Tensor Decomposition via Joint Matrix Schur Decomposition

Nicolò Colombo NICOLO.COLOMBO@UNI.LU
Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur Alzette, Luxembourg

Nikos Vlassis VLASSIS@ADOBE.COM
Adobe Research, San Jose, CA

Abstract

We describe an approach to tensor decomposition that involves extracting a set of observable matrices from the tensor and applying an approximate joint Schur decomposition on those matrices, and we establish the corresponding first-order perturbation bounds. We develop a novel iterative Gauss-Newton algorithm for joint matrix Schur decomposition, which minimizes a nonconvex objective over the manifold of orthogonal matrices, and which is guaranteed to converge to a global optimum under certain conditions. We empirically demonstrate that our algorithm is faster and at least as accurate and robust than state-of-the-art algorithms for this problem.

1. Introduction

Tensors are ubiquitous in the mathematical sciences, with important applications in physics, signal processing, and machine learning (Comon, 2014). The fundamental problem of *tensor decomposition* amounts to writing a given tensor as a sum of simpler tensors that can carry useful interpretation. For instance, the classical CP decomposition of an order-3 tensor reads

$$T = \sum_{r=1}^{R} w_r \, x_r \circ y_r \circ z_r, \tag{1}$$

where w_r are scalar weights and the vectors x_r, y_r, z_r are the factors of the decomposition ($\circ$ denotes outer product). The smallest R for which (1) holds is the *rank* of the tensor. Tensors enjoy widespread use because of their capacity to model higher-order statistics but also their identifiability properties: Contrary to the matrix case, tensor decomposition is unique under mild conditions, for instance when R in (1) is not too large (Kruskal, 1977; Sidiropoulos & Bro, 2000; Comon, 2014).

In machine learning we are typically dealing with tensors that are noisy versions of (1), corresponding to empirical high-order moments (Anandkumar et al., 2014b). In that case a problem of interest is *approximate* tensor decomposition under a low-rank constraint: Given an ‘empirical’ tensor $\hat{T}$, and a desired rank R, find a CP-decomposable ‘ground-truth’ tensor T that approximates $\hat{T}$ in a certain sense. Our approach belongs to the class of semi-algebraic methods, in which an estimation of T is obtained by exploiting a series of algebraic conditions that are satisfied exactly by T but only hold approximately for $\hat{T}$. Specifically, we build on a line of research that reduces tensor decomposition to a (joint) matrix diagonalization problem (Harshman, 1970; Leurgans et al., 1993; Chang, 1996; Mossel & Roch, 2006; De Lathauwer, 2006; Anandkumar et al., 2014b; Montanari & Richard, 2014; Vempala & Xiao, 2015; Kuleshov et al., 2015). The underlying idea of this ‘matricization’ approach to tensor decomposition is to extract a set of ‘observable’ matrices from the tensor, and then estimate the tensor factors through the eigenstructure of this set. The observable matrices are typically obtained by random contractions of the tensor slices, followed by an approximate joint matrix diagonalization step. For the latter, first-order perturbation bounds are available (Cardoso, 1994; Afsari, 2008; Kuleshov et al., 2015), which reveal a much milder dependence on the eigengaps that plagued earlier analyses (Mossel & Roch, 2006).

We depart from the above literature in three key aspects: (1) We describe a way to extract observable matrices from the tensor that does not involve random contractions. (2) We propose an approach for diagonalizing the set of observable matrices via an *approximate joint Schur decomposition*, and we establish a first-order perturbation bound drawing on the proof technique of Cardoso (1994). (3) We describe a novel Gauss-Newton algorithm for approximate joint Schur decomposition that minimizes a nonconvex objective on the manifold of *orthogonal* matrices. The fact

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

that optimization is over a 'nice' manifold overcomes usual problems of methods involving nonorthogonal joint matrix decomposition (Afsari, 2008). Our matrix-manifold algorithm is at least one order of magnitude faster than state-of-the-art Jacobi algorithms (Haardt & Nossek, 1998; Abed-Meraim & Hua, 1998; De Lathauwer, 2006), and it is guaranteed to converge to a globally optimal solution under certain conditions.

Notation. Given a matrix A, $\lambda_i(A)$ is its ith eigenvalue, $\mathrm{vec}(A)$ is its column-wise vectorization, mat is defined by $\mathrm{mat}(\mathrm{vec}(A)) = A$, $\mathrm{low}(A)$ is the strictly lower-diagonal part of A defined by $[\mathrm{low}\,(A)]_{ij} = A_{ij}$ if $i > j$ and 0 otherwise, Low is defined by $\mathrm{vec}(\mathrm{low}(A)) = \mathrm{Low}\,\mathrm{vec}(A)$, and $\kappa(A)$ is the condition number of A. The identity matrix is I, the Frobenius norm is $\|A\|$, the tensor norm is defined by $\|T\|^2 = \sum_{ijk} T_{ijk}^2$, and $\otimes$ is the Kronecker product.

2. Tensor decomposition as a joint matrix diagonalization problem

Consider an order-3 tensor of the form[1]

$$\hat{T} = T + \sigma E, \qquad T_{ijk} = \sum_{r=1}^{R} Z_{ir} Z_{jr} Z_{kr}, \tag{2}$$

where $i, j, k = 1, \dots d$, $\sigma > 0$ and E is an arbitrary (not necessary symmetric) noise term satisfying $\|E\| \leq \varepsilon$. The problem involves recovering the factors Z from the empirical tensor $\hat{T}$. Assume $R = d$, and define the observable matrices $\hat{M}_k$, for $k = 1, \dots d$ as

$$\hat{M}_k = \hat{m}_k (\hat{m})^{-1}, \quad [\hat{m}_k]_{ij} = \hat{T}_{ijk}, \quad \hat{m} = \sum_{k=1}^{d} \hat{m}_k. \tag{3}$$

Note that each matrix $\hat{m}_k$ is just a slice of the empirical tensor $\hat{T}$. It is easy to show the following:

