. Also, with the reasonable assumption that $L \geq 1$, we also have that the 4th term $\bar{c}_0^3 L \leq \bar{c}_0^2 L^2$. Therefore,

$$\mathbb{E}\left[\|\widehat{S}-S\|_F^2\right] \leq \frac{3\nu}{N}\max\left[\Delta_1^2 L^4,\, \bar{c}_0^2 L^2\right] + O\left(\frac{1}{N^2}\right).$$

D.4 Auxiliary expressions

As $\{x_m\}_{m=1}^M$ are conditionally independent given y in the DICA model (3), we have the following expressions by using the law of total expectation for $m \neq m'$ and using the moments of the Poisson distribution with parameter y_m:

$$\begin{aligned}
\mathbb{E}(x_m) &= \mathbb{E}[\mathbb{E}(x_m|y_m)] = \mathbb{E}(y_m),\\
\mathbb{E}(x_m^2) &= \mathbb{E}[\mathbb{E}(x_m^2|y_m)] = \mathbb{E}(y_m^2) + \mathbb{E}(y_m),\\
\mathbb{E}(x_m^3) &= \mathbb{E}[\mathbb{E}(x_m^3|y_m)] = \mathbb{E}(y_m^3) + 3\mathbb{E}(y_m^2) + \mathbb{E}(y_m),\\
\mathbb{E}(x_m^4) &= \mathbb{E}[\mathbb{E}(x_m^4|y_m)] = \mathbb{E}(y_m^4) + 6\mathbb{E}(y_m^3) + 7\mathbb{E}(y_m^2) + \mathbb{E}(y_m),\\
\mathbb{E}(x_m x_{m'}) &= \mathbb{E}[\mathbb{E}(x_m x_{m'}|y)] = \mathbb{E}[\mathbb{E}(x_m|y_m)\mathbb{E}(x_{m'}|y_{m'})] = \mathbb{E}(y_m y_{m'}),\\
\mathbb{E}(x_m x_{m'}^2) &= \mathbb{E}[\mathbb{E}(x_m x_{m'}^2|y)] = \mathbb{E}[\mathbb{E}(x_m|y_m)\mathbb{E}(x_{m'}^2|y_{m'})] = \mathbb{E}(y_m y_{m'}^2) + \mathbb{E}(y_m y_{m'}),\\
\mathbb{E}(x_m^2 x_{m'}^2) &= \mathbb{E}[\mathbb{E}(x_m^2|y_m)\mathbb{E}(x_{m'}^2|y_{m'})] = \mathbb{E}(y_m^2 y_{m'}^2) + \mathbb{E}(y_m^2 y_{m'}) + \mathbb{E}(y_m y_{m'}^2) + \mathbb{E}(y_m y_{m'}).
\end{aligned}$$

Moreover, the moments of $\alpha_k \sim \mathrm{Gamma}(c_k, b)$ are

$$\mathbb{E}(\alpha_k) = \frac{c_k}{b}, \quad \mathbb{E}(\alpha_k^2) = \frac{c_k^2+c_k}{b^2}, \quad \mathbb{E}(\alpha_k^3) = \frac{c_k^3+3c_k^2+2c_k}{b^3}, \quad \mathbb{E}(\alpha_k^4) = \frac{c_k^4+6c_k^3+11c_k^2+6c_k}{b^4}, \quad \text{etc.}$$

D.5 Analysis of the whitening and recovery error

We can follow a similar analysis as in Appendix C of [15] to derive the topic recovery error given the sample estimate error. In particular, if we define the following sampling errors E_S and E_T:

$$\|\widehat{S}-S\| \leq E_S,$$
$$\|\widehat{T}(u)-T(u)\| \leq \|u\|_2 E_T,$$

then the following form of their Lemma C.2 holds for both the LDA moments and the GP/DICA cumulants:

$$\|\widehat{W}\widehat{T}(\widehat{W}^\top u)\widehat{W}^\top - WT(W^\top u)W^\top\| \leq \nu\left[\frac{(\max_k\gamma_k)E_S}{\sigma_K(\widetilde{D})^2} + \frac{E_T}{\sigma_K(\widetilde{D})^3}\right], \tag{34}$$

where $\sigma_k(\cdot)$ denotes the k-th singular value of a matrix, ν is some universal constant, and in both cases $\widetilde{D}$ was defined such that $S = \widetilde{D}\widetilde{D}^\top$. For the LDA moments, $\gamma_k = 2\sqrt{\frac{c_0(c_0+1))}{c_k(c_0+2)^2}}$, whereas for the GP/DICA cumulants, γ_k takes the simpler form $\gamma_k := \mathrm{cum}(\alpha_k)/[\mathrm{var}(\alpha_k)]^{3/2} = 2/\sqrt{c_k}$.

We note that the scaling for S is $O(L^2)$ for the GP/DICA cumulants, in contrast to $O(1)$ for the LDA moments. Thus, to compare the upper bound (34) for the two types of moments, we need to put it in quantities which are common. In the first section of the Appendix C of [15], it was mentioned that $\sigma_K\big(\widetilde{D}\big) \geq \sqrt{\frac{c_{\min}}{c_0(c_0+1)}}\sigma_K(D)$ for the LDA moments, where $c_{\min} := \min_k c_k$. In contrast, for the GP/DICA cumulants, we can show that $\sigma_K\big(\widetilde{D}\big) \geq L\frac{\sqrt{c_{\min}}}{c_0}\sigma_K(D)$, where $L := c_0/b$ is the average length of a document in the GP model. Using this lower bound for the singular vector, we thus get the following bound in the case of the GP cumulant:

$$\|\widehat{W}\widehat{T}(\widehat{W}^\top u)\widehat{W}^\top - WT(W^\top u)W^\top\| \leq \frac{\nu}{c_{\min}^{3/2}}\left[\frac{E_S}{L^2}\frac{2c_0^2}{\big[\sigma_K\big(D\big)\big]^2} + \frac{E_T}{L^3}\frac{c_0^3}{[\sigma_K(D)]^3}\right]. \tag{35}$$

The $c_{\min}^{3/2}$ factor is common for both the LDA moment and GP cumulant, but as we mentioned after Proposition 3.1, the sample error E_S term gets divided by L^2 for the GP cumulant, as expected.

The recovery error bound in [15] is based on the bound (35), and thus by showing that the error E_S/L^2 for the GP cumulant is lower than the E_S term for the LDA moment, we expect to also gain a similar gain for the recovery error, as the rest of the argument is the same for both types of moments (see Appendix C.2, C.3 and C.4 in [15] for the completion).

E Appendix. The LDA moments

E.1 Our notation

The LDA moments were derived in [3]. Note that the full version of the paper with proofs appeared in [15] and a later version of the paper also appeared in [31]. In this section, we recall the form of the LDA moments using our notation. This section does not contain any novel results and is included for the reader's convenience. We also refer to this section when deriving the practical expressions for computation of the sample estimates of the LDA moments in Appendix F.4.

