Hotelling Deflation on Large Symmetric Spiked Tensors

Mohamed El Amine SEDDIK[1] José Henrique de Morais GOULART[2] Maxime GUILLAUD[3] Alexis DECURNINGE[4]

[1]Technology Innovation Institue, PO Box: 9639, Masdar City, Abu Dhabi, UAE

[2]IRIT, Toulouse INP, CNRS, 2 rue Charles Camichel, 31071 Toulouse, France

[3]Inria / CITI Laboratory, 6 avenue des Arts, 69621 Villeurbanne, France

[4]Huawei Technologies, 20 quai du Point du Jour, 92100 Boulogne-Billancourt, France

Résumé – Cet article étudie l'algorithme de déflation appliqué à l'estimation d'un modèle de rang faible symétrique contenu dans un tenseur de grandes dimensions corrompu par un bruit additif gaussien. Plus précisément, nous fournissons une caractérisation précise de la performance en grandes dimensions de la déflation en termes des alignements des vecteurs obtenus par approximations successives de rang 1 et de leurs poids, en supposant des corrélations (fixes) non-triviales entre les composantes du modèle. Notre analyse permet de comprendre le mécanisme de déflation en présence de bruit et peut être exploitée pour concevoir des méthodes d'estimation plus efficaces.

Abstract – This paper studies the deflation algorithm when applied to estimate a low-rank symmetric spike contained in a large tensor corrupted by additive Gaussian noise. Specifically, we provide a precise characterization of the large-dimensional performance of deflation in terms of the alignments of the vectors obtained by successive rank-1 approximation and of their estimated weights, assuming non-trivial (fixed) correlations among spike components. Our analysis allows an understanding of the deflation mechanism in the presence of noise and can be exploited for designing more efficient signal estimation methods.

1 Introduction

By capitalizing on the uniqueness properties of certain tensor decompositions, one can address many parameter estimation or information retrieval problems in signal processing, data sciences, and machine learning by recasting them as the decomposition of some data tensor built from the observations [3]. An archetypal application in signal processing is source separation, which can be formulated as the decomposition of a tensor containing either data acquired by a sensor array or high-order statistics estimated from these data [6].

Among these (essentially) unique tensor decompositions, the canonical polyadic decomposition (CPD) [9] figures prominently. It consists in writing a tensor as a (minimal) sum of rank-one terms, and as such can be seen as one possible extension of the singular value decomposition. Yet, these rank-one terms need not be orthogonal for their uniqueness, and this fact is at the heart of its popularity. In several problems, the tensor of interest is symmetric and the sought information is encoded in a rank-r symmetric CPD, that is,

$$\mathbf{X} = \sum_{i=1}^{r} \beta_i \, \boldsymbol{x}_i^{\otimes d}, \tag{1}$$

where $\beta_i \in \mathbb{R}$, $\|\boldsymbol{x}_i\| = 1$ and $\boldsymbol{x}^{\otimes d}$ denotes the tensor product of vector $\boldsymbol{x} \in \mathbb{R}^n$ with itself $d-1$ times (for instance, $\boldsymbol{x}^{\otimes 3} = \boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}$). One example is in latent variable model learning [3], where the vectors $\boldsymbol{x}_i$ appearing in the decomposition are directly related to the model's parameters (for instance, each $\boldsymbol{x}_i$ is the mean of a Gaussian component in a mixture model).