Lemma 1. *If Z is invertible and $[1^T Z]_r \neq 0$ for all $r = 1, \dots, R$, the observable matrices $\hat{M}_k$ defined in* (3) *can be expanded as follows, for $k = 1, \dots, d$*

$$\begin{aligned} \hat{M}_k &= M_k + \sigma W_k + O(\sigma^2) & (4)\\ M_k &= Z \mathrm{diag}(\mathbf{e}_k^T Z) \left(\mathrm{diag}(1^T Z)\right)^{-1} Z^{-1} & (5)\\ W_k &= e_k m^{-1} + m_k m^{-1} e m^{-1} & (6) \end{aligned}$$

where $\mathbf{e}_k$ is the k-basis vector and the ground-truth matrices are defined by $[e_k]_{ij} = E_{ijk}$, $e = \sum_k e_k$, $[m_k]_{ij} = T_{ijk}$ and $m = \sum_k m_k$.

The lemma implies that, up to a normalization constant, the tensor decomposition problem is equivalent to an (approximate) joint eigenvalues estimation problem. The tensor factors can be estimated from the joint eigenvalue matrices

$$\hat{\lambda}_r(\hat{M}_k) = \frac{Z^*_{kr}}{[1^T Z^*]_r}, \tag{7}$$

since for all $r = 1, \dots, R$ and all $k = 1, \dots, d$, one has $\hat{\lambda}_r(\hat{M}_k) = \frac{Z_{kr}}{[1^T Z]_r} + O(\sigma)$.

2.1. The joint Schur decomposition approach

We propose a way to estimate the joint eigenvalues $\hat{\lambda}_r(\hat{M}_k)$ of the observable matrices in (3) via an approximate joint Schur decomposition. The Schur decomposition of a single matrix A involves finding an orthogonal matrix U such that $U^T A U$ is upper triangular, i.e., $\mathrm{low}(U^T A U) = 0$. Such a decomposition is always possible (Horn & Johnson, 2012). A key property of the Schur decomposition is that the eigenvalues of A appear on the diagonal of the triangularized matrix $U^T A U$, i.e., $\lambda_i(A) = [U^T A U]_{ii}$. When $\sigma = 0$, and under certain conditions on their eigenvalues (Colombo & Vlassis, 2016), the matrices $\hat{M}_k$ admit a finite number of exact joint triangularizers and the estimation of the joint eigenvalues is trivial. For $\sigma > 0$, an exact joint triangularizer of the matrices $\hat{M}_k$ may not exist, but an approximate joint triangularizer can be computed by the following optimization problem over the manifold of orthogonal matrices $\mathrm{O}(d)$

$$\min_{U \in \mathrm{O}(d)} \quad \mathscr{L}(U, \mathscr{M}_\sigma) = \sum_{k=1}^{d} \|\mathrm{low}(U^T \hat{M}_k U)\|^2, \tag{8}$$

with $\mathscr{M}_\sigma = \{\hat{M}_k\}_{k=1}^d$. This problem has always a globally optimal solution, due to the compactness of $\mathrm{O}(d)$.

Let U_* be a (local) minimizer of (8). Due to the continuity of (8) in σ, the approximate triangularizer U_* is expected to be located in a neighbourhood of one of the (exact) triangularizers of the unperturbed matrices M_k. More precisely, there exists an orthogonal matrix $U_\circ$, which is an exact triangularizer of the unperturbed matrices M_k, such that

$$U_* = U_\circ e^{\alpha_* X_*}, \;\; X_* = -X_*^T, \;\; \|X_*\| = 1, \;\; \alpha_* > 0, \tag{9}$$

where the parameter α can be interpreted as the 'distance' between U_* and the ground-truth triangularizer $U_\circ$. Then the estimation of Z is obtained from the approximate joint triangularizer U_* via (7) and the following relation

$$\hat{\lambda}_r(\hat{M}_k) = [U_*^T \hat{M}_k U_*]_{rr} = \lambda_r(M_k) + O(\alpha_* + \sigma) \tag{10}$$

for all $k = 1, \dots, d$ and $r = 1, \dots, R$. Intuitively, the estimation error is due to the noise σ and the perturbation parameter α_* that measures the distance between U_* and $U_\circ$. In practice, the tensor decomposition problem (2) is reduced to (8), which is a matrix optimization problem on the manifold of orthogonal matrices. Under the conditions

[1] Our analysis and algorithm apply also to nonsymmetric T.

that Z is invertible and $[1^T Z]_r \neq 0$ for all $r = 1, \ldots, R$, we can bound the difference between the estimated tensor components (7) and the ground-truth tensor components Z, as we describe next.

2.2. Perturbation theorem

Our main result is the following

Theorem 1. *Let $\hat{T}$ be the tensor defined in* (2) *and assume that $R = d$ and Z is invertible, and $[1^T Z]_r \neq 0$ for all $r = 1, \ldots R$. Then the estimated Z_* from* (7) *and* (10)*, with U_* being a critical point of* (8)*, satisfy*

$$\left|\frac{Z^*_{kr}}{[1^T Z^*]_r} - \frac{Z_{kr}}{[1^T Z]_r}\right| \leq \sigma\left(\frac{4\sqrt{R(R-1)}}{\gamma}\kappa(Z)^4\mu^2 + 1\right)\rho + O(\sigma^2), \quad (11)$$

where $\gamma = \frac{1}{d}\min_{r<r'}\sum_k (Z_{kr} - Z_{kr'})^2$ and

$$\mu = d\kappa(Z)^2\frac{\max|Z|}{\min|1^T Z|}, \quad (12)$$

$$\rho = \varepsilon\frac{\kappa(Z)^2}{\|Z\|^2}\frac{\sqrt{d}}{\min|1^T Z|}\left(1 + d\kappa(Z)^2\frac{\max|Z|}{\min|1^T Z|}\right). \quad (13)$$

Remarks. The proof of Theorem 1 is based on a linear perturbation analysis of the simultaneous triangularizers of a set of nearly commuting matrices. This is a generalization of the analysis carried out by Cardoso (1994) for the simultaneous diagonalization of symmetric nearly joint diagonalizable matrices. In the approach followed here, the matrices M_k in (3) are symmetric only if the tensor (2) has orthogonal factors, i.e., Z is an orthogonal matrix. The orthogonal problem is known to be simpler than the nonorthogonal one addressed here, see, e.g., Kuleshov et al. (2015). The procedure described here is completely general and can be straightforwardly applied also to the orthogonal setting, in which case the error bound (11) will be improved.

