Research Article

Journal of CHEMOMETRICS

Received: 29 November 2008, Revised: 12 April 2009, Accepted: 16 April 2009, Published online in Wiley Interscience: 23 June 2009

(www.interscience.wiley.com) DOI: 10.1002/cem.1244

Nonnegative approximations of nonnegative tensors

Lek-Heng Lim[a*] and Pierre Comon[b]

We study the decomposition of a nonnegative tensor into a minimal sum of outer product of nonnegative vectors and the associated parsimonious naïve Bayes probabilistic model. We show that the corresponding approximation problem, which is central to nonnegative PARAFAC, will always have optimal solutions. The result holds for any choice of norms and, under a mild assumption, even Brègman divergences. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: nonnegative tensors; nonnegative hypermatrices; nonnegative tensor decompositions; nonnegative tensor rank; low-rank tensor approximations; probabilistic latent semantic indexing; CANDECOMP; PARAFAC, tensor norm; tensor Brègman divergence

1. DEDICATION

This paper is dedicated to the memory of our late colleague Richard Allan Harshman. It is loosely organized around two of Harshman's best known works—PARAFAC (parallel factor analysis) [1] and LSI (latent semantic indexing) [2], and answers two questions that he posed. We target this paper at a technometrics readership.

In Section 4, we discussed a few aspects of nonnegative tensor factorization and Hofmann's PLSI, a variant of the LSI model co-proposed by Harshman [2]. In Section 5, we answered a question of Harshman on why the apparently unrelated construction of Bini, Capovani, Lotti, and Romani in Reference [3] should be regarded as the first example of what he called "PARAFAC degeneracy" [4]. Finally in Section 6, we showed that such PARAFAC degeneracy will not happen for nonnegative approximations of nonnegative tensors, answering another question of his.

2. INTRODUCTION

The *decomposition* of a tensor into a minimal sum of outer products of vectors was first studied by Hitchcock [5,6] in 1927. The topic has a long and illustrious history in algebraic computational complexity theory (cf. [7] and the nearly 600 references in its bibliography) dating back to Strassen's celebrated result [8]. It has also recently found renewed interests, coming most notably from algebraic statistics and quantum computing.

However the study of the corresponding *approximation* problem, i.e., the approximation of a tensor by a sum of outer products of vectors, probably first surfaced as data analytic models in psychometrics in the work of Harshman [1], who called his model PARAFAC (for parallel factor analysis), and the work of Carrol and Chang [9], who called their model CANDECOMP (for canonical decomposition).

The CANDECOMP/PARAFAC model, sometimes abbreviated as CP model, essentially asks for a solution to the following problem: given a tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, find an optimal rank-$r$ approximation to A,

$$X_r \in \operatorname*{argmin}_{\operatorname{rank}(X) \le r} \|A - X\| \tag{1}$$

or, more precisely, find scalars λ_p and unit vectors[†] $\mathbf{u}_p, \mathbf{v}_p, \dots, \mathbf{z}_p$, $p = 1, \dots, r$, that minimizes

$$\left\| A - \sum_{p=1}^{r} \lambda_p \, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p \right\| \tag{2}$$

The norm $\|\cdot\|$ here is arbitrary and we will discuss several natural choices in the next section. When $k = 2$, A becomes a matrix and a solution to the problem when $\|\cdot\|$ is unitarily invariant is given by the celebrated Eckart-Young theorem [10]: X_r may be taken to be

$$X_r = \sum_{p=1}^{r} \sigma_p \, \mathbf{u}_p \otimes \mathbf{v}_p$$

where $\sigma_1 \ge \cdots \ge \sigma_r$ are the first r singular values of A and $\mathbf{u}_p$, $\mathbf{v}_p$ the corresponding left and right singular vectors.

