A Nested Matrix-Tensor Model for Noisy Multi-view Clustering

Mohamed El Amine Seddik **Mastane Achab** **Henrique Goulard** **Merouane Debbah**

Abstract

In this paper, we propose a nested matrix-tensor model which extends the spiked rank-one tensor model of order three. This model is particularly motivated by a multi-view clustering problem in which multiple noisy observations of each data point are acquired, with potentially non-uniform variances along the views. In this case, data can be naturally represented by an order-three tensor where the views are stacked. Given such a tensor, we consider the estimation of the hidden clusters via performing a best rank-one tensor approximation. In order to study the theoretical performance of this approach, we characterize the behavior of this best rank-one approximation in terms of the alignments of the obtained component vectors with the hidden model parameter vectors, in the large-dimensional regime. In particular, we show that our theoretical results allow us to anticipate the exact accuracy of the proposed clustering approach. Furthermore, numerical experiments indicate that leveraging our tensor-based approach yields better accuracy compared to a naive unfolding-based algorithm which ignores the underlying low-rank tensor structure. Our analysis unveils unexpected and non-trivial phase transition phenomena depending on the model parameters, "interpolating" between the typical behavior observed for the spiked matrix and tensor models.

1 Introduction

Tensor methods have received growing attention in recent years, especially from a statistical perspective, following the introduction of a statistical model for tensor PCA by Richard & Montanari (2014). In machine learning, these methods are particularly attractive for addressing several unsupervised learning tasks which can be formulated as the extraction of some *low-rank structure* from a (potentially high-dimensional) tensor containing observations or functions thereof (such as high-order moments). Among the many existing examples, we can mention learning latent variable models such as Dirichlet allocation, topic models, multi-view models and Gaussian mixtures (Anandkumar et al., 2014, 2015; Ge et al., 2015; Hsu et al., 2012; Hsu & Kakade, 2013; Janzamin et al., 2019; Khouja et al., 2022; Bakshi et al., 2022; Rahmani et al., 2020); learning probability densities and non-Gaussian mixtures (Kargas & Sidiropoulos, 2019; Singhal et al., 2023; Oseledets & Kharyuk, 2021); detecting communities from interaction data of (possibly multi-view or time-evolving) networks (Anandkumar et al., 2013; Huang et al., 2015; Gujral et al., 2020; Fernandes et al., 2021); and high-order co-clustering (Papalexakis et al., 2012).

Despite its simplicity, the statistical model of Richard & Montanari (2014), sometimes called a *rank-one spiked tensor model*, has raised many theoretical challenges. A significant amount of work has been done to understand the fundamental questions related to this model (Perry et al., 2020; Jagannath et al., 2020; Goulart et al., 2022; Auddy & Yuan, 2022; Ben Arous et al., 2021; Seddik et al., 2021), in particular involving statistical thresholds and the asymptotic performance of estimators in the large-dimensional limit. However, the findings of these works have a somewhat limited practical impact due to the rank-one nature of that model, motivating the development and study of more sophisticated statistical models for the analysis of tensor methods. In particular, phase transitions

Preprint. Under review.

associated with multi-spiked tensor models of rank $r > 1$ have been considered by Chen et al. (2021); Lesieur et al. (2017).

In this work, we take another path towards bridging the gap between theory and practical applications, by proposing a statistical *nested matrix-tensor model* that generalizes the (third-order) rank-one spiked tensor model and is motivated by a problem that we call *noisy multi-view clustering*, which can be formulated as follows. Let $\mathbf{M} = \boldsymbol{\mu}\boldsymbol{y}^\top + \mathbf{Z} \in \mathbb{R}^{p\times n}$ be a data matrix containing n observations of p-dimensional vectors centered around $\pm\boldsymbol{\mu}$ (i.e., data are made of two classes), with $\boldsymbol{y} \in \{-1, 1\}^n$ holding their corresponding labels and $\mathbf{Z}$ a Gaussian matrix modeling data dispersion. Now, suppose that we are given m different noisy observations of $\mathbf{M}$ with potentially different signal-to-noise ratios (SNR), denoted by:

$$\tilde{\mathbf{X}}_k = \boldsymbol{\mu}\boldsymbol{y}^\top + \mathbf{Z} + \tilde{\mathbf{W}}_k, \quad k = 1, \ldots, m,$$

where $\tilde{\mathbf{W}}_k$ is a $p \times n$ matrix comprising independent Gaussian entries drawn from $\mathcal{N}(0, \sigma_k^2)$. Assuming that the variances σ_k^2 are known (or can be accurately estimated), one can build a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times n\times m}$ containing normalized slices $\mathbf{X}_k = h_k\tilde{\mathbf{X}}_k$, with $h_k := 1/\sigma_k$, so that:

$$\boldsymbol{\mathcal{X}} = \left(\boldsymbol{\mu}\boldsymbol{y}^\top + \mathbf{Z}\right) \otimes \boldsymbol{h} + \boldsymbol{\mathcal{W}} \in \mathbb{R}^{p\times n\times m}, \qquad \text{(Nested Matrix-Tensor Model)}$$

where the tensor $\boldsymbol{\mathcal{W}}$ has independent standard Gaussian entries and $\boldsymbol{h} = (h_1, \ldots, h_m)^\top \in \mathbb{R}^m$.

The above model can be seen as a more general version of the rank-one spiked model that incorporates a nested structure allowing for more flexible modeling (Specifically, when the variances of the elements in $\mathbf{Z}$ tend to zero, one recovers the rank-one spiked model). The common low-rank structure in the slices $\mathbf{X}_k$, which can be interpreted as different views of the data, encodes the latent clustering structure that can then be retrieved by using tensor methods applied on $\boldsymbol{\mathcal{X}}$.

In particular, our results precisely quantify the asymptotic performance of a simple estimator of the vectors $\boldsymbol{\mu}$, $\boldsymbol{y}$, and $\boldsymbol{h}$ based on rank-one approximation of $\boldsymbol{\mathcal{X}}$, in the large-dimensional limit where $p, n, m \to \infty$ at the same rate. This is achieved by resorting to the recently developed approach of Goulart et al. (2022) and Seddik et al. (2021), which allows one to use tools from random matrix theory by inspecting *contractions* of the random tensor model in question. Numerical results are given to illustrate the usefulness of such predictions even for moderately large values of p and n, and also to show the superiority of such a tensor-based approach in comparison with a naive spectral method that does not take the tensor structure of the model into account. Quite interestingly, our results show that the performance of such a rank-one spectral estimator exhibits different phase transition behaviors depending on two parameters governing the SNR and the data dispersion, effectively "interpolating" between phase transition curves that are characteristic of matrix and tensor models.

Key contributions: Our main contributions can be summarized as follows:

1. We introduce a nested matrix-tensor model that generalizes the (third-order) spiked tensor model, and we provide a random matrix analysis of its best rank-one tensor approximation in the high-dimensional regime.
2. We provide an application of this model to the problem of clustering multi-view data and show that the developed theory allows the exact characterization of the asymptotic performance of a multi-view clustering approach. Further simulations suggest the superiority of the tensor-based clustering approach compared to a naive unfolding method that ignores the hidden rank-one structure.