We note that Theorem 1 only proves the existence of a matrix Z_* with bounded error. Due to the nonconvexity of (8), finding such a Z_* may not be straightforward. However, since Theorem 1 is obtained from a linear expansion of the stationary condition for (8), it holds for all critical points that admit a bounded first order approximation when expanded around the ground truth solution. More specifically, the bound in Theorem 1 is obtained from an inequality that can be established for the parameter α_* associated to a given critical point of (8). The bound on α_* is given in terms of the ground-truth matrices M_k, the noise parameter σ, and the noise matrices W_k, and it is obtained by expanding the stationarity equation $\nabla\mathscr{L} = 0$, where $\nabla\mathscr{L}$ is the gradient of the objective function (8). In a first order approximation, the stationarity equation can be written as a linear operator $\mathscr{T}$ acting on the projection on the subspace of strictly lower-diagonal matrices of the skew-symmetric matrix X_* defined in (9). Schematically,

$$\alpha_*\mathscr{T}P_{\text{low}}\text{vec}(X_*) = \sigma A + O((\alpha_* + \sigma)^2), \quad (14)$$

where A is a linear function of the ground-truth matrices M_k and the noise matrices W_k, the operator P_{low} is the projector on the subspace of strictly lower-diagonal matrices defined by $P_{\text{low}}^T P_{\text{low}} = \text{Low}$ and $P_{\text{low}}P_{\text{low}}^T = I$, and the operator $\mathscr{T}$ is defined by

$$\mathscr{T} = \sum_{k=1}^{d} t_k^T t_k \quad (15)$$

$$t_k = P_{\text{low}}\big(I \otimes U_\circ^T M_k^T U_\circ - U_\circ^T M_k U_\circ \otimes I\big)P_{\text{low}}^T, \quad (16)$$

where $U_\circ$ is the exact joint triangularizer in (9). Under certain conditions, the operator $\mathscr{T}$ can be shown to be invertible, in which case an inequality on the perturbation parameter α_* can be obtained, up to nonlinear terms, by taking the norm in both sides of $\alpha_* P_{\text{low}}\text{vec}(X_*) = \mathscr{T}^{-1}\sigma A$.

The bound provided by Theorem 1 appears to be less tight that related bounds (Song et al., 2015; Azizzadenesheli et al., 2016) due to the high dependence on the condition number of Z (although our setting is different). In practice, one may consider other types of error estimation techniques such as, for example, *a posteriori* bounds in which the error is computed from a function of the current solution and observed quantities only (Colombo & Vlassis, 2016). These bounds are usually much tighter, and they also overcome the need for global guarantees for the specific algorithm (as those discussed in Section 3).

Finally, a tighter bound can be obtained by considering the spectral norm of the operator $\mathscr{T}$ (we thank a referee for this suggestion). Specifically,

$$\alpha_* \leq \sqrt{2}\|\mathscr{T}^{-1}\|_2 \; \|\sum_{k=1}^{d} t_k^T \text{vec}(U_\circ^T W_k U_\circ)\|, \quad (17)$$

where $\|\cdot\|_2$ denotes spectral norm. The corresponding error in the tensor components estimation is then (Colombo & Vlassis, 2016)

$$\left|\frac{Z^*_{kr}}{[1^T Z^*]_r} - \frac{Z_{kr}}{[1^T Z]_r}\right| \leq 2\alpha_*\mu + \sigma\rho, \quad (18)$$

with μ and ρ given in Theorem 1. Although (18) is expected to be tighter than (11), it lacks an intuitive interpretation in terms of the objects appearing in (2). It would be interesting to compute an upper bound of (17) in which the dependence on the joint triangularizer $U_\circ$ is replaced by an explicit function of the tensor components matrix Z. We leave this for future work.

2.3. Comparison with related results

The generalized Schur decomposition approach of De Lathauwer et al. (2004). The perturbation analysis of the generalized joint Schur decomposition problem used in the method of De Lathauwer et al. (2004) does not provide any bound for the estimation of the tensor components. Despite the state-of-the-art performance of that method, the optimization problem associated to the tensor decomposition is harder than the one in (8), and providing a complete error analysis seems to be a more challenging task. This would require the generalization to the multiple matrices case of the single matrix perturbation analysis for the generalized Schur decomposition given by Sun (1995), but we are not aware of such an extension in the literature.

The matrix-based approach of Kuleshov et al. (2015). Error bounds for the estimation of the tensor components Z have been proposed by Kuleshov et al. (2015). Those perturbation results follow naturally from the perturbation analysis of Cardoso (1994) and Afsari (2008) for the orthogonal and nonorthogonal cases respectively. Our Theorem 1 and the bounds of Kuleshov et al. (2015) share the same general idea of reducing the tensor decomposition problem to a simultaneous matrix decomposition problem, however our result is different from (Kuleshov et al., 2015) in the following three aspects:

(i) The technique of Kuleshov et al. (2015) is based on random projections and the resulting bounds are probabilistic bounds and depend on some failure probability δ. More precisely, the error depends on the logarithm of the inverse of the failure probability, i.e., $\varepsilon \sim \log(\frac{1}{\delta})$. Our approach does not involve any random projections and (11) depends only on the noise and the condition number of Z (see also Section 4.2).