However, when $k \ge 3$ the problem becomes more subtle. In fact, a global minimizer of Equation (2) may not even exist as

[*] *Correspondence to: L.-H. Lim, Department of Mathematics, University of California, Berkeley, CA 94720-3840, USA.*
E-mail: lekheng@math.berkeley.edu

[a] *L.-H. Lim*
Department of Mathematics, University of California, Berkeley, USA

[b] *P. Comon*
University of Nice, Nice, France

[†] Whenever possible, we will use $\mathbf{u}_p, \mathbf{v}_p, \dots, \mathbf{z}_p$ instead of the more cumbersome $\mathbf{u}_p^{(1)}, \mathbf{u}_p^{(2)}, \dots, \mathbf{u}_p^{(k)}$ to denote the vector factors in an outer product. It is to be understood that there are k vectors in "$\mathbf{u}_p, \mathbf{v}_p, \dots, \mathbf{z}_p$," where $k \ge 3$.

soon as $k \geq 3$; in which case the Problem (1) is ill-posed because the set of minimizers is empty. We refer the reader to Section 5 for examples and discussions. Nevertheless we will show that for nonnegative tensors the problem of finding a best nonnegative rank-*r* approximation always has a solution, i.e., Equation (2) will always have a global minimum when A and $\mathbf{u}_p,\mathbf{v}_p,\ldots,\mathbf{z}_p$ are required to be nonnegative. Such nonnegativity arises naturally in applications. For example, in the context of chemometrics, sample concentration and spectral intensity often cannot assume negative values [11–16]. Nonnegativity can also be motivated by the data analytic tenet [17] that the way "basis functions" combine to build "target objects" is an exclusively additive process and should not involve any cancellations between the basis functions. For $k = 2$, this is the motivation behind *nonnegative matrix factorization* (NMF) [16,17], essentially a decomposition of a nonnegative matrix $A \in \mathbb{R}^{m\times n}$ into a sum of outer-products of nonnegative vectors,

$$A = WH^{\top} = \sum_{p=1}^{r} \mathbf{w}_p \otimes \mathbf{h}_p$$

or, in the noisy situation, the approximation of a nonnegative matrix by such a sum:

$$\min_{W\geq 0, H\geq 0} \|A - WH^{\top}\| = \min_{\mathbf{w}_p\geq 0, \mathbf{h}_p\geq 0} \left\| A - \sum_{p=1}^{r} \mathbf{w}_p \otimes \mathbf{h}_p \right\|$$

The generalization of NMF to tensors of higher order yields a model known as *nonnegative* PARAFAC [13–15], which has also been studied more recently under the name *nonnegative tensor factorization* (NTF) [18]. As we have just mentioned, a general tensor can fail to have a best low-rank approximation. So the first question that one should ask in a multilinear generalization of a bilinear model is whether the generalized problem would still have a solution—and this was a question that Harshman posed [19]. More generally, we will show that nonnegative PARAFAC always has a solution for any continuous measure of proximity satisfying some mild conditions, e.g., norms or Brègman divergences. These include the sum-of-squares loss and Kullback-Leibler divergence commonly used in NMF and NTF.

The following will be proved in Sections 6 and 7. Let $\Omega_0 \subseteq \Omega \subseteq \mathbb{R}_+^{d_1\times\cdots\times d_k}$ be closed convex subsets. Let $d : \Omega \times \Omega_0 \to \mathbb{R}$ be a norm or a Brègman divergence. For any nonnegative tensor $A \in \Omega$ and any given $r \in \mathbb{N}$, a *best nonnegative rank-r approximation* always exists in the sense that the following infimum

$$\inf\{d(A, X) \mid X \in \Omega_0, \operatorname{rank}_+(X) \leq r\}$$

is attained by some nonnegative tensor $X_r \in \Omega_0$, $\operatorname{rank}_+(X_r) \leq r$. In particular, the nonnegative tensor approximation problem

$$X_r \in \operatorname*{argmin}_{\operatorname{rank}_+(X)\leq r} \|A - X\|$$

is well-posed. Here $\operatorname{rank}_+(X)$ denotes the *nonnegative rank* of X and will be formally introduced in Section 4.