For deriving the LDA moments, a document is assumed to be composed of at least three tokens: $L \geq 3$. As the LDA generative model (1) is only defined *conditional* on the length L, this is not too problematic. But given that we present models in this paper which also model L, we mention for clarity that we can suppose that all expectations and probabilities defined below are implicitly conditioning on $L \geq 3$.[11] The theoretical LDA moments are derived only using the first three words w_1, w_2 and w_3 of a document. But note that since the words w_ℓ's are conditionally i.i.d. given θ (for $1 \leq \ell \leq L$), we have $M_3 := \mathbb{E}(w_1 \otimes w_2 \otimes w_3) = \mathbb{E}(w_{\ell_1} \otimes w_{\ell_2} \otimes w_{\ell_3})$ for any three distinct tokens ℓ_1, ℓ_2 and ℓ_3. The tensor M_3 is thus symmetric, and could have been defined using any distinct ℓ_1, ℓ_2 and ℓ_3 that are less than L. To highlight this arbitrary choice and to make the links with the U-statistics estimator presented later, we thus use generic distinct ℓ_1, ℓ_2 and ℓ_3 in the definition of the LDA moments below, instead of $\ell_1 = 1$, $\ell_2 = 2$ and $\ell_3 = 1$ as in [3].

[11] Note that another advantage of the DICA cumulants from Section 3.1 is that they do not require such a somewhat artificial condition: they are well-defined for any document length (even a document of length zero!).

Using this notation, then by the law of total expectation and the properties of the Dirichlet distribution, the non-central moments[12] of the LDA model (1) take the form [3]:

$$M_1 = \mathbb{E}(w_{\ell_1}) = D\frac{c}{c_0}, \tag{36}$$

$$M_2 = \mathbb{E}(w_{\ell_1} w_{\ell_2}^\top) = \frac{c_0}{c_0+1} M_1 M_1^\top + \frac{1}{c_0(c_0+1)} D\,\mathrm{diag}\,(c)\, D^\top, \tag{37}$$

$$\begin{aligned} M_3 &= \mathbb{E}(w_{\ell_1} \otimes w_{\ell_2} \otimes w_{\ell_3}) \\ &= \frac{c_0}{c_0+2}\left[\mathbb{E}(w_{\ell_1} \otimes w_{\ell_2} \otimes M_1) + \mathbb{E}(w_{\ell_1} \otimes M_1 \otimes w_{\ell_3}) + \mathbb{E}(M_1 \otimes w_{\ell_2} \otimes w_{\ell_3})\right], \\ &\quad - \frac{2c_0^3}{c_0(c_0+1)(c_0+2)} M_1 \otimes M_1 \otimes M_1 + \frac{2}{c_0(c_0+1)(c_0+2)} \sum_{k=1}^K c_k d_k \otimes d_k \otimes d_k. \end{aligned} \tag{38}$$

where $\otimes$ denotes the tensor product.

Similarly to the GP/DICA cumulants (as discussed in Appendix C.3), moving the terms in the non-central moments (36), (37), (38), the following quantities are defined

$$(Pairs) = S := M_2 - \frac{c_0}{c_0+1} M_1 M_1^\top, \qquad \text{LDA S-moment} \tag{39}$$

$$\begin{aligned} (Triples) = T := M_3 &- \frac{c_0}{c_0+2}\left[\mathbb{E}(w_{\ell_1} \otimes w_{\ell_2} \otimes M_1) + \mathbb{E}(w_{\ell_1} \otimes M_1 \otimes w_{\ell_3}) + \mathbb{E}(M_1 \otimes w_{\ell_2} \otimes w_{\ell_3})\right] \\ &+ \frac{2c_0^2}{(c_0+1)(c_0+2)} M_1 \otimes M_1 \otimes M_1. \qquad \text{LDA T-moment} \end{aligned} \tag{40}$$

Slightly abusing terminology, we refer to the entities S and T as the "LDA moments". They have the following diagonal structure

$$S = \frac{1}{c_0(c_0+1)} \sum_{k=1}^K c_k d_k d_k^\top, \tag{41}$$

$$T = \frac{2}{c_0(c_0+1)(c_0+2)} \sum_{k=1}^K c_k d_k \otimes d_k \otimes d_k. \tag{42}$$

Note however that this form of the LDA moments has a slightly different nature than the similar form (11) and (13) of the GP/DICA cumulants. Indeed, the former is the result of properties of the Dirichlet distribution, while the latter is the result of the independence of α's. However, one can think of the elements of a Dirichlet random vector as being almost independent (as, e.g., a Dirichlet random vector can be obtained from independent gamma variables through dividing each by their sum). Also, this closeness of the structures of the LDA moments and the GP cumulants can be explained by the closeness of the respective models as discussed in Section 2.

E.2 Asymptotically unbiased finite sample estimators for the LDA moments

Given realizations $w_{n\ell}$, $n = 1, \ldots, N$, $\ell = 1, \ldots, L_n$, of the token random variable w_ℓ, we now give the expressions for the finite sample estimates of S (39) and T (40) for the LDA model (and we rewrite them as a function of the sample counts x_n).[13] We use the notation $\widehat{\mathbb{E}}$ below to express a U-statistics empirical expectation over the token within a documents, uniformly averaged over the whole corpus. For example, $\widehat{\mathbb{E}}(w_{\ell_1} \otimes w_{\ell_2} \otimes \widehat{M_1}) := \frac{1}{N}\sum_{n=1}^N \frac{1}{L_n(L_n-1)} \sum_{\ell_1=1}^{L_n} \sum_{\substack{\ell_2=1 \\ \ell_2 \neq \ell_1}}^{L_n} w_{\ell_1} \otimes w_{\ell_2} \otimes$

[12] Note, the difference in the notation for the LDA moments in papers [3] and [4]. In [3], $M_1 = \mathbb{E}(w_{\ell_1})$, $M_2 = \mathbb{E}(w_{\ell_1} \otimes w_{\ell_2})$, and $M_3 = \mathbb{E}(w_{\ell_1} \otimes w_{\ell_2} \otimes w_{\ell_3})$. However, in [4], M_2 is equivalent to S in our notation and to $Pairs$ in the notation of [3]; similarly, M_3 is T in our notation or $Triples$ in the notation of [3].

[13] Note that because non-linear functions of $\widehat{M_1}$ appear in the expression for $\widehat{S}$ (43) and $\widehat{T}$ (44), the estimator is biased, i.e., $\mathbb{E}(\widehat{S}) \neq S$. The bias is small though: $\|\mathbb{E}(\widehat{S}) - S\| = O(1/N)$ and the estimator is asymptotically unbiased. This is in contrast with the estimator for the GP/DICA moments which is easily made unbiased.