If the vectors $\boldsymbol{x}_i$ in (1) were orthogonal, then one could retrieve them by resorting to a greedy deflation procedure, first introduced by Hotelling for matrices [10]: a best rank-1 approximation of $\mathbf{X}$ is computed and then subtracted from $\mathbf{X}$, and the process is repeated r times. Algorithmically, each such approximation can be computed by power iteration [12]. However, this is no longer true when these vectors are not orthogonal—in fact, subtracting a best rank-1 approximation from $\mathbf{X}$ can even yield a tensor of *higher* rank [16]. In some applications, this can in principle be circumvented by transforming $\mathbf{X}$ in such a way that it becomes a rank-r symmetric orthogonal decomposition, as long as $r \leq n$. For instance, in latent variable model learning the eigendecomposition of a matrix of second-order statistics can be exploited to obtain a whitening matrix $\boldsymbol{W} \in \mathbb{R}^{r \times n}$ such that the vectors $\tilde{\boldsymbol{x}}_i = \boldsymbol{W}\boldsymbol{x}_i \in \mathbb{R}^r$, $i = 1, \ldots, r$, are pairwise orthogonal. An analysis of an algorithm employing this technique coupled with tensor power iteration was carried out in [3], including a robust estimation result quantifying the performance in the case one observes $\mathbf{Y} = \bar{\mathbf{X}} + \mathbf{E}$, in terms of the spectral norm of the perturbation $\mathbf{E}$, where $\bar{\mathbf{X}}$ admits a symmetric orthogonal decomposition. Performance bounds were also derived in [1, 2] for an algorithm involving tensor power iteration in the overcomplete regime with $r > n$, where whitening is no longer possible and thus one has to impose additional constraints to control deviation from orthogonality (such as small coherence or uniform sampling from the unit sphere).

Yet, the results of these previous works are not well suited to the *large-dimensional* regime. Specifically, the vectors $\boldsymbol{x}_i$ become nearly orthogonal under the assumptions made in [1, 2] as $n \to \infty$, which can be quite restrictive, while the bound on the spectral norm of the perturbation imposed on [3] may not hold when the data dimension n is of the same order of the number of samples used to estimate the required statistics, which is a typical assumption in this regime.

Our goal here is to study the performance of a deflation

J. H. de M. Goulart's work was supported by the ANR LabEx CIMI (ANR-11-LABX-0040) within the French Programme "Investissements d'Avenir.".

procedure in the case where r is fixed (for simplicity, $r = 2$), the observations are corrupted by noise, $n \to \infty$ but the alignments $|\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle| \neq 0$ are fixed and not $o(1)$. Such an analysis is a first step towards devising more sophisticated algorithms such as orthogonalized deflation, as has been recently done in the asymmetric case [15]. To this end, we build upon a recently developed approach [8, 14] which allows studying random tensor models by deploying tools from random matrix theory. The core idea of this approach is to study partial contractions, which give rise to large random matrices.

More concretely, we study the (random) alignments between the vectors obtained by deflation and the components $\boldsymbol{x}_i$ of our CPD model, in the regime of asymptotically large tensor dimensions. Under the assumption that these alignments (and the estimates of the weights β_i) concentrate and some additional technical conditions, we derive a system of equations that are satisfied by the limiting values of these quantities. Once the deflation procedure is applied, one can plug its output into the equations and numerically solve for the other unknown quantities, including the weights β_i and the alignments $\langle \boldsymbol{x}_i, \boldsymbol{u}_j \rangle$, where the vectors $\boldsymbol{u}_j$ are the estimated components obtained by deflation. This allows evaluating the quality of the obtained decomposition in terms of the alignments $\langle \boldsymbol{x}_i, \boldsymbol{u}_j \rangle$, which is useful in practice. Moreover, our technical results give insights on the properties of the standard deflation algorithm, thus being a first step towards proposing better algorithms based on rank-one approximations. Finally, our numerical results for finite dimensions show that the obtained values closely match the predictions given by the derived equations.

2 Spiked tensor and deflation

We consider the following rank-r and order-d symmetric spiked random tensor

$$\mathbf{S} \equiv \sum_{i=1}^{r} \beta_i \boldsymbol{x}_i^{\otimes d} + \tfrac{1}{\sqrt{n}} \mathbf{W}, \tag{2}$$

with $\boldsymbol{x}_i$ on the unit sphere $\mathbb{S}^{n-1}$ and $\mathbf{W}$ a dth-order symmetric Gaussian tensor (see [8] for a formal definition). The signal part is modeled by the rank-r component with weights $\beta_i > 0$, which collectively determine the signal-to-noise ratio [1] of the model. We further assume that the rank-one components are non-orthogonal and we denote

$$\alpha_{ij} \equiv \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \neq 0 \quad \text{for all } i \neq j. \tag{3}$$