3. TENSORS AS HYPERMATRICES

Let $V_1,\ldots,V_k$ be real vector spaces of dimensions $d_1,\ldots,d_k$, respectively. An element of the tensor product $V_1 \otimes \cdots \otimes V_k$ is called an order-*k tensor*. Up to a choice of bases on $V_1,\ldots,V_k$, such a tensor may be represented by a $d_1 \times \cdots \times d_k$ array of real numbers[‡],

$$A = [\![a_{j_1\cdots j_k}]\!]_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} \in \mathbb{R}^{d_1\times\cdots\times d_k} \tag{3}$$

Gelfand, Kapranov, and Zelevinsky called such coordinate representations of abstract tensors *hypermatrices* [20]. It is worth pointing out that an *array* is just a data structure but like matrices, hypermatrices are more than mere arrays of numerical values. They are equipped with algebraic operations arising from the algebraic structure of $V_1 \otimes \cdots \otimes V_k$:

- *Addition and Scalar Multiplication*: For $[\![a_{j_1\cdots j_k}]\!],[\![b_{j_1\cdots j_k}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_k}$ and $\lambda,\mu \in \mathbb{R}$,

$$\lambda[\![a_{j_1\cdots j_k}]\!] + \mu[\![b_{j_1\cdots j_k}]\!] = [\![\lambda a_{j_1\cdots j_k} + \mu b_{j_1\cdots j_k}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_k} \tag{4}$$

- *Outer Product Decomposition*: Every $A = [\![a_{j_1\cdots j_k}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_k}$ may be decomposed as

$$A = \sum_{p=1}^{r} \lambda_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p, \qquad a_{j_1\cdots j_k} = \sum_{p=1}^{r} \lambda_p u_{pj_1} v_{pj_2} \cdots z_{pj_k} \tag{5}$$

with $\lambda_p \in \mathbb{R}$, $\mathbf{u}_p = [u_{p1},\ldots,u_{pd_1}]^{\top} \in \mathbb{R}^{d_1},\ldots,\mathbf{z}_p = [z_{p1},\ldots, z_{pd_k}]^{\top} \in \mathbb{R}^{d_k}$, $p = 1,\ldots,r$.

The symbol $\otimes$ denotes the *Segre outer product*: For vectors $\mathbf{x} = [x_1,\ldots,x_l]^{\top} \in \mathbb{R}^l$, $\mathbf{y} = [y_1,\ldots,y_m]^{\top} \in \mathbb{R}^m$, $\mathbf{z} = [z_1,\ldots,z_n]^{\top} \in \mathbb{R}^n$, the quantity $\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z}$, is simply the 3-hypermatrix $[\![x_i y_j z_k]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l\times m\times n}$, with obvious generalization to an arbitrary number of vectors.

It follows from Equation (4) that $\mathbb{R}^{d_1\times\cdots\times d_k}$ is a vector space of dimension $d_1\cdots d_k$. The existence of a decomposition (5) distinguishes $\mathbb{R}^{d_1\times\cdots\times d_k}$ from being merely a vector space by endowing it with a tensor product structure. While as real vector spaces, $\mathbb{R}^{l\times m\times n}$ (hypermatrices), $\mathbb{R}^{lm\times n}$, $\mathbb{R}^{ln\times m}$, $\mathbb{R}^{mn\times l}$ (matrices), and $\mathbb{R}^{lmn}$ (vectors) are all isomorphic, the tensor product structure distinguishes them. Note that a different choice of bases on $V_1,\ldots,V_k$ would lead to a different hypermatrix representation of elements in $V_1 \otimes \cdots \otimes V_k$. So strictly speaking, a tensor and a hypermatrix are different in the same way a linear operator and a matrix are different. Furthermore, just as a bilinear functional, a linear operator, and a dyad may all be represented by the same matrix, different types of tensors may be represented by the same hypermatrix if one disregards covariance and contravariance. Nonetheless the term "tensor" has been widely used to mean a hypermatrix in the data analysis communities (including bioinformatics, computer vision, machine learning, neuroinformatics, pattern recognition, signal processing, technometrics), and we will refrain from being perverse and henceforth adopt this naming convention. For the

[‡] The subscripts and superscripts will be dropped when the range of $j_1,\ldots,j_k$ is obvious or unimportant. We use double brackets to delimit hypermatrices.

more pedantic readers, it is understood that what we call a tensor in this paper really means a hypermatrix.