Related work on tensor multi-view methods: In multi-view machine learning (Xu et al., 2013; Zhao et al., 2017; Sun, 2013), one has to deal with data coming from different sources or exhibiting various statistical or physical natures (e.g. documents composed of both text and images). The main challenge consists in jointly leveraging both the agreement and the complementarity of the different views (Blum & Mitchell, 1998; Dasgupta et al., 2001; Nigam & Ghani, 2000), e.g. via learning a shared latent subspace (White et al., 2012) for diverse tasks such as regression (Kakade & Foster, 2007) or clustering (Chaudhuri et al., 2009; Gao et al., 2015; Cao et al., 2015). In this context, multi-view clustering algorithms using a low-rank tensor representation of the multi-view data have already been proposed: among others, Xie et al. (2018); Wu et al. (2020) relied on tensor-SVD (Kilmer et al., 2013) while Liu et al. (2013) favored a Tucker-type tensor decomposition.

However, the usual sense employed for the term "multi-view clustering" is not exactly the same that we adopt here, since in our problem all views essentially hold noisy measurements of the same

quantities. Hence, our work is perhaps closer in spirit to certain tensor-based clustering models comprising an additional diversity (e.g., temporal), such as those of Papalexakis et al. (2012) or those reviewed in Fernandes et al. (2021). Yet, it differs from this literature in that our additional diversity is quite specific (namely, it comes from the availability of multiple measurements for each individual in the sample) and, furthermore, we derive the exact asymptotic performance of our proposed tensor-based method in the large-dimensional limit.

2 Notation and Background

The set $\{1, \dots, n\}$ is denoted by $[n]$. The unit sphere in $\mathbb{R}^p$ is denoted by $\mathbb{S}^{p-1}$. The Dirac measure at some real value x is denoted by δ_x. The support of a measure ν is denoted by $\mathrm{Supp}(\nu)$. The inner product between two vectors $\boldsymbol{u}, \boldsymbol{v}$ is denoted by $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_i u_i v_i$. The imaginary part of a complex number z is denoted by $\Im[z]$. The set of eigenvalues of a matrix $\mathbf{M}$ is denoted by $\mathrm{Sp}(\mathbf{M})$. Almost sure convergence of a sequence of random variables is denoted by $\xrightarrow{\text{a.s.}}$. The arrow $\xrightarrow{\mathcal{D}}$ denotes the convergence in distribution.

2.1 Tensor Notations and Contractions

In this section, we introduce the main tensor notations and definitions used throughout the paper, which we recommend following carefully for a clear understanding of its technical contents.

Three-order tensors: The set of third-order tensors of size $n_1 \times n_2 \times n_3$ is denoted $\mathbb{R}^{n_1 \times n_2 \times n_3}$. The scalar T_{ijk} or $[\mathbf{T}]_{ijk}$ denotes the (i, j, k) entry of a tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$.

Rank-one tensors: A tensor $\mathbf{T}$ is said to be of rank-one if it can be represented as the outer product of three real-valued vectors $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^{n_3}$. In this case, we write $\mathbf{T} = \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z}$, where the outer product is defined such that $[\boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z}]_{ijk} = x_i y_j z_k$.

Tensor contractions: The first mode contraction of a tensor $\mathbf{T}$ with a vector $\boldsymbol{x}$ yields a matrix denoted $\mathbf{T}(\boldsymbol{x}, \cdot, \cdot)$ with entries $[\mathbf{T}(\boldsymbol{x}, \cdot, \cdot)]_{jk} = \sum_{i=1}^{n_1} x_i \mathsf{T}_{ijk}$. Similarly, $\mathbf{T}(\cdot, \boldsymbol{y}, \cdot)$ and $\mathbf{T}(\cdot, \cdot, \boldsymbol{z})$ denote the second and third mode contractions of $\mathbf{T}$ with vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ respectively. We will sometimes denote these contractions by $\mathbf{T}(\boldsymbol{x})$, $\mathbf{T}(\boldsymbol{y})$, and $\mathbf{T}(\boldsymbol{z})$ if there is no ambiguity. The contraction of $\mathbf{T}$ with two vectors $\boldsymbol{x}, \boldsymbol{y}$ is a vector denoted $\mathbf{T}(\boldsymbol{x}, \boldsymbol{y}, \cdot)$ with entries $[\mathbf{T}(\boldsymbol{x}, \boldsymbol{y}, \cdot)]_k = \sum_{ij} x_i y_j \mathsf{T}_{ijk}$. Similarly, the contraction of $\mathbf{T}$ with three vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ is a scalar denoted $\mathbf{T}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{ijk} x_i y_j z_k \mathsf{T}_{ijk}$.

Tensor norms: The Frobenius norm of a tensor $\mathbf{T}$ is denoted $\|\mathbf{T}\|_F$ with $\|\mathbf{T}\|_F^2 = \sum_{ijk} \mathsf{T}_{ijk}^2$. The spectral norm of $\mathbf{T}$ is $\|\mathbf{T}\| = \sup_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=\|\boldsymbol{w}\|=1} |\mathbf{T}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})|$.

Best rank-one approximation: A best rank-one approximation of $\mathbf{T}$ corresponds to a rank-one tensor of the form $\lambda \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}$, where $\lambda > 0$ and $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ are unitary vectors, that minimizes the square loss $\|\mathbf{T} - \lambda \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}\|_F^2$. The latter generalizes to tensors the concept of singular value and vectors (Lim, 2005) and the scalar λ coincides with the spectral norm of $\mathbf{T}$. Such a best rank-one approximation can be computed via *tensor power iteration* which consists of iterating:

$$\boldsymbol{u} \leftarrow \mathbf{T}(\cdot, \boldsymbol{v}, \boldsymbol{w}) / \|\mathbf{T}(\cdot, \boldsymbol{v}, \boldsymbol{w})\|, \quad \boldsymbol{v} \leftarrow \mathbf{T}(\boldsymbol{u}, \cdot, \boldsymbol{w}) / \|\mathbf{T}(\boldsymbol{u}, \cdot, \boldsymbol{w})\|, \quad \boldsymbol{w} \leftarrow \mathbf{T}(\boldsymbol{u}, \boldsymbol{v}, \cdot) / \|\mathbf{T}(\boldsymbol{u}, \boldsymbol{v}, \cdot)\|,$$

starting from some appropriate initialization (Kofidis & Regalia, 2002; Anandkumar et al., 2014).

2.2 Random Matrix Theory

In this section, we recall some necessary tools from random matrix theory (RMT) which are at the core of our main results. Specifically, we will consider the *resolvent* formalism of Hachem et al. (2007) which allows one to characterize the spectral behavior of large symmetric random matrices and the estimation of low-dimensional functionals of such matrices. Given a symmetric matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$, the resolvent of $\mathbf{S}$ is defined as $\mathbf{R}(\xi) = (\mathbf{S} - \xi \boldsymbol{I}_n)^{-1}$ for some $\xi \in \mathbb{C} \setminus \mathrm{Sp}(\mathbf{S})$.