(ii) The bound of Kuleshov et al. (2015) for the nonorthogonal setting includes a parameter describing the 'closeness to orthogonality' of the tensor components, that prescribes how far Z is from an orthogonal matrix. Our approach is designed for the nonorthogonal case, and does not rely on how close to orthogonality the matrix Z is, but only on its conditioning properties.

(iii) Our analysis provides a precise first order expansion, while the bound of Kuleshov et al. (2015) does not specify the form of the term linear in σ, i.e., it reads $\|z_r - z_{r'}\| \leq O(x)\sigma + O(\sigma^2)$ where x is a function of various parameters involved in the perturbation analysis.

The tensor power method of Anandkumar et al. (2014a). Another technique for which precise error bounds are available is the tensor power method of Anandkumar et al. (2014a). In this case, quite involved theoretical bounds are obtained from a convergence analysis of the tensor power iterations and a perturbation bound on the deflation technique that must be applied after the recovery of each tensor component. In practice, comparisons with the tensor decomposition techniques of De Lathauwer et al. (2004) and Kuleshov et al. (2015) show that this method can be less accurate in the recovery of the tensor factors Z. In the nonorthogonal setting, suboptimal performance can often be due to the whitening technique used to approximately transform the nonorthogonal problem into an orthogonal optimization problem (Souloumiac, 2009). Our method can be regarded as a non-approximate alternative to such a whitening technique, where the nonorthogonal decomposition problem, i.e., the simultaneous diagonalization of the nonsymmetric matrices M_k in (3), is mapped into a (slightly more complicated) orthogonal optimization problem, i.e., the simultaneous Schur decomposition (8).

3. A Gauss-Newton algorithm for approximate joint Schur decomposition

In the method proposed here the non-convex optimization problem (8) is solved via an iterative Gauss-Newton algorithm on the manifold of orthogonal matrices. The Gauss-Newton algorithm can be initialized by the triangularizer of a random linear combination of the input matrices in $\mathscr{M}_\sigma$. The computation of each update consists of two steps:

(i) the descent Gauss-Newton direction is computed in the tangent space of $\mathrm{O}(d)$ (extrinsic step)

(ii) a line search is performed on the manifold (intrinsic step) to find the best step size at each iteration.

Compared to the more popular Jacobi approach for joint Schur decomposition, one of the main advantages of Gauss-Newton is its speed. This is mainly due to the fact that no polynomial rooting is required. The second important feature of the Gauss-Newton procedure is its provable convergence to local optima (Absil et al., 2009). Moreover, global guarantees on the obtained solutions can be obtained under certain condition on the initialization (see next).

3.1. The Gauss-Newton algorithm

Let $f : \mathrm{O}(d) \to \mathbf{R}$ be a function of the form $f = \frac{1}{2}\langle g, g\rangle$, where $g : \mathrm{O}(d) \to \mathbf{R}^{d\times d}$ and $\langle A, B\rangle = \mathrm{Tr}(A^T B)$ is the inner product on $\mathbf{R}^{d\times d}$. Its Taylor expansion along the curve $\gamma(t) = Ue^{tX}$, where the tangent space element X obeys $X = -X^T$, is given by

$$f(t) = f(0) + t\dot{f}(0) + \frac{t^2}{2}\ddot{f}(0) + O(t^3) \qquad (19)$$

$$= \langle g(0), g(0)\rangle + t\langle \dot{g}(0), g(0)\rangle + \frac{t^2}{2}\left(\langle \dot{g}(0), \dot{g}(0)\rangle + \langle \ddot{g}(0), g(0)\rangle\right) + O(t^3), \qquad (20)$$

where $f(t) = f(\gamma(t)) = f(Ue^{tX})$. In particular, the difference $f(t) - f(0)$ can be written as an explicit function of the tangent space element X and maximized to find an optimal descent direction. In a gradient descent approach only the first order term $\langle \dot{g}(0), g(0) \rangle$ is considered, while in the Gauss-Newton scheme only the term containing the second order derivative $\ddot{g}(0)$ is neglected. It can be shown that this is equivalent to iteratively minimize a simplified quadratic objective $\tilde{f}(t) = \langle g(0) + t\dot{g}(0), g(0) + t\dot{g}(0) \rangle$. The second term speeds up the convergence in the region where $\dot{f}$ is small, while preserving the descent properties of the simple gradient updates. (It can be proven that the Gauss-Newton direction is always a descent direction since $\langle \dot{g}(0), \dot{g}(0) \rangle$ is positive definite.) The damped Gauss-Newton algorithm is defined by the update

$$U_{m+1} = U_m e^{\alpha_m X_m}, \tag{21}$$

where X_m and α_m are obtained by the extrinsic and intrinsic optimization as explained below and $U_{m=0}$ is a suitable initialization.

Tangent space optimization (extrinsic step). Let U_m be the previous update, then the objective function (8) is rewritten as

$$\mathscr{L}(U_m) = \frac{1}{2} \sum_k \langle g_k(U_m), g_k(U_m) \rangle, \tag{22}$$

where $g_k(U_m) = \mathrm{low}(U_m^T M_k U_m)$ and the Gauss-Newton approximate loss is

$$\begin{aligned} \delta\mathscr{L}(tX) = t \sum_r \langle \dot{g}_k(U_m e^{tX}), g_k(U_m e^{tX}) \rangle|_{t=0} \\ + \frac{t^2}{2} \sum_r \langle \dot{g}_k(U_m e^{tX}), \dot{g}_k(U_m e^{tX}) \rangle|_{t=0} \end{aligned} \tag{23}$$

with $\dot{g}_k(U_m e^{tX})|_{t=0} = \mathrm{low}(X\tilde{M}_k - \tilde{M}_k X)$ and where $\tilde{M}_k = U_m^T M_k U_m$. The tangent space direction X_m is the solution of

$$\min_x \quad x^T A x + b^T x \tag{24}$$

$$\text{s.t.} \quad Cx = 0 \tag{25}$$

where $x = \mathrm{vec}(tX)$, $A = \sum_k T_{\tilde{M}_k}^T \mathrm{Low} T_{\tilde{M}_k}$, and $b = \sum_k T_{\tilde{M}_k}^T \mathrm{Low}\,\mathrm{vec}(\tilde{M}_k)$, and $T_{\tilde{M}_k} = I \otimes \tilde{M}_k - \tilde{M}_k^T \otimes I$, and $C = (1 + \mathrm{Transp})$ with Transp being defined by $\mathrm{Transp}(\mathrm{vec}(A)) = \mathrm{vec}(A^T)$. This is a convex quadratic optimization problem with an equality constraint, whose solution is

$$X_m = -\mathrm{mat}(Y^T A Y)^{-1} Y b, \tag{26}$$

where Y is defined by $CY = 0$.