A non-zero tensor that can be expressed as an outer product of vectors is called a rank-1 tensor. More generally, the *rank* of a tensor $A = [\![a_{j_1\cdots j_k}]\!]_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} \in \mathbb{R}^{d_1\times\cdots\times d_k}$, denoted rank($A$), is defined as the minimum r for which A may be expressed as a sum of r rank-1 tensors [5,6],

$$\operatorname{rank}(A) := \min\left\{ r \,\middle|\, A = \sum_{p=1}^{r} \lambda_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p \right\} \tag{6}$$

The definition of rank in Equation (6) agrees with the definition of matrix rank when applied to an order-2 tensor.

The *Frobenius norm* or *F-norm* of a tensor $A = [\![a_{j_1\cdots j_k}]\!]_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} \in \mathbb{R}^{d_1\times\cdots\times d_k}$ is defined by

$$\|A\|_F = \left[\sum_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} |a_{j_1\cdots j_k}|^2 \right]^{\frac{1}{2}} \tag{7}$$

The F-norm is by far the most popular choice of norms for tensors in data analytic applications. However when A is nonnegative valued, then there is a more natural norm that allows us to interpret the normalized values of A as probability distribution values, as we will see in the next section. With this in mind, we define the *E-norm* and *G-norm* by

$$\|A\|_E = \sum_{i_1,\ldots,i_k=1}^{d_1,\ldots,d_k} |a_{j_1\cdots j_k}| \tag{8}$$

and

$$\|A\|_G = \max\{|a_{j_1\cdots j_k}| \mid j_1 = 1,\ldots,d_1;\cdots;j_k = 1,\ldots,d_k\}$$

Observe that the E-, F-, and G-norms of a tensor A are simply the l^1-, l^2-, and l^∞-norms of A regarded as a vector of dimension $d_1\cdots d_k$. Furthermore they are multiplicative on rank-1 tensors in the following sense:

$$\begin{aligned}
\|\mathbf{u}\otimes\mathbf{v}\otimes\cdots\otimes\mathbf{z}\|_E &= \|\mathbf{u}\|_1\|\mathbf{v}\|_1\cdots\|\mathbf{z}\|_1,\\
\|\mathbf{u}\otimes\mathbf{v}\otimes\cdots\otimes\mathbf{z}\|_F &= \|\mathbf{u}\|_2\|\mathbf{v}\|_2\cdots\|\mathbf{z}\|_2,\\
\|\mathbf{u}\otimes\mathbf{v}\otimes\cdots\otimes\mathbf{z}\|_G &= \|\mathbf{u}\|_\infty\|\mathbf{v}\|_\infty\cdots\|\mathbf{z}\|_\infty
\end{aligned} \tag{9}$$

The F-norm has the advantage of being induced by an inner product on $\mathbb{R}^{d_1\times\cdots\times d_k}$, namely,

$$\langle A, B\rangle = \sum_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} a_{j_1\cdots j_k} b_{j_1\cdots j_k} \tag{10}$$