$\widehat{M}_1$.

$$\widehat{S} := \widehat{M}_2 - \frac{c_0}{c_0+1}\widehat{M}_1\widehat{M}_1^\top, \tag{43}$$

$$\widehat{T} := \widehat{M}_3 - \frac{c_0}{c_0+2}\left[\widehat{\mathbb{E}}(w_{\ell_1}\otimes w_{\ell_2}\otimes \widehat{M}_1) + \widehat{\mathbb{E}}(w_{\ell_1}\otimes \widehat{M}_1\otimes w_{\ell_3}) + \widehat{\mathbb{E}}(\widehat{M}_1\otimes w_{\ell_2}\otimes w_{\ell_3})\right] + \frac{2c_0^2}{(c_0+1)(c_0+2)}\widehat{M}_1\otimes\widehat{M}_1\otimes\widehat{M}_1, \tag{44}$$

where, as suggested in [4], unbiased U-statistics estimates of M_1, M_2 and M_3 are:

$$\widehat{M}_1 := \widehat{\mathbb{E}}(w_\ell) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L_n}\sum_{\ell=1}^{L_n} w_{n\ell} = \frac{1}{N}\sum_{n=1}^{N}[\delta_1]_n x_n = \frac{1}{N}X\delta_1, \tag{45}$$

$$\begin{aligned}\widehat{M}_2 := \widehat{\mathbb{E}}(w_{\ell_1}w_{\ell_2}^\top) &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L_n(L_n-1)}\sum_{\ell_1=1}^{L_n}\sum_{\substack{\ell_2=1\\ \ell_2\neq\ell_1}}^{L_n} w_{n\ell_1}w_{n\ell_2}^\top \\ &= \frac{1}{N}\sum_{n=1}^{N}[\delta_2]_n\left(x_n x_n^\top - \sum_{\ell=1}^{L_n} w_{n\ell}w_{n\ell}^\top\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}[\delta_2]_n\left(x_n x_n^\top - \mathrm{diag}(x_n)\right) \\ &= \frac{1}{N}\left[X\,\mathrm{diag}(\delta_2)X^\top - \mathrm{diag}(X\delta_2)\right], \end{aligned} \tag{46}$$

$$\tag{47}$$

$$\begin{aligned}\widehat{M}_3 := \widehat{\mathbb{E}}(w_{\ell_1}\otimes w_{\ell_2}\otimes w_{\ell_3}) &= \frac{1}{N}\sum_{n=1}^{N}\delta_{3n}\sum_{\ell_1=1}^{L_n}\sum_{\substack{\ell_2=1\\ \ell_2\neq\ell_1}}^{L_n}\sum_{\substack{\ell_3=1\\ \ell_3\neq\ell_2\\ \ell_3\neq\ell_1}}^{L_n} w_{n\ell_1}\otimes w_{n\ell_2}\otimes w_{n\ell_3} \\ &= \frac{1}{N}\sum_{n=1}^{N}[\delta_3]_n\left(x_n\otimes x_n\otimes x_n - \sum_{\ell=1}^{L_n} w_{n\ell}\otimes w_{n\ell}\otimes w_{n\ell}\right. \\ &\left. - \sum_{\ell_1=1}^{L_n}\sum_{\substack{\ell_2=1\\ \ell_2\neq\ell_1}}^{L_n}\left(w_{n\ell_1}\otimes w_{n\ell_1}\otimes w_{n\ell_2} + w_{n\ell_1}\otimes w_{n\ell_2}\otimes w_{n\ell_1} + w_{n\ell_1}\otimes w_{n\ell_2}\otimes w_{n\ell_2}\right)\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}[\delta_3]_n\left(x_n\otimes x_n\otimes x_n + 2\sum_{m=1}^{M} x_{nm}(e_m\otimes e_m\otimes e_m)\right. \\ &\left. - \sum_{m_1=1}^{M}\sum_{m_2=1}^{M} x_{nm_1}x_{nm_2}(e_{m_1}\otimes e_{m_1}\otimes e_{m_2} + e_{m_1}\otimes e_{m_2}\otimes e_{m_1} + e_{m_1}\otimes e_{m_2}\otimes e_{m_2})\right). \end{aligned} \tag{48}$$

Here, the vectors δ_1, δ_2 and $\delta_3 \in \mathbb{R}^N$ are defined as $[\delta_1]_n := L_n^{-1}$; $[\delta_2]_n := (L_n(L_n-1))^{-1}$, i.e., $[\delta_2]_n = \left[\binom{L_n}{2}2!\right]^{-1}$ is the number of times to choose an ordered pair of tokens out of L_n tokens; $[\delta_3]_n := (L_n(L_n-1)(L_n-2))^{-1}$, i.e., $[\delta_3]_n = \left[\binom{L_n}{3}3!\right]^{-1}$ is the number of times to choose an ordered triple of tokens out of L_n tokens. Note that the vectors δ_1, δ_2, and δ_3 have nothing to do with the Kronecker delta δ.

For a vector $a \in \mathbb{R}^N$, we sometimes use notation $[a]_n$ to denote its n-th element. Similarly, for a matrix $A \in \mathbb{R}^{M\times N}$ we use notation $[A]_{mn}$ to denote its (m,n)-th element.

There is a slight abuse of notation in the expressions above as w_ℓ is sometimes treated as a random variable (i.e., in $\widehat{\mathbb{E}}(w_\ell)$, $\widehat{\mathbb{E}}(w_{\ell_1} w_{\ell_2}^\top)$, etc.) and sometimes as its realization. However, the difference is clear from the context.

F Appendix. Practical aspects and implementation details

F.1 Whitening of S and dimensionality reduction

The algorithms from Section 4 require the computation of a whitening matrix W of S. Due to the similar diagonal structure ((41) and (11)) of the matrix S for both the LDA moments (39) and the GP/DICA cumulants (10), the computation of a whitening matrix is exactly the same in both cases.

By a whitening matrix, we mean a matrix $W \in \mathbb{R}^{K \times M}$ (in practice, $M \gg K$) that does not only whiten $S \in \mathbb{R}^{M \times M}$, but also reduces its dimensionality such that[14] $WSW^\top = I_K$.

Let $S = U \Sigma U^\top$ be an orthogonal eigendecomposition of the symmetric matrix S. Let $\Sigma_{1:K}$ denotes the diagonal matrix that contains the largest K eigenvalues[15] of S on its diagonal and let $U_{1:K}$ be a matrix with the respective eigenvalues in its columns. Then, a whitening matrix is

$$W = \Sigma_{1:K}^{\dagger 1/2} U_{1:K}^\top, \tag{49}$$

where $\Sigma_{1:K}^{\dagger 1/2}$ is a diagonal matrix constructed from $\Sigma_{1:K}$ by taking the inverse and the square root of its non-zero diagonal values († stands for the pseudo-inverse).

In practice, when only a finite sample estimator $\widehat{S}$ of S is available, the following finite sample estimator $\widehat{W}$ of W can be introduced

$$\widehat{W} := \widehat{\Sigma}_{1:K}^{\dagger 1/2} \widehat{U}_{1:K}^\top, \tag{50}$$

where $\widehat{S} = \widehat{U} \widehat{\Sigma} \widehat{U}^\top$.

F.2 Computation of the finite sample estimators of the GP/DICA cumulants

In this section, we present efficient formulas for computation of the finite sample estimate (see Appendix C.4 for the definition of $\widehat{T}$) of $\widehat{W}\widehat{T}(v)\widehat{W}^\top$ for the GP/DICA models. The construction of the finite sample estimator $\widehat{W}$ is discussed in Appendix F.1, while the computation of $\widehat{S}$ (29) is straightforward.