Exhaustive line search (intrinsic step). Given X_m, the optimal scaling is the solution of

$$\min_\alpha \quad \mathscr{L}(U_m e^{\alpha X_m}) \tag{27}$$

$$\text{s.t.} \quad 0 \leq \alpha \leq 1 \tag{28}$$

which is obtained by direct search on the discretized $[0, 1]$.

3.2. Global guarantees

In this section we argue how to combine the convergence properties of the GN algorithm and the perturbation theorem (1) to obtain certain global guarantees on tensor decomposition obtained by our approach. Due to its iterative nature, the GN algorithm is not expect to converge to the global optimum of (8) and the solution can depend on the particular initialization. However, our experiments have shown that the quality of the solution, in terms of the value of the objective function at convergence, is almost insensitive to the choice of the initialization. Moreover, it is possible to define convex relaxations to show that the final value is often 'close' to the global optimum (see Section (4.3)). These features seem to be shared by various optimization problems involving orthogonal matrices and may be related to the particular landscape defined by the quartic objective involved in simultaneous matrix decompositions.

However, we can provide a stronger and more quantitative argument pertaining to the convergence of the algorithm. The local convergence properties of the damped GN algorithm are easy to prove for the case of scalar variables and can be extended to the matrix manifold framework (see, for example, Absil et al. (2009)). This implies that the convergence to a given local optimum U_* is guaranteed if the algorithm is started in the basin of attraction of U_*. In particular, such a basin of attraction always includes a convex region containing U_*, i.e., a neighbourhood of U_* where the Hessian is positive definite. Thanks to the perturbation bounds obtained in Section 2.2, it is possible to characterize such a convex region via an upper bound on the perturbation parameter α_* defined in (9).

In particular, if the noise parameter σ is small enough, one can use a linear expansion of $\mathscr{L}(U)$ around $U_\circ$. A condition on σ is necessary at this point to guarantee that $U_\circ$ belongs to the convex region of U_*. The convergence of the algorithm to U_* is then ensured if it is possible to construct an initial solution inside the convex region. Since a good initialization can always be chosen by computing the single-matrix Schur decomposition of a linear combination of the input matrices $\hat{M}_k$, an initialization in the convex region is available if the noise σ is not too big. By combining the condition defining the basin of attraction of U_* and the initialization condition, one obtains a bound on σ that guarantees the algorithm to converge to the U_* that is the closest minimum to the ground truth triangularizer $U_\circ$. In this case, the corresponding estimation of the tensor components satisfies the bound in Theorem 1 and the existence statement of Theorem 1 is converted to a success guarantee for the recovered components. We refer to a forthcoming article (Colombo & Vlassis, 2016) for an explicit form of the sufficient condition on the noise level σ for guaranteed convergence to bounded solutions and a detailed proof.

Here we provide an informal sketch of the construction described above. An important point is the following: Under certain conditions on the joint eigenvalues of the matrices $M_k \in \mathscr{M}_0 = \{\hat{M}_k|_{\sigma=0}\}_{k=1}^d$, it can be shown that $\mathscr{L}(U, \mathscr{M}_0)$ has isolated minima for which the value of the objective is zero. Due to the continuity of $\mathscr{L}(U, \mathscr{M}_\sigma)$ in the parameter σ, one can expect that there exists a local minimum U_* of $\mathscr{L}(U, \mathscr{M}_\sigma)$ for each such isolated minimum of $\mathscr{L}(U, \mathscr{M}_0)$. Let $U_\circ$ be the exact joint triangularizer of $\mathscr{M}_0$ associated to U_*, then $U_* = U_\circ e^{\alpha_* X_*}$, with X_* and α_* characterized by Theorem 1. By definition, each U_* defines a convex region in which the Hessian of (8) is positive definite. If σ is small enough, $U_\circ$ belongs to the same convex region and it makes sense to consider a linear expansion around it. This is equivalent to assuming that σ is such that the Hessian of $\mathscr{L}(U, \mathscr{M}_\sigma)$ at $U_\circ$ is positive definite. Then the global guarantees on the convergence of the algorithm can be obtained as follows:

1. Find the local convexity condition $H(U) > 0$, or equivalently $\langle X, H(U)X\rangle > 0$ for all X, where $H(U)$ is the Hessian at the point U and $\langle A, B\rangle = \mathrm{Tr}(A^T B)$.