In the following, we will study a deflation approach aimed at approximately recovering the low-rank signal tensor, which consists in performing successive rank-one approximations and subtracting the result at each iteration. Specifically, at iteration $i \in [r]$, we compute the best rank-one approximation of $\mathbf{S}_i$, denoted $\hat{\lambda}_i \boldsymbol{u}_i^{\otimes d}$, and subtract it from $\mathbf{S}_i$. Starting with $\mathbf{S}_0 = \mathbf{S}$, this yields the sequence of tensors

$$\mathbf{S}_i = \mathbf{S}_{i-1} - \hat{\lambda}_{i-1} \boldsymbol{u}_{i-1}^{\otimes d}, \tag{4}$$

where $\hat{\lambda}_0 = 0$ by convention, and

$$\boldsymbol{u}_i \equiv \underset{\|\boldsymbol{u}\|=1}{\arg\max}\ \mathbf{S}_i \cdot \boldsymbol{u}^d, \tag{5}$$

$$\hat{\lambda}_i \equiv \mathbf{S}_i \cdot \boldsymbol{u}_i^d, \tag{6}$$

1. Here we assume for simplicity that $\beta_i > 0$ for all $i \in [r]$, which implies no loss of generality for odd d. The case with arbitrary signs can be treated similarly, at the expense of more cumbersome derivations.

where $\mathbf{S} \cdot \boldsymbol{u}^m$ denotes m-fold contraction of the tensor $\mathbf{S}$ with the vector $\boldsymbol{u}$. It follows that each $\mathbf{S}_i$ is also a low-rank spiked random tensor given by

$$\mathbf{S}_i = \textstyle\sum_{j=1}^{r} \beta_j \boldsymbol{x}_j^{\otimes d} - \sum_{j=1}^{i-1} \hat{\lambda}_j \boldsymbol{u}_j^{\otimes d} + \frac{1}{\sqrt{n}} \mathbf{W}. \tag{7}$$

Note that the solution to the best rank-one tensor approximation problem (5) is in general *not* a component of the CPD of $\mathbf{S}$. This is due to the fact that the Eckhart-Young theorem is not applicable in the non-orthogonally decomposable setting [7]. Thus, there is admittedly a mismatch between the objective of estimating the components of the CPD and the strategy of computing successive rank-one approximations. Nonetheless, the deflation approach is algorithmically simple and easier to analyze than joint optimization schemes (as it relies on rank-1 approximation), and can also provide acceptable approximate solutions when cross-component correlations are small. In the sequel, we give analytical tools to characterize and improve the accuracy achieved by Hotelling-type tensor deflation.

To understand the performance of this procedure in the large-dimensional regime, our main task consists in estimating the following quantities, which we refer to as *summary statistics* as introduced by [4], when $n \to \infty$, as functions of the parameters α_{ij} and β_i:

$$\hat{\lambda}_i, \quad \hat{\rho}_{ij} \equiv \langle \boldsymbol{u}_i, \boldsymbol{x}_j \rangle, \quad \hat{\eta}_{ij} \equiv \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle \quad \text{for } i, j \in [r]. \tag{8}$$

We will see in the sequel how this problem can be addressed through the analysis of certain random matrices, built from contractions of the tensors $\mathbf{S}_i$.

3 Main results

3.1 Associated random matrices

For $r = 1$, the problem (5) is tantamount to the maximum likelihood estimation (MLE) of $\boldsymbol{x}_1$. In this setting, [8] introduced an approach for studying the performance of MLE by borrowing tools from random matrix theory. This approach is based upon two crucial observations: (i) critical points $\boldsymbol{u}$ of (5) are eigenvectors of $\mathbf{S}_i$ satisfying [13]

$$\mathbf{S}_i \cdot \boldsymbol{u}^{d-1} = \lambda \boldsymbol{u}, \tag{9}$$

with the eigenvalue λ given by $\mathbf{S}_i \cdot \boldsymbol{u}^d$; (ii) every eigenpair $(\lambda, \boldsymbol{u})$ of the tensor $\mathbf{S}_i$ is also an eigenpair of the matrix resulting from the contraction $\mathbf{S}_i \cdot \boldsymbol{u}^{d-2}$, since by (9)

$$\left(\mathbf{S}_i \cdot \boldsymbol{u}^{d-2}\right) \boldsymbol{u} = \mathbf{S}_i \cdot \boldsymbol{u}^{d-1} = \lambda \boldsymbol{u}. \tag{10}$$