[14] Note that such a whitening matrix $W \in \mathbb{R}^{K \times M}$ is not uniquely defined as left multiplication by any orthogonal matrix $V \in \mathbb{R}^{K \times K}$ does not change anything. Indeed, let $\widetilde{W} = VW$, then $\widetilde{W} S \widetilde{W}^\top = VWSW^\top V^\top = I_K$.

[15] We mean the largest non-negative eigenvalues. In theory, S have to be PSD. In practice, when we deal with finite number of samples, respective estimate of S can have negative eigenvalues. However, for K sufficiently small, S should have enough positive eigenvalues. Moreover, it is standard practice to use eigenvalues of S for estimation of a good value of K, e.g., by thresholding all negative and close to zero eigenvalues.

By plugging the definition of the tensor $\widehat{T}$ (30) in the formula (16) for the projection of a tensor onto a vector, we obtain for a given $v \in \mathbb{R}^M$:

$$\begin{aligned}
\left[\widehat{T}(v)\right]_{m_1 m_2} &= \sum_{m_3} \widehat{\mathrm{cum}}(x_{m_1}, x_{m_2}, x_{m_3}) v_{m_3} + 2 \sum_{m_3} \delta(m_1, m_2, m_3) \widehat{\mathbb{E}}(x_{m_3}) v_{m_3} \\
&- \sum_{m_3} \delta(m_2, m_3) \widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) v_{m_3} \\
&- \sum_{m_3} \delta(m_1, m_3) \widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) v_{m_3} \\
&- \sum_{m_3} \delta(m_1, m_2) \widehat{\mathrm{cov}}(x_{m_1}, x_{m_3}) v_{m_3} \\
&= \sum_{m_3} \widehat{\mathrm{cum}}(x_{m_1}, x_{m_2}, x_{m_3}) v_{m_3} + 2\delta(m_1, m_2) \widehat{\mathbb{E}}(x_{m_1}) v_{m_1} \\
&- \widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) v_{m_2} - \widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) v_{m_1} - \delta(m_1, m_2) \sum_{m_3} \widehat{\mathrm{cov}}(x_{m_1}, x_{m_3}) v_{m_3}.
\end{aligned}$$

This gives the following for the expression $\widehat{W}\widehat{T}(v)\widehat{W}^\top$:

$$\begin{aligned}
\left[\widehat{W}\widehat{T}(v)\widehat{W}^\top\right]_{k_1 k_2} &= \widehat{W} {k_1}^\top \widehat{T}(v) \widehat{W}_{k_2} \\
&= \sum_{m_1, m_2, m_3} \widehat{\mathrm{cum}}(x_{m_1}, x_{m_2}, x_{m_3}) v_{m_3} \widehat{W}_{k_1 m_1} \widehat{W}_{k_2 m_2} \\
&+ 2 \sum_{m_1, m_2} \delta(m_1, m_2) \widehat{\mathbb{E}}(x_{m_1}) v_{m_1} \widehat{W}_{k_1 m_1} \widehat{W}_{k_2 m_2} \\
&- \sum_{m_1, m_2} \widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) v_{m_2} \widehat{W}_{k_1 m_1} \widehat{W}_{k_2 m_2} \\
&- \sum_{m_1, m_2} \widehat{\mathrm{cov}}(x_{m_1}, x_{m_2}) v_{m_1} \widehat{W}_{k_1 m_1} \widehat{W}_{k_2 m_2} \\
&- \sum_{m_1, m_3} \widehat{\mathrm{cov}}(x_{m_1}, x_{m_3}) v_{m_3} \widehat{W}_{k_1 m_1} \widehat{W}_{k_2 m_1},
\end{aligned}$$

where $\widehat{W}_k$ denotes the k-th row of $\widehat{W}$ as a column vector. By further plugging in the expressions (31) for the unbiased finite sample estimates of $\widehat{\mathrm{cov}}$ and $\widehat{\mathrm{cum}}$, we further get

$$\begin{aligned}
\left[\widehat{W}\widehat{T}(v)\widehat{W}^\top\right]_{k_1 k_2} &= \frac{N}{(N-1)(N-2)} \sum_n \left\langle \widehat{W}_{k_1}, x_n - \widehat{\mathbb{E}}(x) \right\rangle \left\langle \widehat{W}_{k_2}, x_n - \widehat{\mathbb{E}}(x) \right\rangle \left\langle v, x_n - \widehat{\mathbb{E}}(x) \right\rangle \\
&+ 2 \sum_m \widehat{\mathbb{E}}(x_m) v_m \widehat{W}_{k_1 m} \widehat{W}_{k_2 m} \\
&- \frac{1}{N-1} \sum_n \left\langle \widehat{W}_{k_1}, x_n - \widehat{\mathbb{E}}(x) \right\rangle \left\langle v \circ \widehat{W}_{k_2}, x_n - \widehat{\mathbb{E}}(x) \right\rangle \\
&- \frac{1}{N-1} \sum_n \left\langle v \circ \widehat{W}_{k_1}, x_n - \widehat{\mathbb{E}}(x) \right\rangle \left\langle \widehat{W}_{k_2}, x_n - \widehat{\mathbb{E}}(x) \right\rangle \\
&- \frac{1}{N-1} \sum_n \left\langle \widehat{W}_{k_1} \circ \widehat{W}_{k_2}, x_n - \widehat{\mathbb{E}}(x) \right\rangle \left\langle v, x_n - \widehat{\mathbb{E}}(x) \right\rangle,
\end{aligned}$$

where $\circ$ denotes the elementwise Hadamard product. Introducing the counts matrix $X \in \mathbb{R}^{M \times N}$ where each element X_{mn} is the count of the m-th word in the n-th document (note, the matrix X

contain the vector x_n in the n-th column), we further simplify the above expression

$$\begin{aligned}
\widehat{W}\widehat{T}(v)\widehat{W}^\top &= \frac{N}{(N-1)(N-2)}(\widehat{W}X)\mathrm{diag}[X^\top v](\widehat{W}X)^\top \\
&+ \frac{N}{(N-1)(N-2)}\left\langle v, \widehat{\mathbb{E}}(x)\right\rangle \left[2N(\widehat{W}\widehat{\mathbb{E}}(x))(\widehat{W}\widehat{\mathbb{E}}(x))^\top - (\widehat{W}X)(\widehat{W}X)^\top\right] \\
&- \frac{N}{(N-1)(N-2)}\left[\widehat{W}X(X^\top v)(\widehat{W}\widehat{\mathbb{E}}(x))^\top + \widehat{W}\widehat{\mathbb{E}}(x)(\widehat{W}X(X^\top v))^\top\right] \\
&+ 2\widehat{W}\mathrm{diag}[v\circ\widehat{\mathbb{E}}(x)]\widehat{W}^\top \\
&- \frac{1}{N-1}\left[(\widehat{W}X)(\widehat{W}\mathrm{diag}(v)X)^\top + (\widehat{W}\mathrm{diag}(v)X)(\widehat{W}X)^\top + \widehat{W}\mathrm{diag}[X(X^\top v)]\widehat{W}^\top\right] \\
&+ \frac{N}{N-1}\left[(\widehat{W}\widehat{\mathbb{E}}(x))(\widehat{W}\mathrm{diag}[v]\widehat{\mathbb{E}}(x))^\top + (\widehat{W}\mathrm{diag}[v]\widehat{\mathbb{E}}(x))(\widehat{W}\widehat{\mathbb{E}}(x))^\top\right] \\
&+ \frac{N}{N-1}\left\langle v, \widehat{\mathbb{E}}(x)\right\rangle \widehat{W}\mathrm{diag}[\widehat{\mathbb{E}}(x)]\widehat{W}^\top.
\end{aligned} \tag{51}$$