2. Given σ, find an $\alpha_{max} = \alpha_{max}(\sigma, \mathscr{M}_0)$ such that $H(U_\circ e^{\alpha Y}) > 0$ for all (α, Y) such that $\|Y\| = 1$ and all $\alpha \leq \alpha_{max}$. This is possible by expanding $H(U)$ in a neighbourhood of $U_\circ$ and using the fact that, under certain conditions on the joint eigenvalues of M_k, the Hessian of $\mathscr{L}(U, \mathscr{M}_0)$ at $U_\circ$ is positive definite. In particular, given the Schur decomposition of a ground-truth matrix M_k, it can be shown that the distance in norm between the unperturbed triangularizers $U_\circ$ and the triangularizers U_* of $\hat{M}_k$ is proportional to the perturbation level σ and the inverse of the eigengap $\gamma = \min_{i\neq j} |\lambda_i(M_k) - \lambda_j(M_k)|$ of the unperturbed matrix M_k. Now, given a set of nearly joint diagonalizable matrices this non-degeneracy condition generalizes to

$$\forall r \neq r' \quad \exists\, k \in 1, \ldots, d \quad \text{s.t.} \quad [\Lambda_k]_{rr} \neq [\Lambda_k]_{r'r'}. \tag{29}$$

   It can be shown that in this case the operator $\mathscr{T}$ defined in (15) and the Hessian of $\mathscr{L}(U, \mathscr{M}_0)$ at $U = U_\circ$, where $U_\circ$ is an exact joint triangularizer of all M_k, are both positive definite.

3. Let $U_{init} = U_\circ e^{\alpha_{init} X_{init}}$, where $\|X_{init}\| = 1$ and $\alpha_{init} > 0$, and find $\bar{\alpha}_{init} = \bar{\alpha}_{init}(\sigma, \mathscr{M}_0)$ such that $\alpha_{init} \leq \bar{\alpha}_{init}$, up to second order terms. This is possible by assuming that U_{init} is obtained from the Schur decomposition of a linear combination of the matrices $\hat{M}_n$ that has real-separated eigenvalues. Perturbation bounds on single matrix Schur decomposition are known (Konstantinov et al., 1994).

4. Finally, require that U_{init} belongs to the local convexity region of $U_\circ$ by imposing $\alpha_{init} \leq \alpha_{max}$. This condition is satisfied if σ is small enough to satisfy $\bar{\alpha}_{init} \leq \alpha_{max}$.

4. Experiments

We have compared our method with two other methods for tensor decomposition, the `cpd3-sgsd` algorithm of De Lathauwer et al. (2004) and the `no-tenfact` algorithm of Kuleshov et al. (2015). For both methods, we have used the Matlab codes available online with default settings, except for the option `options.Symmetry` $=$ $\{[123]\}$ in the `cpd3-sgsd` algorithm. In the following 'GN' denotes the proposed Gauss-Newton algorithm.

4.1. Comparison on synthetic data

We first tested the algorithms on synthetic data with known ground truth. For two sets of dimensionality-rank settings, namely $d = \{10, 20\}$ and $p = \{\frac{d}{2}, d\}$, we have randomly generated 10 distinct ground-truth models

$$T_{ijk} = \sum_{r=1}^{p} w_r Z_{ir} Z_{jr} Z_{kr}, \tag{30}$$

where $w_r > 0$ for all $r = 1, \ldots, p$, $1^T w = 1$ and $\sum_{i=1}^d Z_{ir}^2 = 1$ for all r, and the corresponding noise tensors E such that $E_{ijk} = \mathscr{N}(0, 1)$, for all $i, j, k = 1, \ldots, d$. The input tensor for a given experiment was given by

$$\hat{T} = T + \epsilon \frac{\|T\|}{\|E\|} E, \tag{31}$$

where $\epsilon \in [0, 10^{-2}]$. The distance in norm between the (rescaled) recovered matrix $Z_* = [z_1^*, \ldots z_p^*]$ and the ground-truth $Z = [z_1, \ldots z_p]$ has been used to evaluated the quality of the decomposition via the score function

$$\text{error} = \frac{\|Z - Z_*\|}{\|Z + Z_*\|}. \tag{32}$$

Since all algorithms can estimate Z up to permutations of the columns, a reordering step was performed before the evaluation. In Figure 1 we show the average performance over the 10 experiments for different amounts of noise. For all the considered levels of noise and all dimensionality-rank settings our algorithms obtained results that are statistically equivalent to `cpd3-sgsd`.

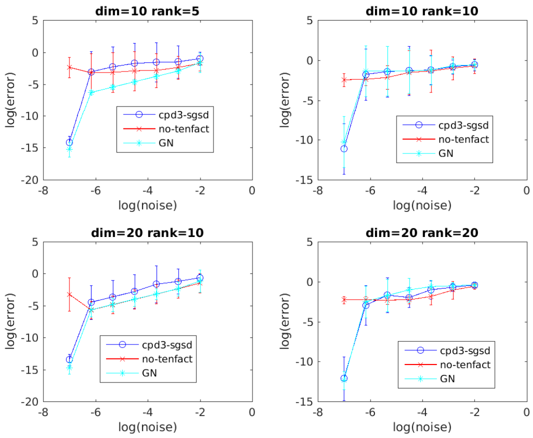

Figure 1. Decomposition of a symmetric nonorthogonal tensor. On the y axis, the logarithm of the recovery error defined by $s = \frac{\|Z-Z_*\|}{\|Z+Z_*\|}$. On the x axis, the logarithm noise parameter $\epsilon \in [0, 10^{-2}]$ (see (31)). For each algorithm and each considered noise level, a point represents the average $\bar{s}$ over 10 experiments. The corresponding errorbars define the range $\log(\bar{s}) \pm \frac{\sigma}{\bar{s}}$ where σ is the standard deviation over the distinct experiments.

4.2. Random Projections

Unlike most matrix-based tensor decomposition algorithms (Chang, 1996; Mossel & Roch, 2006; Anandkumar et al., 2014b; Hsu & Kakade, 2013; Kuleshov et al., 2015) the proposed method does not make use of random projections of the tensor slides. Single matrix perturbation bounds usually depend inversely on the eigenvalues spacing of the ground-truth matrix and the contraction to random vectors has been used to prove, via usual concentration bounds on the normal distribution, that a minimum eigengap can be guaranteed. In the tensor decomposition via a single matrix decomposition, as for example in Chang (1996) and Mossel & Roch (2006), the random projection is a statistically natural choice for compressing the information of all the tensor slides into a single matrix diagonalization problem. For a simultaneous matrix decomposition approach this is not required, since all unprojected slices can be decomposed simultaneously as described in Section 2.1. Moreover, the inverse dependence on the eigengap is replaced by the softer dependence on averaging factors as $\gamma = \frac{1}{d}\min_{r<r'}\sum_k (Z_{kr} - Z_{kr'})^2$. In the framework of matrix-based approaches, methods with (Kuleshov et al., 2015) or without (Cardoso, 1991; De Lathauwer et al., 2004) random projections have been proposed.