In essence, RMT focuses on describing the distribution of eigenvalues of large random matrices. Typically, under certain technical assumptions on some random matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \dots, \lambda_n$, the *empirical spectral measure* of $\mathbf{S}$, defined as $\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i}$, converges in the weak sense (Van Der Vaart & Wellner, 1996) to some deterministic probability measure ν as $n \to \infty$ and RMT aims at describing such a ν. To this end, one widely considered (so-called analytical)

approach relies on the *Stieltjes transform* (Widder, 1938). Given a probability measure ν, the Stieltjes transform of ν is defined as $g_\nu(\xi) = \int \frac{d\nu(\lambda)}{\lambda - \xi}$ with $\xi \in \mathbb{C} \setminus \mathrm{Supp}(\nu)$, and the inverse formula allows one to describe the density of ν as $\nu(dx) = \frac{1}{\pi} \lim_{\varepsilon \to 0} \Im[g_\nu(x + i\varepsilon)]$ (assuming it admits one).

The Stieltjes transform of the empirical spectral measure, $\hat{\nu}_n$, is closely related to the resolvent of $\mathbf{S}$ through the normalized trace operator. In fact, $g_{\hat{\nu}_n}(\xi) = \frac{1}{n} \operatorname{Tr} \mathbf{R}(\xi)$ and the point-wise *almost sure* convergence of $g_{\hat{\nu}_n}(\xi)$ to some deterministic Stieltjes transform $g_\nu(\xi)$ (where ν is defined on $\mathbb{R}$) on the upper-half complex plane is equivalent to the weak convergence of $\hat{\nu}_n$ to ν (Tao, 2012). Our analysis relies on estimating quantities involving $\frac{1}{n} \operatorname{Tr} \mathbf{R}(\xi)$, making the use of the resolvent approach a natural choice (see Appendix A for the derivation of our results).

3 Main Results

3.1 The Nested Matrix-Tensor Model

We start by defining our considered nested matrix-tensor model in a general form since it might have applications beyond the multi-view data model in Eq. (Nested Matrix-Tensor Model). Let $n_1, n_2, n_3 \in \mathbb{N}_+$ and further denote $n_M = n_1 + n_2$ and $n_T = n_1 + n_2 + n_3$. We consider the following statistical model:

$$\mathbf{T} = \beta_T \mathbf{M} \otimes \boldsymbol{z} + \frac{1}{\sqrt{n_T}} \mathbf{W} \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \quad \mathbf{M} = \beta_M \boldsymbol{x} \otimes \boldsymbol{y} + \frac{1}{\sqrt{n_M}} \mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}, \tag{1}$$

where we assume that $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = \|\boldsymbol{z}\| = 1$ and that the entries of $\mathbf{W}$ and $\mathbf{Z}$ are independent Gaussian random variables, with $\boldsymbol{W}_{ijk} \sim \mathcal{N}(0, \sigma_T^2)$ and $\boldsymbol{Z}_{ij} \sim \mathcal{N}(0, \sigma_M^2)$. For the sake of simplicity, we consider the unit variance case $\sigma_T = \sigma_M = 1$ in the remainder of the paper while we defer the general variance case to Appendix A.

Remark 1 (Spectral normalization) *Note that the normalization of $\mathbf{W}$ by $\sqrt{n_T}$ (resp. $\boldsymbol{Z}$ by $\sqrt{n_M}$) in Eq. (1) ensures that the spectral norm of $\mathbf{T}$ is of order $O(1)$ when the dimensions n_i grow to infinity. This follows from a standard concentration result (Seddik et al., 2021, Lemma 4).*

Best rank-one tensor estimator: We consider the analysis of the best rank-one approximation of $\mathbf{T}$ which corresponds to the following problem (Lim, 2005):

$$\underset{\lambda > 0,\, \|\boldsymbol{u}\|\|\boldsymbol{v}\|\|\boldsymbol{w}\| = 1}{\arg\min} \|\mathbf{T} - \lambda \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}\|_{\mathrm{F}}^2 \quad \Leftrightarrow \quad \underset{\|\boldsymbol{u}\|\|\boldsymbol{v}\|\|\boldsymbol{w}\| = 1}{\arg\max} \mathbf{T}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \,. \tag{2}$$

In particular, the solution for the scalar λ in the left-hand problem coincides with the spectral norm of $\mathbf{T}$, i.e., $\lambda = \|\mathbf{T}\|$. Given a critical point $(\lambda, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ of that problem, it holds that (Lim, 2005):

$$\mathbf{T}(\cdot, \boldsymbol{v}, \boldsymbol{w}) = \lambda \boldsymbol{u}, \quad \mathbf{T}(\boldsymbol{u}, \cdot, \boldsymbol{w}) = \lambda \boldsymbol{v}, \quad \mathbf{T}(\boldsymbol{u}, \boldsymbol{v}, \cdot) = \lambda \boldsymbol{w}, \quad \lambda = \mathbf{T}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}). \tag{3}$$

In essence, for sufficiently large β_M and β_T, the triplet $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ will start to align with the signal components $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ and our main goal is to quantify these alignments (i.e., the inner products $\langle \boldsymbol{u}, \boldsymbol{x} \rangle$, $\langle \boldsymbol{v}, \boldsymbol{y} \rangle$ and $\langle \boldsymbol{w}, \boldsymbol{z} \rangle$) in the large dimensional regime when $n_i \to \infty$. To this end, we need a typical set of assumptions that we formulate as follows (see (Goulart et al., 2022; Seddik et al., 2021) for similar assumptions in the case of spiked random tensors).

Assumption 1 *There exists a sequence of critical points $(\lambda, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ satisfying Eq. (3) such that, when $n_i \to \infty$ with $\frac{n_1}{n_T} \to c_1 > 0$, $\frac{n_2}{n_T} \to c_2 > 0$, $\frac{n_3}{n_T} \to c_3 > 0$, we have the following:*

$$\lambda \xrightarrow{a.s.} \bar{\lambda}, \quad |\langle \boldsymbol{u}, \boldsymbol{x} \rangle| \xrightarrow{a.s.} \alpha_1, \quad |\langle \boldsymbol{v}, \boldsymbol{y} \rangle| \xrightarrow{a.s.} \alpha_2, \quad |\langle \boldsymbol{w}, \boldsymbol{z} \rangle| \xrightarrow{a.s.} \alpha_3.$$

In the remainder of the paper, we refer to the quantities $(\lambda, \langle \boldsymbol{u}, \boldsymbol{x} \rangle, \langle \boldsymbol{v}, \boldsymbol{y} \rangle, \langle \boldsymbol{w}, \boldsymbol{z} \rangle)$ as *summary statistics* as per the formalism introduced by Ben Arous et al. (2022) since the asymptotic limits of these scalar quantities fully describe the asymptotic behavior of the considered best rank-one tensor estimator applied to $\mathbf{T}$.

Remark 2 (On Assumption 1) *The almost sure convergence of the summary statistics has been demonstrated in (Jagannath et al., 2020) in the case of the spiked tensor model. We believe similar arguments can be extended to our proposed nested matrix-tensor model to validate Assumption 1.*

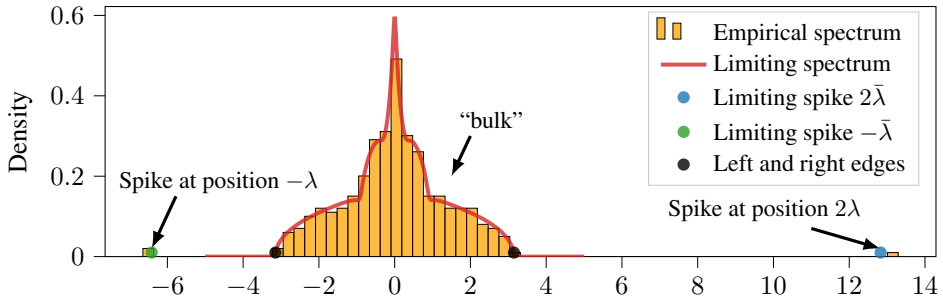

Figure 1: Empirical versus limiting spectrum of $\mathbf{\Phi}$ for $\beta_T = 2, \beta_M = 3, n_1 = 130, n_2 = 80, n_3 = 140$. In addition to the "bulk" of eigenvalues, the spectrum of $\mathbf{\Phi}$ exhibits two isolated spikes at positions 2λ and $-\lambda$ as discussed in Remark 3. In particular, the isolated spikes are accurately estimated by the limiting singular value $\bar{\lambda}$ as per Theorem 2 and Algorithm 2.