A more compact way to write down expression (51) is as follows

$$\begin{aligned}
\widehat{W}\widehat{T}(v)\widehat{W}^\top &= \frac{N}{(N-1)(N-2)}\left[T_1 + \langle v, \widehat{\mathbb{E}}(x)\rangle (T_2 - T_3) - (T_4 + T_4^\top)\right] \\
&+ \frac{1}{N-1}\left[T_5 + T_5^\top - T_6 - T_6^\top + \widehat{W}\mathrm{diag}(a)\widehat{W}^\top\right],
\end{aligned} \tag{52}$$

where

$$\begin{aligned}
T_1 &= (\widehat{W}X)\mathrm{diag}[X^\top v](\widehat{W}X)^\top, \\
T_2 &= 2N(\widehat{W}\widehat{\mathbb{E}}(x))(\widehat{W}\widehat{\mathbb{E}}(x))^\top, \\
T_3 &= (\widehat{W}X)(\widehat{W}X)^\top, \\
T_4 &= \widehat{W}X(X^\top v)(\widehat{W}\widehat{\mathbb{E}}(x))^\top, \\
T_5 &= (\widehat{W}X)(\widehat{W}\mathrm{diag}(v)X)^\top, \\
T_6 &= (\widehat{W}\mathrm{diag}(v)\widehat{\mathbb{E}}(x))(\widehat{W}\widehat{\mathbb{E}}(x))^\top, \\
a &= 2(N-1)[v\circ\widehat{\mathbb{E}}(x)] + \langle v, \widehat{\mathbb{E}}(x)\rangle\widehat{\mathbb{E}}(x) - X(X^\top v).
\end{aligned}$$

F.3 Computational complexity of the GP/DICA T-cumulant estimator (52)

When computing the T-cumulant P times with the formula above, the following terms are dominant: $O(RNK)+O(NK^2)+O(MK)$, where R is the largest number of unique words (non-zero counts) in a document over the corpus. In practice, almost always $K < M < N$, which gives the overall complexity of P computations of the estimator (52) to be equal to $O(PRNK)+O(PNK^2)$.

F.4 Computation of the finite sample estimators of the LDA moments

In this section, we present efficient formulas for computation of the finite sample estimate (see Appendix E.2 for the definition of $\widehat{T}$) of $\widehat{W}\widehat{T}(v)\widehat{W}^\top$ for the LDA model. Note that the construction of the sample estimator $\widehat{W}$ of a whitening matrix W is discussed in Appendix F.1). The computation of $\widehat{S}$ (43) is straightforward. This approach to efficient implementation was discussed in [4], however, to the best of our knowledge, the final expressions were not explicitly stated before. All derivations are straightforward, but quite tedious.

By analogy with the GP/DICA case, a projection (16) of the tensor $\widehat{T} \in \mathbb{R}^{M\times M\times M}$ (44) onto some vector $v \in \mathbb{R}^M$ in the LDA is

$$\left[\widehat{T}(v)\right]_{m_1m_2} = \sum_{m_3=1}^{M} \left[\widehat{M}_3\right]_{m_1m_2m_3} v_{m_3} + \frac{2c_0^2}{(c_0+1)(c_0+2)} \sum_{m_3} [\widehat{M}_1]_{m_1}[\widehat{M}_1]_{m_2}[\widehat{M}_1]_{m_3} v_{m_3}$$

$$- \frac{c_0}{c_0+2} \sum_{m_3=1}^{M} \left[\widehat{\mathbb{E}}(w_{\ell_1}\otimes w_{\ell_2}\otimes \widehat{M}_1) + \widehat{\mathbb{E}}(w_{\ell_1}\otimes \widehat{M}_1\otimes w_{\ell_3}) + \widehat{\mathbb{E}}(\widehat{M}_1\otimes w_{\ell_2}\otimes w_{\ell_3})\right]_{m_1m_2m_3} v_{m_3}.$$

Plugging in the expression (48) for an unbiased sample estimate $\widehat{M}_3$ of M_3, we get

$$\begin{aligned}
\left[\widehat{T}(v)\right]_{m_1m_2} &= \frac{1}{N}\sum_{n=1}^{N}[\delta_3]_n \left(x_{nm_1}x_{nm_2}\langle x_n, v\rangle + 2\sum_{m_3}\delta(m_1,m_2,m_3)x_{nm_3}v_{m_3}\right) \\
&\quad - \frac{1}{N}\sum_{n=1}^{N}[\delta_3]_n \sum_{m_3=1}^{M}\left[\sum_{i,j=1}^{M} x_{ni}x_{nj}\left(e_i\otimes e_i\otimes e_j + e_i\otimes e_j\otimes e_i + e_i\otimes e_j\otimes e_j\right)\right]_{m_1m_2m_3} v_{m_3} \\
&\quad + \frac{2c_0^2}{(c_0+1)(c_0+2)}[\widehat{M}_1]_{m_1}[\widehat{M}_1]_{m_2}\left\langle \widehat{M}_1, v\right\rangle \\
&\quad - \frac{c_0}{c_0+2}\left([\widehat{M}_2]_{m_1m_2}\left\langle \widehat{M}_1, v\right\rangle + \sum_{m_3=1}^{M}\left([\widehat{M}_2]_{m_1m_3}[\widehat{M}_1]_{m_2}v_{m_3} + [\widehat{M}_2]_{m_2m_3}[\widehat{M}_1]_{m_1}v_{m_3}\right)\right),
\end{aligned}$$

where e_1, e_2, …, e_M denote the canonical vectors of $\mathbb{R}^M$ (i.e., the columns of the identity matrix I_M). Further, this gives the following for the expression $\widehat{W}\widehat{T}(v)\widehat{W}^\top$:

$$\begin{aligned}
\left[\widehat{W}\widehat{T}(v)\widehat{W}^\top\right]_{k_1k_2} &= \frac{1}{N}\sum_{n=1}^{N}[\delta_3]_n\left(\langle x_n, v\rangle\left\langle x_n, \widehat{W}_{k_1}\right\rangle\left\langle x_n, \widehat{W}_{k_2}\right\rangle + 2\sum_{m=1}^{M} x_{nm}v_m\widehat{W}_{k_1m}\widehat{W}_{k_2m}\right) \\
&\quad - \frac{1}{N}\sum_{n=1}^{N}\delta_{3n}\sum_{i,j=1}^{M} x_{ni}x_{nj}\left(\widehat{W}_{k_1i}\widehat{W}_{k_2i}v_j + \widehat{W}_{k_1i}\widehat{W}_{k_2j}v_i + \widehat{W}_{k_1i}\widehat{W}_{k_2j}v_j\right) \\
&\quad - \frac{c_0}{c_0+2}\left(\left\langle \widehat{W}_{k_1}, \left[\widehat{M}_2\right]\widehat{W}_{k_2}\right\rangle + \left\langle \widehat{W}_{k_1}, \widehat{M}_2v\right\rangle\left\langle \widehat{M}_1\widehat{W}_{k_2}\right\rangle + \left\langle \widehat{W}_{k_2}, \widehat{M}_2v\right\rangle\left\langle \widehat{M}_1, \widehat{W}_{k_1}\right\rangle\right) \\
&\quad + \frac{2c_0^2}{(c_0+1)(c_0+2)}\left\langle \widehat{M}_1, \widehat{W}_{k_1}\right\rangle\left\langle \widehat{M}_1, \widehat{W}_{k_2}\right\rangle\left\langle \widehat{M}_1, v\right\rangle,
\end{aligned}$$

where $\widehat{W}_k$ denotes the k-th row of $\widehat{W}$ as a column-vector. This further simplifies to

$$\begin{aligned}
\widehat{W}\widehat{T}(v)\widehat{W}^\top &= \frac{1}{N}(\widehat{W}X)\mathrm{diag}\left[(X^\top v)\circ\delta_3\right](\widehat{W}X)^\top \\
&\quad + \frac{1}{N}\widehat{W}\mathrm{diag}\left[2[(X\delta_3)\circ v] - X[(X^\top v)\circ\delta_3]\right]\widehat{W}^\top \\
&\quad - \frac{1}{N}(\widehat{W}\mathrm{diag}[v]X)\mathrm{diag}[\delta_3](\widehat{W}X)^\top \\
&\quad - \frac{1}{N}(\widehat{W}X)\mathrm{diag}[\delta_3](\widehat{W}\mathrm{diag}[v]X)^\top \\
&\quad - \frac{c_0}{c_0+2}\left[\left\langle \widehat{M}_1, v\right\rangle(\widehat{W}\widehat{M}_2\widehat{W}^\top) + (\widehat{W}(\widehat{M}_2v))(\widehat{W}\widehat{M}_1)^\top + (\widehat{W}\widehat{M}_1)(\widehat{W}(\widehat{M}_2v))^\top\right] \\
&\quad + \frac{2c_0^2}{(c_0+1)(c_0+2)}\left\langle \widehat{M}_1, v\right\rangle(\widehat{W}\widehat{M}_1)(\widehat{W}\widehat{M}_1)^\top.
\end{aligned} \tag{53}$$

A more compact representation gives:

$$\begin{aligned}
\widehat{W}\widehat{T}(v)\widehat{W}^\top &= \frac{1}{N}\left[T_1 + T_2 - T_3 - T_3^\top\right] - \frac{c_0}{c_0+2}\left[\langle\widehat{M}_1, v\rangle(\widehat{W}\widehat{M}_2\widehat{W}^\top) + T_4 + T_4^\top\right] \\
&\quad + \frac{2c_0^2}{(c_0+1)(c_0+2)}\langle\widehat{M}_1, v\rangle(\widehat{W}\widehat{M}_1)(\widehat{W}\widehat{M}_1)^\top,
\end{aligned} \tag{54}$$

where

$$T_1 = (\widehat{W}X)\text{diag}\left[(X^\top v)\circ\delta_3\right](\widehat{W}X)^\top,$$
$$T_2 = \widehat{W}\text{diag}\left[2[(X\delta_3)\circ v] - X[(X^\top v)\circ\delta_3]\right]\widehat{W}^\top,$$
$$T_3 = [\widehat{W}\text{diag}(v)X]\text{diag}(\delta_3)(\widehat{W}X)^\top,$$
$$T_4 = [\widehat{W}(\widehat{M}_2 v)](\widehat{W}\widehat{M}_1)^\top.$$

F.5 Computational complexity of the LDA T-moment estimator (54)

By analogy with Appendix F.3, the computational complexity of the T-moment is $O(RNK) + O(NK^2)$. However, in practice we noticed that the computation of (52) is slightly faster for larger datasets than the computation of (54) (although the code for both was equally well optimized). This means that the constants in $O(RNK) + O(NK^2)$ for the LDA T-moment are, probably, slightly larger than for the GP/DICA T-cumulant.

F.6 Estimation of the model parameters for GP/DICA model

Below we briefly discuss the recovery of the model parameters for the GP/DICA and LDA models from a joint diagonalization matrix $A \in \mathbb{R}^{K\times M}$ estimated in Algorithm 1. This matrix has the property that AD should be approximately diagonal up to a permutation of the columns of D. The standard approach [3] of taking the pseudo-inverse of A to get an estimate of the topic matrix D has a problem that it does not preserve the simplex constraint of the topics (in particular, the non-negativity of $\widetilde{D}$). Due to the space constraints, we do not discuss this issue here, but we observed experimentally that this can potentially significantly deteriorate performance of all moment matching algorithms for topic models considered in this paper. We made an attempt to solve this problem by integrating the non-negativity constraint into the Jacobi-updates procedure of the orthogonal joint diagonalization algorithm, but the obtained results did not lead to any significant improvement. Therefore, in our experiments for both GP/DICA cumulants and LDA moments, we estimate the topic matrix by thresholding the negative values of the pseudo-inverse of A:

$$\widehat{d}_k := \tau_k \max(0, [A^\dagger]_{:k})/\|\max(0, [A^\dagger]_{:k})\|_1,$$

where $[A^\dagger]_{:k}$ is the k-th column of the pseudo-inverse $A^\dagger$ of A, and $\tau_k = \pm 1$ set to -1 if $[A^\dagger]_{:k}$ has more negative than positive values. This might not be the best option, and we leave this issue for the future research.