To test the effect of random projections in a simultaneous decomposition approach, we have modified our algorithm in order to allow for an arbitrary number N_θ of random projections, and we have tested the performance for different $N_\theta = [0, 500]$. Figure 2 shows the performance of the algorithm on a set of synthetic experiments analogous to the ones described in the previous section. It is clear that increasing the number of random projections does not improve the accuracy of the output while it increases its runtime (data not shown). A visual comparison between the result of the `no-tenfact` algorithm (Kuleshov et al., 2015) in the previous set of experiments and the performance of our algorithm for $N_\theta > 200$ seems to suggest that a high number of random projection can even reduce the quality of the output when the noise is small.

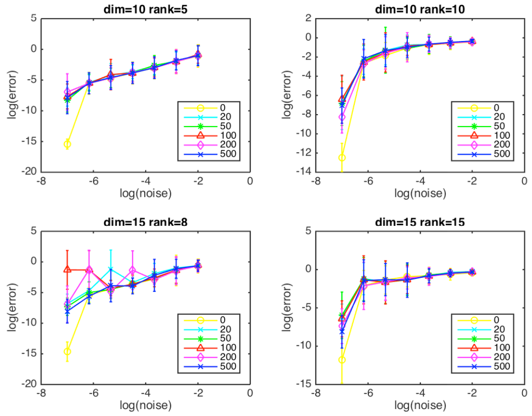

Figure 2. Performance of a modified version of our algorithm supporting an arbitrary number N_θ of random projections. The x-axis represents the noise level and the y-axis the recovery error as defined in Figure 1, points indicate the average score over 10 experiments.

4.3. Jacobi vs Gauss-Newton

The most popular algorithm for solving the problem of approximate joint Schur decomposition in (8) is the Jacobi algorithm by Haardt & Nossek (1998). A related method has been proposed by Abed-Meraim & Hua (1998). The Jacobi algorithm was first proposed for the problem of simultaneous diagonalization of symmetric matrices (Cardoso & Souloumiac, 1996). In the framework of simultaneous diagonalization, various alternative methods have been proposed. An interesting matrix manifold approach is the gradient-based method proposed by Afsari & Krishnaprasad (2004), but we are not aware of any alternative to the Jacobi algorithm for simultaneous Schur Decomposition.

We have compared the performance of the Jacobi algorithm of Haardt & Nossek (1998) (our implementation) and the Gauss-Newton algorithm proposed here on the simultaneous triangularization of 100 randomly generated nearly commuting matrices. The nearly commuting matrices have been generated by choosing a random eigenvector matrix

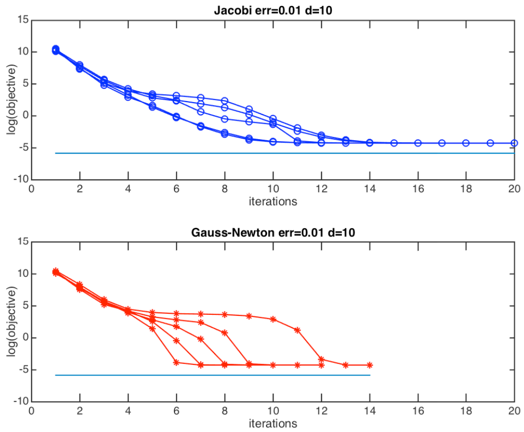

Figure 3. Comparison of the Jacobi algorithm described by Haardt & Nossek (1998) (our implementation) and the Gauss-Newton presented in this work on an approximate join Schur decomposition task. 100 nearly commuting matrices have been generated and approximately triangularized by the two algorithms. For five random initializations, the plot shows the value of the objective function at each iteration. The flat line is a lower bound on the global optimum, computed via the relaxation (34).

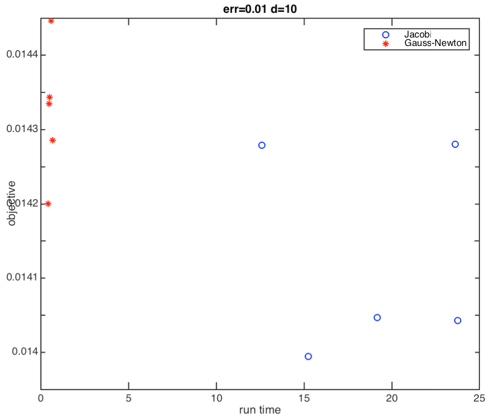

Figure 4. Quality of the output and runtime of the Jacobi algorithm of Haardt & Nossek (1998) (our implementation) and the Gauss-Newton on the approximate joint triangularization problem described in Section 4.3. The algorithms were randomly initialized five times and each dot represents the corresponding final objective value (y-axis) and the total runtime (x-axis). For an essentially equivalent quality of the result, the Gauss-Newton algorithm turns out to be much faster. The larger runtime of the Jacobi algorithm seems not to be due to a larger number of iterations but to the higher computational complexity per iteration.