3.2 Associated Random Matrix

As discussed in the previous section, our primary goal is to compute the asymptotic summary statistics $(\bar{\lambda}, \alpha_1, \alpha_2, \alpha_3)$ in terms of the model's parameters, namely, the signal-to-noise ratios (β_M, β_T) and the dimension ratios (c_1, c_2, c_3). To this end, we follow the approach developed by Seddik et al. (2021), who studied the *asymmetric* spiked tensor model, and where it has been shown that the estimation of $(\bar{\lambda}, \alpha_1, \alpha_2, \alpha_3)$ boils down to the analysis of the *block-wise contraction random matrix* $\mathbf{\Phi}$ in Eq.(4), which can be done by deploying tools from random matrix theory.

Given the model in Eq. (1), it can be easily noticed that $\mathbf{\Phi}$ decomposes as a sum of two matrices $\mathbf{H} + \mathbf{L}$ where $\mathbf{L}$ is a low-rank matrix related to the signal part in the nested matrix-tensor model (the expression of $\mathbf{L}$ is provided in Eq. (17) in Appendix A), and $\mathbf{H}$ corresponds to the noise part of the model, being given by:

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{0}_{n_1 \times n_1} & \mathbf{T}(\boldsymbol{w}) & \mathbf{T}(\boldsymbol{v}) \\ \mathbf{T}(\boldsymbol{w})^\top & \mathbf{0}_{n_2 \times n_2} & \mathbf{T}(\boldsymbol{u}) \\ \mathbf{T}(\boldsymbol{v})^\top & \mathbf{T}(\boldsymbol{u})^\top & \mathbf{0}_{n_3 \times n_3} \end{bmatrix} \tag{4}$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{0}_{n_1 \times n_1} & \frac{\langle \boldsymbol{w}, \boldsymbol{z} \rangle \beta_T}{\sqrt{n_M}} \mathbf{Z} + \frac{1}{\sqrt{n_T}} \mathbf{W}(\boldsymbol{w}) & \frac{1}{\sqrt{n_T}} \mathbf{W}(\boldsymbol{v}) \\ \frac{\langle \boldsymbol{w}, \boldsymbol{z} \rangle \beta_T}{\sqrt{n_M}} \mathbf{Z}^\top + \frac{1}{\sqrt{n_T}} \mathbf{W}(\boldsymbol{w})^\top & \mathbf{0}_{n_2 \times n_2} & \frac{1}{\sqrt{n_T}} \mathbf{W}(\boldsymbol{u}) \\ \frac{1}{\sqrt{n_T}} \mathbf{W}(\boldsymbol{v})^\top & \frac{1}{\sqrt{n_T}} \mathbf{W}(\boldsymbol{u})^\top & \mathbf{0}_{n_3 \times n_3} \end{bmatrix}. \tag{5}$$

Remark 3 (On the spectrum of $\mathbf{\Phi}$) *In terms of spectrum, we will see subsequently that the matrices $\mathbf{\Phi}$ and $\mathbf{H}$ share the same "bulk" of eigenvalues while the spectrum of $\mathbf{\Phi}$ exhibits two isolated eigenvalues at positions 2λ and $-\lambda$ if β_M, β_T are large enough. In fact, one can quickly check that, given the identities in Eq. (3), the scalars 2λ and $-\lambda$ are eigenvalues of $\mathbf{\Phi}$ with respective multiplicities 1 and 2, and respective eigenvectors $(\boldsymbol{u}^\top, \boldsymbol{v}^\top, \boldsymbol{w}^\top)^\top$ for the eigenvalue 2λ and $(\boldsymbol{u}^\top, \mathbf{0}^\top, -\boldsymbol{w}^\top)^\top, (\mathbf{0}^\top, \boldsymbol{v}^\top, -\boldsymbol{w}^\top)^\top$ corresponding to the eigenvalue $-\lambda$.*

3.3 Limiting Spectrum

We will find subsequently that the asymptotic summary statistics $(\bar{\lambda}, \alpha_1, \alpha_2, \alpha_3)$ are closely related to the limiting spectral measure of the random matrix $\mathbf{H}$. Therefore, our first result characterizes precisely this limiting distribution using the Stieltjes transform formalism (Widder, 1938).

Theorem 1 (Limiting spectrum) *Under Assumption 1, the empirical spectral measure of $\mathbf{H}$ or $\mathbf{\Phi}$ converges weakly almost surely to a deterministic distribution ν whose Stieltjes transform is given by $g(\xi) = \sum_{i=1}^{3} g_i(\xi)$ such that $\Im[g(\xi)] > 0$ for $\Im[\xi] > 0$, and where $(g_i(\xi))_{i \in [3]}$ satisfy the following equations:*

$$g_1(\xi) = \frac{c_1}{g_1(\xi) - g(\xi) - \bar{\gamma} g_2(\xi) - \xi}, \quad g_2(\xi) = \frac{c_2}{g_2(\xi) - g(\xi) - \bar{\gamma} g_1(\xi) - \xi}, \quad g_3(\xi) = \frac{c_3}{g_3(\xi) - g(\xi) - \xi},$$

with $\bar{\gamma} = \frac{\beta_T^2 \alpha_3^2}{c_1 + c_2}$. In particular, the density function of ν is given by $\nu(dx) = \frac{1}{\pi} \lim_{\varepsilon \to 0} \Im\left[g(x + i\varepsilon)\right]$.

Theorem 1 generalizes the limiting spectral measure obtained by (Seddik et al., 2021) in the sense that the latter corresponds to the particular case when $\bar{\gamma} = 0$ (e.g. if $\beta_T = 0$). Moreover, in the