Hence the solution $\boldsymbol{u}_i$ of (5) is an eigenvector (in fact, the dominant eigenvector [8]) of the matrix $\mathbf{S}_i \cdot \boldsymbol{u}_i^{d-2}$. The matrix eigenproblem (10) does not provide a constructive way to solve for $\boldsymbol{u}$ since $\mathbf{S}_i \cdot \boldsymbol{u}^{d-2}$ itself depends on $\boldsymbol{u}$; however, its analysis through random matrix theory allows to characterize the properties of the solutions. Specifically, by analyzing contractions of this form, combined with the tensor eigenvalue equation (9), [8] derived an asymptotic expression for the performance of MLE in terms of the alignment of $\boldsymbol{u}_1$ and $\boldsymbol{x}_1$ in the regime where estimation is possible (that is, beyond the phase transition characterized by [11]).

Here, assuming now that r is a fixed integer such that $r > 1$, we carry out a similar study of the random tensor models $\mathbf{S}_i$ through the analysis of the contractions $\mathbf{S}_i \cdot \boldsymbol{u}_i^{d-2}$.

$$\begin{cases} \lambda_i + \dfrac{1}{d-1} g\left(\dfrac{\lambda_i}{d-1}\right) = \displaystyle\sum_{j=1}^{r} \beta_j \rho_{ij}^d - \sum_{j=1}^{i-1} \lambda_j \eta_{ij}^d, \quad h(\lambda_i)\rho_{ij} = \sum_{k=1}^{r} \beta_k \alpha_{jk} \rho_{ik}^{d-1} - \sum_{k=1}^{i-1} \lambda_k \rho_{kj} \eta_{ik}^{d-1}, \quad (i,j) \in [r]^2, \\ \left[h(\lambda_i) + q(\lambda_j)\eta_{ij}^{d-2}\right] \eta_{ij} = \displaystyle\sum_{k=1}^{r} \beta_k \rho_{jk} \rho_{ik}^{d-1} - \sum_{k=1}^{i-1} \lambda_k \eta_{kj} \eta_{ik}^{d-1}, \quad i \in [r],\ j < i. \end{cases} \tag{S}$$