$V \in \mathbf{R}^{d\times d}$, $d = 10$ and letting

$$\hat{M}_k = V \Lambda_k V^{-1} + \epsilon \frac{E_k}{\|E_k\|}, \tag{33}$$

$k = 1, \dots 100$, and where Λ_k were 100 random diagonal matrices. In Figure 3 we show the objective value as a function of the number of iteration for five different initializations, and in Figure 4 we show a scatterplot of the obtained final values against runtime. The large runtime of the Jacobi algorithm is not due to a larger number of iterations but to the higher computational complexity per iteration. To show that optimality is almost attained, we also plot in Figure 3 the objective value of the relaxed objective

$$L(V, \mathscr{M}_\epsilon) = \sum_{k=1}^{100} \|P_{\text{low}} V \text{vec}(\hat{M}_k)\|^2. \tag{34}$$

The global optimum of (34) can be computed from the singular value decomposition of a $d^2 \times d^2$ matrix formed by stacking together the 100 vectorized matrices $\hat{M}_k$. This relaxation is based on the observation that $L(U \otimes U, \mathscr{M}_\epsilon) = \mathscr{L}(U, \mathscr{M}_\epsilon)$, where $\mathscr{L}$ is the objective function defined in (8).

4.4. Real data experiment

To test the performance of our algorithm on real-world data we have chosen a label prediction problem from crowdsourcing data. The problem and the dataset are described by Zhang et al. (2014) where an estimator based on order-three moments is also proposed. This technique allows one to solve the label prediction problem by means of a canonical nonorthogonal tensor decomposition, via the inference of suitable confusion matrices. We have integrated the above three algorithms in this estimator and have computed the corresponding scores in predicting the label of the 'birds' and 'dog' datasets considered in Zhang et al. (2014). We report the average score over 10 prediction trials, since the spectral estimator seems to depend on the choice of the random partition of 'workers' used for the inference. The following table shows that the three algorithms on this task are statistically equivalent.

	GN	cpd3-sgsd	no-tenfact
birds dataset	0.70 ± 0.08	0.70 ± 0.08	0.66 ±0.10
dogs dataset	0.64 ± 0.28	0.64 ± 0.28	0.64 ±0.27

5. Conclusions

We presented a new algorithm for tensor decomposition that performs joint matrix Schur decomposition on a set of nearly-commuting observable matrices extracted from the slices of the tensor, and we carried out a first-order perturbation analysis. Our algorithm is faster and at least as accurate and robust than existing tensor decomposition algorithms. Ongoing work involves extending our algorithm and analysis to the case of nonnegative tensors.

References

Abed-Meraim, K and Hua, Y. A least-squares approach to joint Schur decomposition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 4, pp. 2541–2544. IEEE, 1998.

Absil, P-A, Mahony, R, and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Afsari, B. Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1148–1171, 2008.

Afsari, B and Krishnaprasad, PS. Some gradient based joint diagonalization methods for ICA. In *Independent Component Analysis and Blind Signal Separation*, pp. 437–444. Springer, 2004.

Anandkumar, A, Ge, R, Hsu, D, and Kakade, SM. A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15:2239–2312, 2014a.

Anandkumar, A, Ge, R, Hsu, D, Kakade, SM, and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014b.

Azizzadenesheli, K, Lazaric, A, and Anandkumar, A. Reinforcement learning of POMDPs using spectral methods. In *Proceedings of the 29th Conference on Learning Theory*, 2016.

Cardoso, J-F. Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors. In *International Conference on Acoustics, Speech, and Signal Processing*, pp. 3109–3112. IEEE, 1991.

Cardoso, J-F. Perturbation of joint diagonalizers. *Telecom Paris, Signal Department, Technical Report 94D023*, 1994.

Cardoso, J-F and Souloumiac, A. Jacobi angles for simultaneous diagonalization. *SIAM journal on matrix analysis and applications*, 17(1):161–164, 1996.

Chang, JT. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math Biosci*, 137(1):51–73, October 1996.

Colombo, N and Vlassis, N. Approximate joint matrix triangularization. *arXiv*, 2016.

Comon, P. Tensors: a brief introduction. *IEEE Signal Processing Magazine*, 31(3):44–53, 2014.

De Lathauwer, L. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 28 (3):642–666, 2006.

De Lathauwer, L, De Moor, B, and Vandewalle, J. Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition. *SIAM Journal on Matrix Analysis and Applications*, 26(2):295–327, 2004.

Haardt, M and Nossek, JA. Simultaneous Schur decomposition of several nonsymmetric matrices to achieve automatic pairing in multidimensional harmonic retrieval problems. *Signal Processing, IEEE Transactions on*, 46(1):161–169, 1998.

Harshman, R. Foundations of the PARAFAC procedure: Model and conditions for an explanatory factor analysis. Technical report, UCLA Working Papers in Phonetics 16, University of California, Los Angeles CA, 1970.

Horn, RA and Johnson, CR. *Matrix analysis*. Cambridge University Press, 2nd edition, 2012.

Hsu, D and Kakade, SM. 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. ACM, 2013.

Konstantinov, MM, Petkov, PHr, and Christov, ND. Nonlocal perturbation analysis of the Schur system of a matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(2):383–392, 1994.

Kruskal, JB. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

Kuleshov, V, Chaganty, A, and Liang, P. Tensor factorization via matrix factorization. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Leurgans, SE, Ross, RT, and Abel, RB. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993.

Montanari, A and Richard, E. A statistical model for tensor PCA. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2897–2905. 2014.

Mossel, E and Roch, S. Learning nonsingular phylogenies and hidden Markov models. *The Annals of Applied Probability*, 16 (2):583–614, May 2006.

Sidiropoulos, ND and Bro, R. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of chemometrics*, 14 (3):229–239, 2000.

Song, L, Anandkumar, A, Dai, B, and Xie, B. Nonparametric estimation of multi-view latent variable models. In *Proceedings of the 31st International Conference on Machine Learning*. 2015.

Souloumiac, A. Joint diagonalization: Is non-orthogonal always preferable to orthogonal? In *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009.

Sun, J-G. Perturbation bounds for the generalized Schur decomposition. *SIAM journal on matrix analysis and applications*, 16(4):1328–1340, 1995.

Vempala, SS and Xiao, Y. Max vs Min: Tensor decomposition and ICA with nearly linear sample complexity. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

Zhang, Y, Chen, X, Zhou, D, and Jordan, MI. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, pp. 1260–1268, 2014.