To estimate the parameters for the prior distribution over the topic intensities α_k for the DICA model (4), we use the diagonalized form of the projected tensor from (17) and relate it to the output diagonal elements a_p for the p-th projection:

$$[a_p]_k = \widetilde{t}_k\langle z_k, u_p\rangle = \frac{t_k}{s_k^{3/2}}\langle z_k, u_p\rangle = \frac{\text{cum}(\alpha_k,\alpha_k,\alpha_k)}{[\text{var}(\alpha_k)]^{3/2}}\left\langle \tau_k\widetilde{d}_k, W^\top u_p\right\rangle, \tag{55}$$

where $\widetilde{d}_k = \tau_k \max(0, [A^\dagger]_{:k})$. This formula is valid for any prior on α_k in the DICA model. For the GP model (3) where $\alpha_k \sim \text{Gamma}(c_k, b)$, we have that $\text{var}(\alpha_k) = \frac{c_k}{b^2}$ and $\text{cum}(\alpha_k,\alpha_k,\alpha_k) = \frac{2c_k}{b^3}$, and thus $\widetilde{t}_k = \frac{2}{\sqrt{c_k}}$, which enables us to estimate c_k. Plugging this value of $\widetilde{t}_k$ in (55), and solving for c_k gives the following expression:

$$c_k = \frac{4\left\langle \widetilde{d}_k, W^\top u_p\right\rangle^2}{[a_p]_k^2}.$$

By replacing the quantities on the RHS with their estimated ones, we get one estimate for c_k per projection. We use as our final estimate the average estimate over the projections:

$$\widehat{c}_k := \frac{1}{P}\sum_{p=1}^{P}\frac{4\left\langle \widetilde{d}_k, \widehat{W}^\top u_p\right\rangle^2}{[a_p]_k^2}. \tag{56}$$

Reusing the properties of the length of documents for the GP model as described in Appendix B.2, we finally use the following estimates for rate parameter b of the gamma distribution:

$$\widehat{b} := \frac{\widehat{c}_0}{\widehat{L}}, \tag{57}$$

where $\widehat{c}_0 := \sum_k \widehat{c}_k$ and $\widehat{L}$ is the average document length in the corpus.

By analogy, similar formulas for the estimation of the Dirichlet parameter c of the LDA model can be derived and are a straightforward extension of the expression in [3].

G Appendix. Complexity of algorithms and details on the experiments

G.1 Code and complexity

Our (mostly Matlab) implementations of the diagonalization algorithms (JD, Spec, and TPM) for both the GP/DICA cumulants and LDA moments are available online.[16] Moreover, all datasets and the code for reproducing our experiments are available.[17] To our knowledge, no efficient implementation of these algorithms was available for LDA. Each experiment was run in a single thread.

The bottleneck for the spectral, JD, and TPM algorithms is the computation of the cumulants/moments. However, the expressions (52) and (54) provide efficient formulas for fast computation of the GP/DICA cumulants and LDA moments ($O(RNK + NK^2)$, where R is the largest number of non-zeros in the count vector x over all documents, see Appendix F.3 and F.5), which makes even the Matlab implementation fast for large datasets. Since all diagonalization algorithms (spectral, JD, TPM) perform the whitening step once, it is sufficient to compare their complexities by the number of times the cumulants/moments are computed.

Spectral. The spectral algorithm estimates the cumulants/moments only once leading to $O(NK(R + K))$ complexity and, therefore, is the fastest.

JD. For JD, rather than estimating P cumulants/moments separately, one can jointly estimate these values by precomputing and reusing some terms (e.g., WX). However, the complexity is still $O(PNK(R + K))$, although in practice it is sufficient to have $P = K$ or even smaller.

TPM. For TPM some parts of the cumulants/moments can also be precomputed, but as TPM normally does many more iterations than P, it can be significantly slower. In general, the complexity of TPM can be significantly influenced by the initialization of the parameters of the algorithm. There are two main parameters: L_{tpm} is the number of random restarts within one deflation step and N_{tpm} is the maximum number of iterations for each of L_{tpm} random restarts (different from N and L). Some restarts converge very fast (in much less than N_{tpm} iterations), while others are slow. Moreover, as follows from theoretical results [4] and, as we observed in practice, the restarts which converge to a good solution converge fast, while slow restarts, normally, converge to a worse solution. Nevertheless, in the worst case, the complexity is $O(N_{tpm}L_{tpm}NK(R + K))$.

Note that for the experiment in Figure 1, $L_{tpm} = 10$ and $N_{tpm} = 100$ and the run with the best objective is chosen. We believe that these values are reasonable in a sense that they provide a good accuracy solution ($\varepsilon = 10^{-5}$ for the norm of the difference of the vectors from the previous and the current iteration) in a little number of iterations, however, they may not be the best ones.

JD implementation. For the orthogonal joint diagonalization algorithm, we implemented a faster C++ version of the previous Matlab implementation[18] by J.-F. Cardoso. Moreover, the orthogonal joint diagonalization routine can be initialized in different ways: (a) with the $K \times K$ identity matrix or (b) with a random orthogonal $K \times K$ matrix. We tried different options and in nearly all cases the algorithm converged to the same solution, implying that initialization with the identity matrix is sufficient.

Whitening matrix. For the large vocabulary size M, computation of a whitening matrix can be expensive (in terms of both memory and time). One possible solution would be to reduce the vocabulary size with, e.g., TF-IDF score, which is a standard practice in the topic modeling context.

[16] https://github.com/anastasia-podosinnikova/dica-light

[17] https://github.com/anastasia-podosinnikova/dica

[18] http://perso.telecom-paristech.fr/~cardoso/Algo/Joint_Diag/joint_diag_r.m

	min	mean	max
JD-GP	148	192	247
JD-LDA	252	284	366
JD(k)-GP	157	190	247
JD(k)-LDA	264	290	318
JD(f)-GP	1628	1846	2058
JD(f)-LDA	2545	2649	2806
Spec-GP	101	107	111
Spec-LDA	107	140	193
TPM-GP	1734	2393	2726
TPM-LDA	12723	16460	19356

Table 1: The running times in seconds of the algorithms from Figure 1, corresponds to the case when $N = 50,000$. Each algorithm was run 5 times, so the times in the table display the minimum (min), mean, and maximum (max) time.

Another option is using a stochastic eigendecomposition (see, e.g., [33]) to approximate the whitening matrix.

Variational inference. For variational inference, we used the code of D. Blei and modified it for the estimation of a non-symmetric Dirichlet prior c, which is known to be important [35]. The default values of the tolerance/maximum number of iterations parameters are used for variational inference. The computational complexity of one iteration for one document of the variational inference algorithm is $O(RK)$, where R is the number of non-zeros in the count vector for this document, which is then performed a significant number of times for each document.

G.2 Runtimes of the algorithms

In Table 1, we present the running times of the algorithms from Section 5.1. JD and JD(k) are significantly faster than JD(f) as expected, although the performance in terms of the ℓ_1-error is nearly the same for all of them. This indicates that preference should be given to the JD or JD(k) algorithms.

The running time of all LDA-algorithms is higher than the one of the GP/DICA-algorithms. This indicates that the computational complexity of the LDA-moments is slightly higher than the one of the GP/DICA-cumulants (compare, e.g., the times for the spectral algorithm which almost completely consist of the computation of the moments/cumulants). Moreover, the runtime of TPM-LDA is significantly higher (half an hour vs. several hours) than the one of TPM-GP/DICA. This can be explained by the fact that the LDA-moments have more noise than the GP/DICA-cumulants and, hence, the algorithm is slower. Interestingly, all versions of JD algorithm are not that sensitive to noise.

Computation of a whitening matrix is roughly 30 sec (this time is the same for all algorithms and is included in the numbers above).