Algorithm 1 Limiting Stieltjes transform as per Theorem 1

Input: Complex number $\xi \in \mathbb{C} \setminus \mathrm{Supp}(\nu)$, ratios $c_1, c_2, c_3 \in [0,1]$, $\beta_T, \beta_M \geq 0$ and `option`.
Output: Limiting Stieltjes transform $g(\xi)$ and $g_i(\xi)$ for $i \in [3]$.
Initialize g_1, g_2, g_3 and set $g \leftarrow g_1 + g_2 + g_3$.
if `option` **is** *"compute $\bar{\gamma}$"* **then**
  Compute the asymptotic summary statistics $(\bar{\lambda}, \alpha_1, \alpha_2, \alpha_3)$ with Algo. 2 and set $\bar{\gamma} \leftarrow \frac{\beta_T^2 \alpha_3^2}{c_1 + c_2}$.
end if
while *"g has not converged"* **do**
  if `option` **is** *"approximate $\bar{\gamma}$"* **then**
    Update $\bar{\gamma} \leftarrow \frac{\beta_T^2}{c_1+c_2}\left(1 - \frac{g_3^2}{c_3}\right)$.
  end if
  Update $g_1 \leftarrow \frac{c_1}{g_1 - g - \bar{\gamma} g_2 - \xi}$, $g_2 \leftarrow \frac{c_2}{g_2 - g - \bar{\gamma} g_1 - \xi}$, $g_3 \leftarrow \frac{c_3}{g_3 - g - \xi}$, $g \leftarrow g_1 + g_2 + g_3$.
end while

specific case $\beta_T = 0$ and $c_1 = c_2 = c_3 = \frac{1}{3}$, the distribution ν describes a *semi-circle law* of compact support $[-2\sqrt{2/3}, 2\sqrt{2/3}]$, and the corresponding Stieltjes transform is explicitly given by $g(\xi) = \frac{3}{4}(-\xi + \sqrt{\xi^2 - 8/3})$ with $g_i(\xi) = g(\xi)/3$ for all $i \in [3]$. We refer the reader to (Seddik et al., 2021) for more details and a full description of various particular cases. Moreover, an explicit formula for $g(\xi)$ can be derived in the case $c_1 = c_2$ using a formal calculation tool (e.g. SymPy).

However, for arbitrary values of β_T, β_M and of the dimension ratios (c_1, c_2, c_3), the limiting spectral measure of $\mathbf{H}$ or $\mathbf{\Phi}$ can be computed numerically as per Algorithm 1 which implements the equations in Theorem 1. Figure 1 shows that the empirical spectral measure of $\mathbf{\Phi}$ is accurately predicted by the limiting measure of Theorem 1 (further examples are depicted in Figure 6 in the Appendix). We note that the computation of $\bar{\gamma}$ (which is closely related to the alignment α_3) is a key step in the numerical evaluation of g, which we will address next by computing the asymptotic alignments α_i's.

3.4 Asymptotic Summary Statistics

In the previous subsection, we have shown that the empirical spectral measure of $\mathbf{H}$ or $\mathbf{\Phi}$ converges to some deterministic measure ν as we depicted in Figure 1. Specifically, we notice that the measure ν has a compact support that depends on the various parameters of the model. In what follows, we will need to evaluate the corresponding Stieltjes transform g at the asymptotic spectral norm $\bar{\lambda}$, and therefore the latter must lie outside the support of ν as per the following assumption. In fact, this assumption has also been made by (Goulart et al., 2022; Seddik et al., 2021).

Assumption 2 *Assume that $\bar{\lambda} \notin \mathrm{Supp}(\nu)$ and $\alpha_i > 0$ for all $i \in [3]$, with ν given by Theorem 1.*

Remark 4 (On Assumption 2) *For any critical point $(\lambda, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ of problem (2), as we saw in Remark 3, $\mathbf{\Phi}$ has an eigenvalue 2λ. In particular, for a local maximum, 2λ is in fact its largest eigenvalue (Seddik et al., 2021). Furthermore, by studying the Hessian of that problem (which is related to $\mathbf{\Phi}$) at a maximum one can also show that λ is at least as large as the second largest eigenvalue of $\mathbf{\Phi}$ (which is almost surely close to the right edge of the measure ν). Hence, the above condition in Assumption 2 is slightly stronger, only requiring that inequality to hold strictly. See also (Goulart et al., 2022) for a similar discussion in the case of a symmetric spiked tensor model.*

We are now in place to provide our main result which characterizes the asymptotic summary statistics $(\bar{\lambda}, \alpha_1, \alpha_2, \alpha_3)$ given the signal-to-noise ratios (β_M, β_T) and the dimension ratios (c_1, c_2, c_3).

Theorem 2 (Asymptotic summary statistics) *Let us define the following functions for $i \in [2]$:*

$$q_i(\xi) = \sqrt{1 - \frac{[1+\gamma(\xi)] g_i^2(\xi)}{c_i}}, \quad q_3(\xi) = \sqrt{1 - \frac{g_3^2(\xi)}{c_3}}, \quad \gamma(\xi) = \frac{\beta_T^2 q_3^2(\xi)}{c_1 + c_2},$$

$$f(\xi) = \xi + [1+\gamma(\xi)] g(\xi) - \gamma(\xi) g_3(\xi) - \beta_T \beta_M \prod_{i=1}^{3} q_i(\xi).$$

Then, under Assumptions 1 and 2, the asymptotic spectral norm $\bar{\lambda}$ satisfies $f(\bar{\lambda}) = 0$ and the asymptotic alignments are given by $\alpha_i = q_i(\bar{\lambda})$ (in particular, $\bar{\gamma} = \gamma(\bar{\lambda})$).

Algorithm 2 Asymptotic summary statistics as per Theorem 2

Input: Dimension ratios $c_1, c_2, c_3 \in [0, 1]$ and signal-to-noise ratios $\beta_T, \beta_M \geq 0$.
Output: Asymptotic summary statistics $(\bar{\lambda}, \alpha_1, \alpha_2, \alpha_3)$.

Define $q_i(\xi) = \sqrt{1 - \frac{[1+\gamma(\xi)]g_i^2(\xi)}{c_i}}$ for $i \in [2]$, $q_3(\xi) = \sqrt{1 - \frac{g_3^2(\xi)}{c_3}}$ and $\gamma(\xi) = \frac{\beta_T^2 q_3^2(\xi)}{c_1+c_2}$ where $(g_i(\xi))_{i\in[3]}$ and $g(\xi)$ are obtained by Algorithm 1 for some $\xi \in \mathbb{R} \setminus \mathrm{Supp}(\nu)$ by setting the parameter `option` **to** *"approximate $\bar{\gamma}$"*.

Define the function f as $f(\xi) = \xi + [1 + \gamma(\xi)]g(\xi) - \gamma(\xi)g_3(\xi) - \beta_T\beta_M \prod_{i=1}^{3} q_i(\xi)$.

Solve $f(\bar{\lambda}) = 0$ where the root corresponds to $\bar{\lambda}$ and set $\alpha_i \leftarrow q_i(\bar{\lambda})$ for $i \in [3]$.

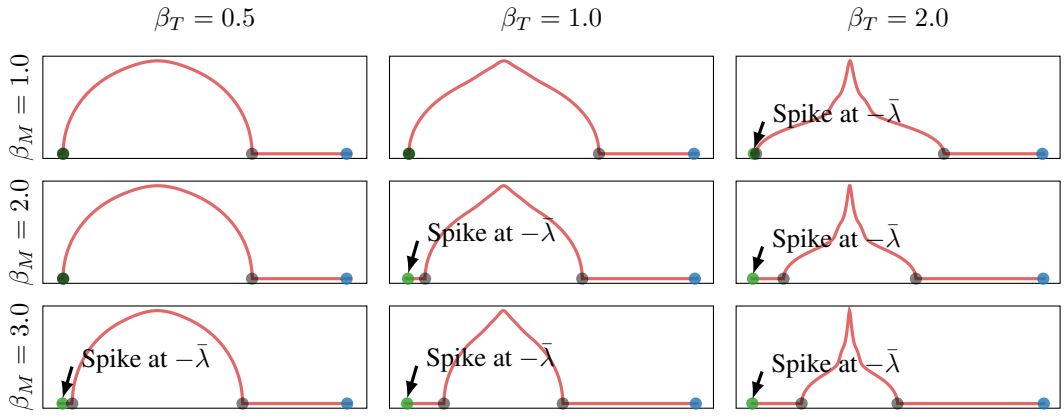

Figure 2: Limiting spectrum and isolated spikes of $\mathbf{\Phi}$ for $n_1 = 80, n_2 = 100, n_3 = 90$ and varying (β_T, β_M). For small values of (β_T, β_M) the "bulk" is described by a semi-circle-like distribution. Large values of β_T affect the "shape" of the limiting measure while larger values of β_M control the position of the isolated spikes.