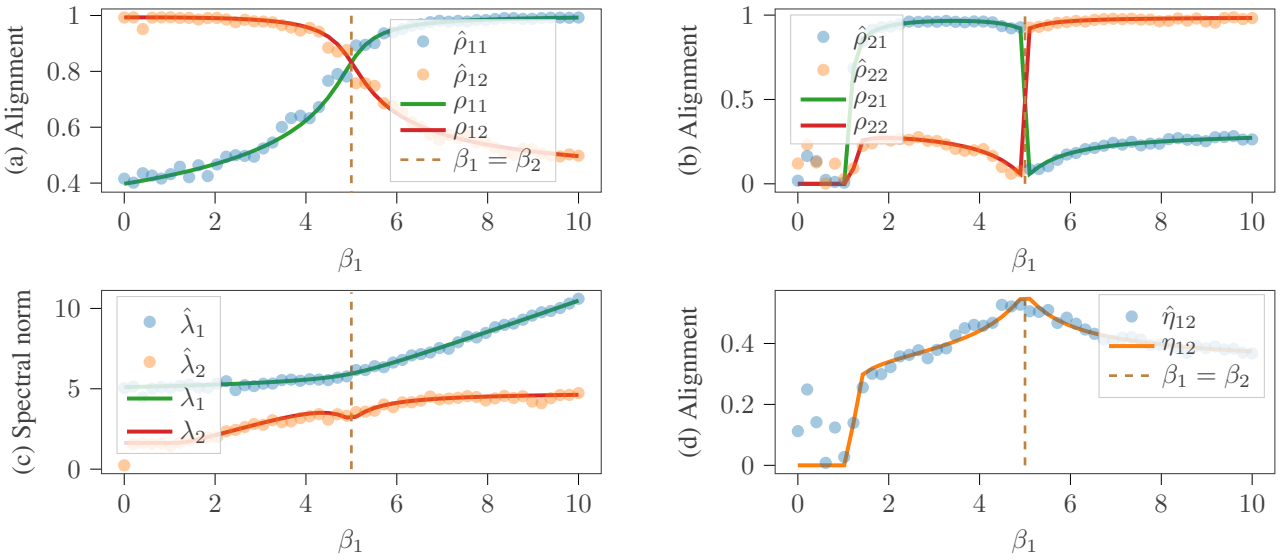

Figure 1 – Empirical (dots, Monte-Carlo simulations) versus asymptotic (lines, as per Corollary 3.4) summary statistics of the symmetric Hotelling deflation, for parameters $r = 2$, $d = 3$, $n = 100$ and $\alpha = 0.4$, for a range of β_1 and a fixed $\beta_2 = 5$. (a) First deflation step: alignments of $\boldsymbol{u}_1$ with $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. (b) Second deflation step: alignments of $\boldsymbol{u}_2$ with $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. (c) Eigenvalues $\hat{\lambda}_i$ and their limits λ_i resp. (d) Alignment η_{12} between the eigenvectors estimated at the first and second deflation step. The asymptotic curves are obtained by solving numerically $\Psi(\cdot, \boldsymbol{\beta}, \cdot) = \mathbf{0}$ in Corollary 3.4 initialized with the simulated summary statistics for one realization of the noise tensor $\mathbf{W}$.

3.2 Limiting spectrum

Our first result characterizes the limiting spectral measure of the contractions $\mathbf{S}_i \cdot \boldsymbol{u}_i^{d-2}$, and is instrumental in proving our main result, which is an asymptotic characterization of the summary statistics in (8).

Theorem 3.1. *The empirical spectral measures of* $\mathbf{S}_i \cdot \boldsymbol{u}_i^{d-2}$ *and of* $\frac{1}{\sqrt{n}}\mathbf{W} \cdot \boldsymbol{u}_i^{d-2}$ *converge weakly almost surely to the semi-circle distribution* μ *whose Stieltjes transform is given by*

$$g(z) \equiv \frac{2}{\gamma_d^2}\left(-z + \sqrt{z^2 - \gamma_d^2}\right),$$

and whose density reads $\mu(dx) = \frac{2}{\pi\gamma_d^2}\sqrt{\gamma_d^2 - x^2}\,dx$ *and is supported on* $[-\gamma_d, \gamma_d]$.

Proof sketch: The proof starts by noticing that the random matrix $\mathbf{S}_i \cdot \boldsymbol{u}_i^{d-2}$ can be written as $\boldsymbol{L} + \frac{1}{\sqrt{n}}\mathbf{W} \cdot \boldsymbol{u}_i^{d-2}$ where $\boldsymbol{L}$ is a low-rank matrix. Therefore, involving classical random matrix arguments, the matrices $\mathbf{S}_i \cdot \boldsymbol{u}_i^{d-2}$ and $\frac{1}{\sqrt{n}}\mathbf{W} \cdot \boldsymbol{u}_i^{d-2}$ share the same limiting spectrum, and the former is characterized similarly to the rank-one case from [8].

3.3 Limiting summary statistics

We now look into the asymptotic values of the summary statistics introduced in (8). To derive them, we start from the tensor eigenvalue equations relating the pairs $(\hat{\lambda}_i, \boldsymbol{u}_i)$ and the tensors $\mathbf{S}_i$, that is

$$\hat{\lambda}_i \boldsymbol{u}_i = \mathbf{S}_i \cdot \boldsymbol{u}_i^{d-1} = \sum_{j=1}^{r} \beta_j \langle \boldsymbol{u}_i, \boldsymbol{x}_j \rangle^{d-1} \boldsymbol{x}_j - \sum_{j=1}^{i-1} \hat{\lambda}_j \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^{d-1} \boldsymbol{u}_j + \frac{1}{\sqrt{n}}\mathbf{W} \cdot \boldsymbol{u}_i^{d-1}, \tag{11}$$

where we used (7). Then, we can have access to $\hat{\lambda}_i$ by taking the scalar product of both sides with $\boldsymbol{u}_i$, since this vector has a unit norm. Similarly, $\hat{\rho}_{ij}$ and $\hat{\eta}_{ij}$ are obtained by taking scalar products with $\boldsymbol{x}_j$ and $\boldsymbol{u}_j$, respectively. Next, one can compute the expectations of these quantities by invoking Stein's lemma (a.k.a. Gaussian integration by parts) to handle the dependence between $\mathbf{W}$ and each $\boldsymbol{u}_i$, and take the limit $n \to \infty$. Finally, similarly to [8, 14, 15] we assume that these random quantities concentrate around their expectations, and impose some technical conditions on their limiting values, as follows.