G.3 Initialization of the parameter c_0 for the LDA moments

The construction of the LDA moments requires the parameter c_0, which is not trivial to set in the unsupervised setting of topic modeling, especially taking into account the complexity of the evaluation for topic models [16]. For the semi-synthetic experiments, the true value of c_0 is provided to the algorithms. It means that the LDA moments, in this case, have access to some oracle information, which in practice is never available. For real data experiments, c_0 is set to the value obtained with variational inference. The experiments in Appendix G.4 show that this choice was somewhat important. However, this requires more thorough investigation.

G.4 The LDA moments vs parameter c_0

In this section, we experimentally investigate dependence of the LDA moments on the parameter c_0. In Figure 5, the joint diagonalization algorithm with the LDA moment is compared for different values of c_0 provided to the algorithm. The data is generated similarly to Figure 2. The experiment indicates that the LDA moments are somewhat sensitive to the choice of c_0. For example, the

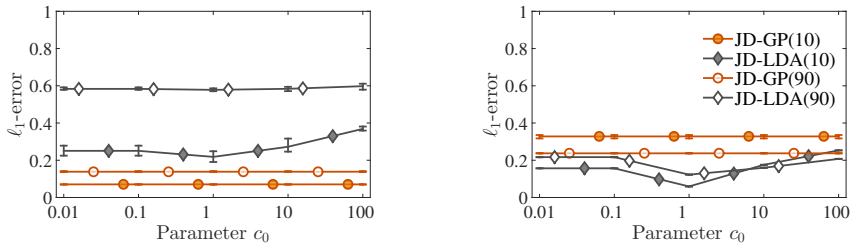


Figure 5: Performance of the LDA moments depending on the parameter c_0. D and c are learned from the AP dataset for $K = 10$ and $K = 50$ and true $c_0 = 1$. JD-GP(10) for $K = 10$ and JD-GP(50) for $K = 50$. Number of sampled documents $N = 20,000$. For the error bars, each dataset is resampled 5 times. Data **(left)**: *GP* sampling; **(right)**: *LDAfix(*200*)* sampling. *Note*: a smaller value of the ℓ_1-error is better.

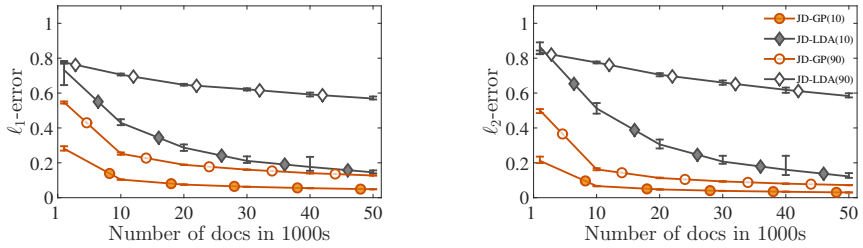


Figure 6: Comparison of the ℓ_1- and ℓ_2- errors on the NIPS semi-synthetic dataset as in Figure 2 (top, left). The ℓ_2 norms of the topics were normalized to [0,1] for the computation of the ℓ_2 error.

recovery ℓ_1-error doubles when moving from the correct choice $c_0 = 1$ to the plausible alternative $c_0 = 0.1$ for $K = 10$ on the *LDAfix(*200*)* dataset (JD-LDA(10) line on the right of Figure 5).

G.5 Comparison of the ℓ_1- and ℓ_2-errors

The sample complexity results [3] for the spectral algorithm for the LDA moments allow straightforward extension to the GP/DICA cumulants, if the results from Proposition 3.1 are taken into account. The analysis is, however, in terms of the ℓ_2-norm. Therefore, in Figure 6, we provide experimental comparison of the ℓ_1- and ℓ_2-errors to verify that they are indeed behaving similarly.

G.6 Evaluation of the real data experiments

For the evaluation of topic recovery in the real data case, we use an approximation of the log-likelihood for held out documents as the metric. The approximation is computed using a Chib-style method as described by [16] using the implementation by the authors.[19] Importantly, this evaluation methods is applicable for both the LDA model as well as the GP model. Indeed, as it follows from Section 2 and Appendix B.1, the GP model is equivalent to the LDA model when conditioning on the length of a document L (with the same c_k hyper parameters), while the LDA model does not make any assumption on the document length. For the test log-likelihood comparison, we thus treat the GP model as a LDA model (we do not include the likelihood of the document length).

G.7 More on the real data experiments

The detailed experimental setup is as follows. Each dataset is separated into 5 training/evaluation pairs, where the documents for evaluation are chosen randomly and non-repetitively among the folds (600 documents are held out for KOS; 400 documents are held out for AP; 450 documents are held out for NIPS). Then, the model parameters are learned for a different number of topics. The evaluation of the held-out documents is performed with averaging over 5 folds. In Figure 3 and Figure 7, on the y-axis, the predictive log-likelihood in bits averaged per token is presented.

[19] http://homepages.inf.ed.ac.uk/imurray2/pub/09etm

In addition to the experiments with AP and KOS in Figure 3, we demonstrate one more experiment with the NIPS dataset in Figure 7 (right).

Note that, as the LDA moments require at least 3 tokens in each document, 1 document from the NIPS dataset and 3 documents from the AP dataset, which did not fulfill this requirement, were removed.

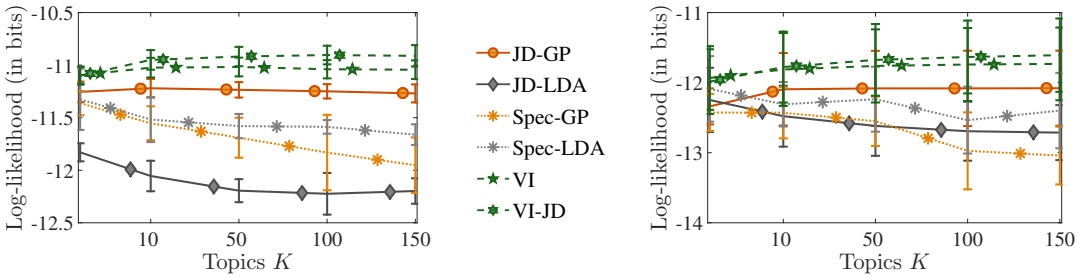


Figure 7: Experiments with real data. **Left:** the KOS dataset. **Right:** the NIPS dataset. *Note*: a higher value of the log-likelihood is better.

Importantly, we observed that VI when initialized with the output of the JD-GP is consistently better in terms of the predictive log-likelihood. Therefore, the new algorithm can be used for more clever initialization of other LDA/GP inference methods.

We also observe that the joint diagonalization algorithm for the LDA moments is worse than the spectral algorithm. This indicates that the diagonal structure (41) and (42) might not be present in the sample estimates (43) and (44) due to either model misspecification or to finite sample complexity issues.

Supplementary References

[31] A. Anandkumar, D.P. Foster, D. Hsu, S.M. Kakade, and Y.-K. Liu. A spectral algorithm for latent Dirichlet allocation. *Algorithmica*, 72(1):193–214, 2015.

[32] B.A. Frigyik, A. Kapila, and M.R. Gupta. Introduction to the Dirichlet distribution and related processes. Technical report, University of Washington, 2010.

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

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

[35] H.M. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: why priors matter. In *NIPS*, 2009.