Theorems 1 and 2 show that the spectral behavior of the random matrix $\mathbf{\Phi}$ is fully described by its limiting spectral measure ν and the position of the limiting singular value $\bar{\lambda}$. This is illustrated by Figure 1 which depicts the empirical spectrum of $\mathbf{\Phi}$ along with the limiting measure ν as per Theorem 1 and the asymptotic spikes computed via Theorem 2. As we discussed earlier in Remark 3, the spectrum of $\mathbf{\Phi}$ consists of a "bulk" of eigenvalues spread around 0 and two isolated eigenvalues at positions 2λ and $-\lambda$ with multiplicities 1 and 2 respectively. In fact, the spike at position $-\lambda$ is only visible when the signal-to-noise ratios (β_M, β_T) are large enough, and this basically corresponds to the situation where it is theoretically possible to estimate the signal components $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ from the tensor $\mathbf{T}$. In addition, note that Assumption 2 holds when a spike is visible at the position $-\lambda$. This *phase transition* phenomenon is highlighted in Figure 2 where we vary the signal-to-noise ratios (β_M, β_T). In particular, roughly speaking, the parameter β_T affects the "shape" of limiting distribution ν while β_M determines the position of the isolated spikes. Besides, note that in the situations where $\bar{\lambda}$ lies inside the support of ν, we solve numerically the equation $f(\bar{\lambda} + i\varepsilon) = 0$ for some small value ε (and take the real parts of g and g_i's), which allows us to circumvent Assumption 2 in this case.

Figure 3 in turn depicts the empirical versus asymptotic summary statistics when varying the parameter β_M (with β_T being fixed) and shows that the empirical quantities are accurately predicted by the theoretical counterparts. Moreover, as in standard spiked random matrix models, our results show that there exists a *phase transition*, i.e., a minimum value for β_M above which the singular vectors along the modes 1 and 2 $(\boldsymbol{u}, \boldsymbol{v})$ start to correlate with the matrix signal components $(\boldsymbol{x}, \boldsymbol{y})$. However, below this critical value of β_M, α_1 and α_2 are vanishing while $\alpha_3 \approx 1$. The continuity of the curves of α_1 and α_2 when varying β_M is a typical characteristic of spiked matrices as per the classical BBP phase transition phenomenon (Baik et al., 2005). Besides, for smaller values of β_T (below some critical value), the curves of α_1 and α_2 start to become discontinuous as per Figure 7 in Appendix B which is commonly observed in spiked tensor models (Jagannath et al., 2020). In this sense, the nested matrix-tensor model is a sort of "interpolating model" between spiked matrices and tensors (see Appendix B for additional simulations), as far as a spectral estimator of the spike is concerned.

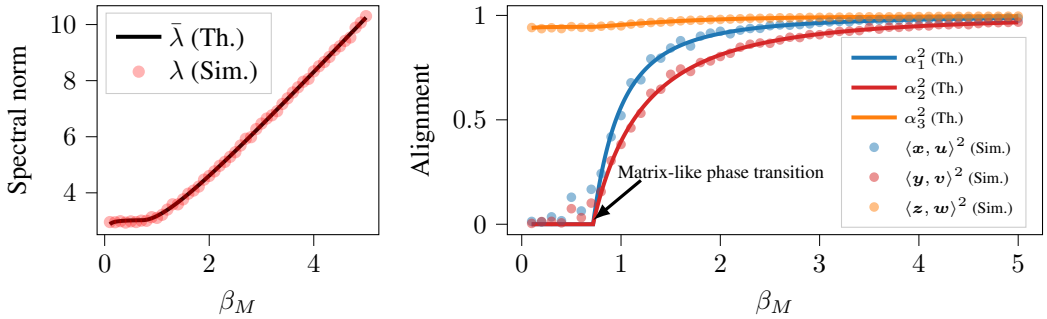

Figure 3: Empirical versus asymptotic summary statistics for $n_1 = 40, n_2 = 110, n_3 = 90, \beta_T = 2$ and varying β_M. Simulations are obtained by averaging over 10 independent realizations of the random matrix $\mathbf{Z}$ and tensor $\mathbf{W}$. Our results exhibit a phase transition when varying β_M above which the matrix components $(\boldsymbol{x}, \boldsymbol{y})$ become estimable.

Remark 5 (Computation of α_3 below the phase transition) *Even though Assumption 2 is not valid in the regime where β_M is below its critical value (because an isolated spike at position $-\bar{\lambda}$ outside the support of ν is not present in this case), numerically computed solutions for $f(\bar{\lambda} + i\varepsilon) = 0$ with a small $\varepsilon > 0$ seem to accurately estimate α_3 as per Fig. 3 (whereas $f(\xi)$ is not defined at $\bar{\lambda}$ since it depends on $g(\xi)$, which is undefined inside the support of ν). Yet, we currently do not have a rigorous justification for this intriguing property.*

4 Application to Multi-view Clustering

Now we illustrate the application of Theorem 2 to the assessment of the performance of a simple multi-view spectral clustering approach. As we presented in the introduction, we consider that we observe a tensor $\mathbf{X}$ of n data points of dimension p along m different views:

$$\mathbf{X} = \left(\boldsymbol{\mu}\bar{\boldsymbol{y}}^\top + \mathbf{Z}\right) \otimes \boldsymbol{h} + \mathbf{W}, \quad Z_{ij} \sim \mathcal{N}\left(0, \frac{1}{p+n}\right), \; W_{ijk} \sim \mathcal{N}\left(0, \frac{1}{p+n+m}\right), \tag{6}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ models the cluster means ($-\boldsymbol{\mu}$ or $\boldsymbol{\mu}$), $\bar{\boldsymbol{y}} = \boldsymbol{y}/\sqrt{n}$ with $\boldsymbol{y} \in \{-1, 1\}^n$ corresponding to the data labels (either -1 or 1) and $\boldsymbol{h} \in \mathbb{R}_+^m$ is related to the variances along the different views. In particular, the case $m = 1$ corresponds to the classical binary Gaussian isotropic model of centroids $\pm\boldsymbol{\mu}$ in which case the tensor $\mathbf{X}$ becomes a matrix of the form $\mathbf{X} = \boldsymbol{\mu}\bar{\boldsymbol{y}}^\top + \mathbf{Z}$. Figure 4 depicts the multi-view model in Eq. (6) for $p = 2$ and $m = 4$, where the first class is represented by dots and the second class is depicted by crosses, while the different views are illustrated with different colors. Observing the tensor $\mathbf{X}$, the clustering of the different data points would consist in estimating the labels vector $\boldsymbol{y}$. Indeed, this can be performed by computing the best rank-one approximation of $\mathbf{X}$ (denoted $\lambda \boldsymbol{u} \otimes \hat{\boldsymbol{y}} \otimes \boldsymbol{w}$), and depending on the class separability condition (i.e. if $\|\boldsymbol{\mu}\|$ and $\|\boldsymbol{h}\|$ are large enough), the 2-mode singular vector of $\mathbf{X}$ will start to correlate with $\boldsymbol{y}$ thereby providing a clustering of the data samples. Our aim is to quantify the performance of this multi-view spectral clustering approach in terms of the different parameters, i.e., the dimensions n, p, m and the quantities $\|\boldsymbol{\mu}\|$ and $\|\boldsymbol{h}\|$. We precisely have the subsequent proposition which characterizes the theoretical performance of the multi-view spectral clustering method under the following growth rate assumptions.