Assumption 3.2 (Almost sure convergence)**.** We suppose that for each tensor $\mathbf{S}_i$ involved in the deflation there exists a sequence of eigenpairs $\{(\hat{\lambda}_i, \boldsymbol{u}_i)\}_{n\in\mathbb{N}}$ of $\mathbf{S}_i$ such that

$$\hat{\lambda}_i \xrightarrow[n\to\infty]{\text{a.s.}} \lambda_i, \quad \hat{\rho}_{ij} \xrightarrow[n\to\infty]{\text{a.s.}} \rho_{ij}, \quad \hat{\eta}_{ij} \xrightarrow[n\to\infty]{\text{a.s.}} \eta_{ij},$$

with $\lambda_i > \gamma_d(d-1)$, $\rho_{ij} \neq 0$ and $\eta_{ij} \neq 0$ where $\gamma_d = 2/\sqrt{d(d-1)}$.

Under these assumptions, we can derive a system of equations characterizing the summary statistics in the limit $n \to \infty$. As our numerical results will show, the solutions to these equations match the empirical observations for n large enough.

Theorem 3.3. *Suppose that Assumption 3.2 holds, then the limiting summary statistics λ_i, ρ_{ij} and η_{ij} satisfy the system of equations shown in (S) on the preceding page with $i, j \in [r]$, where $h(z) \equiv z + g\left(z/(d-1)\right)/d$ and $q(z) \equiv g(z/(d-1))/(d(d-1))$.*

Note that Theorem 3.3 only states that if the summary statistics converge to their respective limits, then the latter are solutions to the system of equations in (S); the converse is not necessarily true. In fact, studying the existence and uniqueness of the solutions of (S) is still an open question.

As we will see in the next section, when the system (S) is solved numerically with proper initialization (e.g., with the empirical summary statistics), the obtained solutions describe well the asymptotic behavior of the *maximizers* of (5), despite the fact that the tensor eigenvalue equations characterize *all critical points* of these problems. We refer the reader to [8, 14] for a discussion on similar phenomena observed in the rank-one case, whose rigorous explanation remains open.

Corollary 3.4. *Suppose $r = 2$, $d = 3$ and denote $\boldsymbol{\lambda} \equiv (\lambda_1, \lambda_2, \eta_{12})$, $\boldsymbol{\beta} \equiv (\beta_1, \beta_2, \alpha_{12})$ and $\boldsymbol{\rho} \equiv (\rho_{ij})_{i,j\in[2]}$. Then, the limiting summary statistics $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ satisfy $\Psi(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\rho}) = \mathbf{0}$, where the mapping $\Psi : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^4 \to \mathbb{R}^7$ is defined by*

$$\Psi\begin{pmatrix}\boldsymbol{\lambda}\\ \boldsymbol{\beta}\\ \boldsymbol{\rho}\end{pmatrix} \equiv \begin{pmatrix} \sum_{i=1}^2 \beta_i \rho_{1i}^3 - f(\lambda_1) \\ \sum_{i=1}^2 \beta_i \alpha_{1i} \rho_{1i}^2 - h(\lambda_1)\rho_{11} \\ \sum_{i=1}^2 \beta_i \alpha_{2i} \rho_{1i}^2 - h(\lambda_1)\rho_{12} \\ \sum_{i=1}^2 \beta_i \rho_{2i}^3 - f(\lambda_2) - \lambda_1 \eta_{12}^3 \\ \sum_{i=1}^2 \beta_i \alpha_{1i} \rho_{2i}^2 - h(\lambda_2)\rho_{21} - \lambda_1 \rho_{11}\eta_{12}^2 \\ \sum_{i=1}^2 \beta_i \alpha_{2i} \rho_{2i}^2 - h(\lambda_2)\rho_{22} - \lambda_1 \rho_{12}\eta_{12}^2 \\ \sum_{i=1}^2 \beta_i \rho_{1i}\rho_{2i}^2 - h(\lambda_2)\eta_{12} - [\lambda_1 + q(\lambda_1)]\eta_{12}^2 \end{pmatrix}.$$