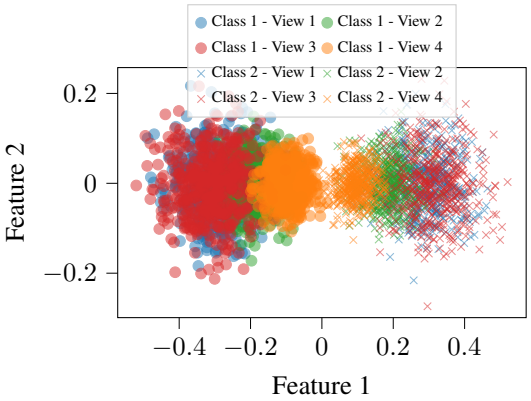

Figure 4: Illustration of the multi-view model in Eq. (6) for $p = 2, n = 1000, m = 4, \|\boldsymbol{\mu}\| = 5$ and $\|\boldsymbol{h}\| = 3$. The first class is represented by dots and the second class by crosses. The different colors represent the views.

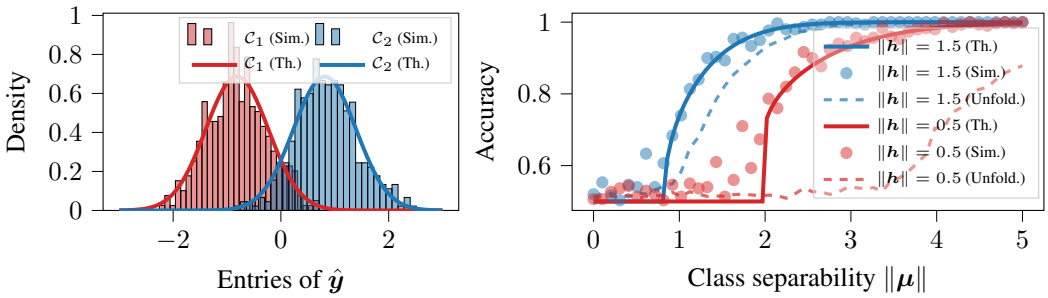

Figure 5: **(Left)** Histogram of the entries of $\sqrt{n}\hat{\boldsymbol{y}}$ for $p = 200, n = 800, m = 100, \|\boldsymbol{\mu}\| = 1.5$ and $\|\boldsymbol{h}\| = 2$ with the corresponding Gaussian limit as per Proposition 1. **(Right)** Empirical versus theoretical multi-view clustering performance as per Proposition 1 for $p = 150, n = 300, m = 60$ and varying $\|\boldsymbol{\mu}\|, \|\boldsymbol{h}\|$. The dashed curves correspond to tensor unfolding which discards the rank-one structure of the data and therefore yields sub-optimal accuracy.

Assumption 3 (Growth rate) *Assume that as* $p, n, m \to \infty$, $\|\boldsymbol{\mu}\|, \|\boldsymbol{h}\| = O(1)$ *and denote* $c_p = \lim \frac{p}{N} > 0, c_n = \lim \frac{n}{N} > 0, c_m = \lim \frac{m}{N} > 0$ *with* $N = p + n + m$.

Proposition 1 (Performance of multi-view spectral clustering) *Let* $\hat{\boldsymbol{y}}$ *be the* 2^{nd} *mode vector of the best rank-one approximation of the data tensor* $\mathbf{X}$. *The estimated label for the sample* $\mathbf{X}_{:,i,j}$ *is given by* $\hat{\ell}_i = \operatorname{sign}(\hat{y}_i)$ *for all* $j \in [m]$ *and let* $\mathcal{L}_{0/1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\hat{\ell}_i \neq y_i\}$ *be the corresponding* $0/1$*-loss. We have under Assumption 3:*

$$(1-\alpha^2)^{-\frac{1}{2}}\left[\sqrt{n}\hat{y}_i - \alpha y_i\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\alpha = q_2(\bar{\lambda})$ *with* $q_2(\cdot)$ *and* $\bar{\lambda}$ *defined as per Theorem 2 for* $(c_1, c_2, c_3) = (c_p, c_n, c_m)$ *and* $(\beta_M, \beta_T) = (\|\boldsymbol{\mu}\|, \|\boldsymbol{h}\|)$. *Moreover, the clustering accuracy* $\max\left(\mathcal{L}_{0/1}, 1 - \mathcal{L}_{0/1}\right)$ *converges almost surely to* $\varphi\left(\alpha/\sqrt{1-\alpha^2}\right)$ *with* $\varphi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt$.

Proposition 1 states that the entries of the 2-mode singular vector corresponding to the largest singular value of $\mathbf{X}$ are Gaussian random variables, with mean and variance depending on the dimension ratios (c_p, c_n, c_m) and the parameters $(\|\boldsymbol{\mu}\|, \|\boldsymbol{h}\|)$ through the asymptotic alignment α obtained thanks to Theorem 2. In fact, Figure 5 (Left) illustrates this Gaussianity by depicting the entries of the vector $\sqrt{n}\hat{\boldsymbol{y}}$ and the corresponding normal distributions. Furthermore, the theoretical accuracy $\varphi\left(\alpha/\sqrt{1-\alpha^2}\right)$ is also depicted in Figure 5 (Right) from which we notice that the empirical performance is accurately anticipated. Essentially, for a fixed value of $\|\boldsymbol{h}\|$, our results show that there exists a minimal value of the class separability $\|\boldsymbol{\mu}\|$ below which the obtained accuracy is no better than a random guess, in fact, the such minimal value of $\|\boldsymbol{\mu}\|$ is related to the *phase transition phenomenon* discussed in the previous section. In addition, we highlight that the considered tensor-based multi-view clustering approach provides better accuracy compared to a tensor unfolding approach, which consists in computing the top left singular vector of the unfolding of $\mathbf{X}$ along the second mode (Ben Arous et al., 2021), and therefore does not consider the hidden rank-one structure.

5 Conclusion & Perspectives

We introduced the nested matrix-tensor model and provided a high-dimensional analysis of its best rank-one approximation, relying on random matrix theory. Our analysis has brought theoretical insights into the problem of muti-view clustering and demonstrates the ability of random matrix tools to assess the theoretical performance of the considered clustering method. This paves the way for an elaborated theoretical assessment and improvement of more sophisticated tensor-based methods. In particular, our present findings address only the case of binary clustering by considering the rank-one matrix model $\boldsymbol{\mu}\boldsymbol{y}^\top + \mathbf{Z}$ which can be extended to higher ranks, thereby modeling a multi-class problem. Besides, such an extension would require the analysis of more sophisticated tensor methods (e.g. the block-term decomposition (De Lathauwer, 2008)) which is more challenging compared to the present best rank-one estimator. Nevertheless, we believe our present work constitutes a fundamental basis for the development of more general results.

References

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. A tensor spectral approach to learning mixed membership community models. In Shalev-Shwartz, S. and Steinwart, I. (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 867–881, Princeton, NJ, USA, June 2013.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

Anandkumar, A., Ge, R., and Janzamin, M. Learning overcomplete latent variable models through tensor methods. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 36–112, Paris, France, 03–06 Jul 2015.

Auddy, A. and Yuan, M. On estimating rank-one spiked tensors in the presence of heavy tailed errors. *IEEE Transactions on Information Theory*, 2022.

Baik, J., Ben Arous, G., and Péché, S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. 2005.

Bakshi, A., Diakonikolas, I., Jia, H., Kane, D. M., Kothari, P. K., and Vempala, S. S. Robustly learning mixtures of k arbitrary Gaussians. In *Proceedings of the 54th Annual ACM Symposium on Theory of Computing*, pp. 1234–1247, Rome, Italy, June 2022.

Ben Arous, G., Huang, D. Z., and Huang, J. Long random matrices and tensor unfolding. *arXiv preprint arXiv:2110.10210*, 2021.

Ben Arous, G., Gheissari, R., and Jagannath, A. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

Cao, X., Zhang, C., Fu, H., Liu, S., and Zhang, H. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–594, 2015.

Chaudhuri, K., Kakade, S. M., Livescu, K., and Sridharan, K. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pp. 129–136, 2009.

Chen, W.-K., Handschy, M., and Lerman, G. Phase transition in random tensors with multiple independent spikes. *The Annals of Applied Probability*, 31(4):1868–1913, 2021.

Couillet, R. and Benaych-Georges, F. Kernel spectral clustering of large dimensional data. 2016.

Dasgupta, S., Littman, M., and McAllester, D. Pac generalization bounds for co-training. *Advances in neural information processing systems*, 14, 2001.

De Lathauwer, L. Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1033–1066, 2008.

Fernandes, S., Fanaee-T, H., and Gama, J. Tensor decomposition for analysing time-evolving social networks: An overview. *Artificial Intelligence Review*, 54:2891–2916, 2021.

Gao, H., Nie, F., Li, X., and Huang, H. Multi-view subspace clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 4238–4246, 2015.

Ge, R., Huang, Q., and Kakade, S. M. Learning mixtures of Gaussians in high dimensions. In *Proceedings of the 47th annual ACM Symposium on Theory of Computing*, pp. 761–770, Portland, OR, USA, June 2015.

Goulart, J. H. de M., Couillet, R., and Comon, P. A random matrix perspective on random tensors. *Journal on Machine Learning Research*, 23(264):1–36, 2022.

Gujral, E., Pasricha, R., and Papalexakis, E. Beyond rank-1: Discovering rich community structure in multi-aspect graphs. In *Proceedings of The Web Conference 2020*, pp. 452–462, Taipei, Taiwan, April 2020.

Hachem, W., Loubaton, P., and Najim, J. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

Hsu, D. and Kakade, S. M. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 11–20, Berkeley, CA, USA, January 2013.

Hsu, D., Kakade, S. M., and Zhang, T. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Huang, F., Niranjan, U. N., H., M. U., and Anandkumar, A. Online tensor methods for learning latent variable models. *Journal of Machine Learning Research*, 16:2797–2835, 2015.

Jagannath, A., Lopatto, P., and Miolane, L. Statistical thresholds for tensor PCA. *The Annals of Applied Probability*, 30(4):1910–1933, 2020.

Janzamin, M., Ge, R., Kossaifi, J., and Anandkumar, A. Spectral learning on matrices and tensors. *Foundations and Trends in Machine Learning*, 12(5-6):393–536, 2019.

Kakade, S. M. and Foster, D. P. Multi-view regression via canonical correlation analysis. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings 20*, pp. 82–96. Springer, 2007.

Kargas, N. and Sidiropoulos, N. D. Learning mixtures of smooth product distributions: Identifiability and algorithm. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 388–396, Naha, Okinawa, Japan, Apr 2019.

Khouja, R., Mattei, P. A., and Mourrain, B. Tensor decomposition for learning Gaussian mixtures from moments. *Journal of Symbolic Computation*, 113:193–210, 2022.

Kilmer, M. E., Braman, K., Hao, N., and Hoover, R. C. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications*, 34(1):148–172, 2013.

Kofidis, E. and Regalia, P. A. On the best rank-1 approximation of higher-order supersymmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(3):863–884, 2002.

Lesieur, T., Miolane, L., Lelarge, M., Krzakala, F., and Zdeborová, L. Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 511–515. IEEE, 2017.

Lim, L.-H. Singular values and eigenvalues of tensors: a variational approach. In *Proc. IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 129–132, 2005.

Liu, X., Ji, S., Glänzel, W., and De Moor, B. Multiview partitioning via tensor methods. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):1056–1069, 2013. doi: 10.1109/TKDE.2012.95.

Nigam, K. and Ghani, R. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pp. 86–93, 2000.

Oseledets, I. V. and Kharyuk, P. V. Structuring data with block term decomposition: Decomposition of joint tensors and variational block term decomposition as a parametrized mixture distribution model. *Computational Mathematics and Mathematical Physics*, 61(5):816–835, 2021.

Papalexakis, E. E., Sidiropoulos, N. D., and Bro, R. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE Transactions on Signal Processing*, 61 (2):493–506, 2012.

Perry, A., Wein, A. S., and Bandeira, A. S. Statistical limits of spiked tensor models. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pp. 230–264. Institut Henri Poincaré, 2020.

Rahmani, D., Niranjan, M., Fay, D., Takeda, A., and Brodzki, J. Estimation of Gaussian mixture models via tensor moments with application to online learning. *Pattern Recognition Letters*, 131: 285–292, 2020.

Richard, E. and Montanari, A. A statistical model for tensor PCA. *Advances in neural information processing systems*, 27, 2014.

Seddik, M. E. A., Guillaud, M., and Couillet, R. When random tensors meet random matrices. *arXiv preprint arXiv:2112.12348*, 2021.

Singhal, P., Mirza, W., Rajwade, A., and Gurumoorthy, K. S. Estimating joint probability distribution with low-rank tensor decomposition, Radon transforms and dictionaries. *arXiv:2304.08740*, 2023.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Sun, S. A survey of multi-view machine learning. *Neural computing and applications*, 23:2031–2038, 2013.

Tao, T. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

White, M., Zhang, X., Schuurmans, D., and Yu, Y.-l. Convex multi-view subspace learning. *Advances in neural information processing systems*, 25, 2012.

Widder, D. V. The stieltjes transform. *Transactions of the American Mathematical Society*, 43(1): 7–60, 1938.

Wu, J., Xie, X., Nie, L., Lin, Z., and Zha, H. Unified graph and low-rank tensor learning for multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6388–6395, 2020.

Xie, Y., Tao, D., Zhang, W., Liu, Y., Zhang, L., and Qu, Y. On unifying multi-view self-representations for clustering by tensor multi-rank minimization. *International Journal of Computer Vision*, 126: 1157–1179, 2018.

Xu, C., Tao, D., and Xu, C. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Zhao, J., Xie, X., Xu, X., and Sun, S. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.