4 Discussion

Fig. 1 illustrates the accuracy of using the deflation approach of (4)–(6) to estimate the spike components $\boldsymbol{x}_i$ and weights β_i in the correlated case $\alpha = 0.4$. As the result depends critically on the relative values of β_1 and β_2, we let β_1 vary for a fixed $\beta_2 = 5$. In Figs. 1(a–b), as expected from the deflation procedure, when $\beta_1 < \beta_2$, $\boldsymbol{u}_1$ tends to correlate with $\boldsymbol{x}_2$, the strongest component, hence ρ_{12} is high; conversely, for $\beta_1 > \beta_2$, $\boldsymbol{u}_1$ tends to correlate with $\boldsymbol{x}_1$ and ρ_{11} is high. Naturally, ρ_{21} and ρ_{22} behave symmetrically. Interestingly, in the regime $\beta_1 \approx \beta_2$, $\boldsymbol{u}_1$ aligns fully neither with $\boldsymbol{x}_1$ nor with $\boldsymbol{x}_2$. This indicates a significant weakness in the deflation approach with non-orthogonally decomposable tensors when several components have comparable power, since improperly estimating and subtracting the first component has the detrimental effect of *increasing* the rank of the non-noise component in $\mathbf{S}_1$ with respect to $\mathbf{S}_0$ (see eq. (7)). We also note that, during the second deflation step, the estimator fails to achieve positive correlation of $\boldsymbol{u}_2$ with either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ for very low values of β_1. Fig. 1(c) shows that λ_1 fairly accurately tracks the power of the strongest component (equal to $\max(\beta_1, 5)$), while λ_2 is affected by a noise floor at the low range of β_1 and constitutes a poor estimator of the power of the weakest component (equal to $\min(\beta_1, 5)$).

As is common with random matrix theory, the asymptotic results from Theorem 3.3 hold approximately with remarkable accuracy for finite dimension problems thanks to the concentration of measure phenomenon[5].

References

[1] A. Anandkumar, R. Ge, and M. Janzamin. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory*, pages 36–112. PMLR, 2015.

[2] A. Anandkumar, R. Ge, and M. Janzamin. Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research*, 18(22):1–40, 2017.

[3] A. Anandkumar, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[4] G. Ben Arous, R. Gheissari, and A. Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25349–25362, 2022.

[5] F. Benaych-Georges, A. Guionnet, and M. Maida. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electronic Journal of Probability*, 16:1621–1662, 2011.

[6] J.-F. Cardoso. Super-symmetric decomposition of the fourth-order cumulant tensor. blind identification of more sources than sensors. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 91, pages 3109–3112, 1991.

[7] J. Draisma, G. Ottaviani, and A. Tocino. Best rank-k approximations for tensors: generalizing Eckart–Young. *Research in the Mathematical Sciences*, 5(2):27, 2018.

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

[9] F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[10] H. Hotelling. Analysis of a complex of statistical variables into principal components. volume 24, pages 417–441. Warwick & York, September 1933.

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

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

[13] L.-H. Lim. Singular values and eigenvalues of tensors: a variational approach. In *CAMSAP*, pages 129–132, Puerto Vallarta, Mexico, December 2005.

[14] M. El A. Seddik, M. Guillaud, and R. Couillet. When random tensors meet random matrices. *Annals of Applied Probability*, 2023. (accepted).

[15] M. El A. Seddik, M. Mahfoud, and M. Debbah. Optimizing orthogonalized tensor deflation via random tensor theory. *preprint arXiv:2302.05798*, 2023.

[16] A. Stegeman and P. Comon. Subtracting a best rank-1 approximation may increase tensor rank. *Linear Algebra and its Applications*, 433(7):1276–1300, 2010.