As usual, it is straightforward to deduce a Cauchy-Schwarz inequality

$$|\langle A, B\rangle| \le \|A\|_F\|B\|_F$$

and a Hölder inequality

$$|\langle A, B\rangle| \le \|A\|_E\|B\|_G$$

Many other norms may be defined on a space of tensors. For any $1 \le p \le \infty$, one may define the l^p-equivalent of Equation (7), of which E-, F-, and G-norms are special cases. Another common class of tensor norms generalizes operator norms of matrices: For example if $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l\times m\times n}$ and

$$A(\mathbf{x},\mathbf{y},\mathbf{z}) := \sum_{i,j,k=1}^{l,m,n} a_{ijk} x_i y_j z_k$$

denotes the associated trilinear functional, then

$$\|A\|_{p,q,r} := \sup_{\mathbf{x},\mathbf{y},\mathbf{z}\neq\mathbf{0}} \frac{|A(\mathbf{x},\mathbf{y},\mathbf{z})|}{\|\mathbf{x}\|_p\|\mathbf{y}\|_q\|\mathbf{z}\|_r}$$

defines a norm for any $1 \le p, q, r \le \infty$. Nevertheless all these norms are equivalent (and thus induce the same topology) since the tensor product spaces here are finite-dimensional. In particular, the results in this paper apply to any choice of norms since they pertain to the convergence of sequences of tensors.

The discussion in this section remains unchanged if $\mathbb{R}$ is replaced by $\mathbb{C}$ throughout (apart from a corresponding replacement of the Euclidean inner product in Equation (10) by the Hermitian inner product) though a minor caveat is that the tensor rank as defined in Equation (6) depends on the choice of base field (see Reference [21] for a discussion).

4. NONNEGATIVE DECOMPOSITION OF NONNEGATIVE TENSORS

We will see that a finite collection of discrete random variables satisfying both the naïve Bayes hypothesis and the Ockham principle of parsimony has a joint probability distribution that, when regarded as a nonnegative tensor on the probability simplex, decomposes in a nonnegative rank-revealing manner that parallels the matrix singular value decomposition. This generalizes Hofmann's probabilistic variant [22] of *latent semantic indexing* (LSI), a well-known technique in natural language processing and information retrieval that Harshman played a role in developing [2]. Nonnegative tensor decompositions were first studied in the context of PARAFAC with nonnegativity constraints by the technometrics communities [11–15]. The interpretation as a naïve Bayes decomposition of probability distributions into conditional distributions was due to Garcia, Stillman, and Sturmfels [23] and Sashua and Hazan [18]. It is perhaps worth taking this opportunity to point out a minor detail that had somehow been neglected in References [18,23]: the naïve Bayes hypothesis is not sufficient to guarantee a nonnegative rank-revealing decomposition, one also needs the Ockham principle of parsimony, i.e., the hidden variable in question has to be minimally supported.

A tensor $A = [\![a_{j_1\cdots j_k}]\!]_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} \in \mathbb{R}^{d_1\times\cdots\times d_k}$ is *nonnegative*, denoted $A \ge 0$, if all $a_{j_1\cdots j_k} \ge 0$. We will write $\mathbb{R}_+^{d_1\times\cdots\times d_k} := \{A \in \mathbb{R}^{d_1\times\cdots\times d_k} \mid A \ge 0\}$. For $A \ge 0$, a *nonnegative outer-product decomposition* is one of the form

$$A = \sum_{p=1}^{r} \delta_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p \tag{11}$$

where $\delta_p \geq 0$ and $\mathbf{u}_p,\mathbf{v}_p,\ldots,\mathbf{z}_p \geq 0$ for $p = 1,\ldots,r$. It is clear that such a decomposition exists for any $A \geq 0$. The minimal r for which such a decomposition is possible will be called the *nonnegative rank*. For $A \geq 0$, this is denoted and defined via

$$\operatorname{rank}_+(A) := \min\left\{ r \,\middle|\, A = \sum_{p=1}^{r} \delta_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p, \quad \delta_p, \mathbf{u}_p, \mathbf{v}_p, \ldots, \mathbf{z}_p \geq 0 \text{ for all } p \right\}$$

Let Δ^d denote the *unit d-simplex*, i.e., the convex hull of the standard basis vectors in $\mathbb{R}^{d+1}$. Explicitly,

$$\Delta^d := \left\{ \sum_{p=1}^{d+1} \delta_p \mathbf{e}_p \in \mathbb{R}^{d+1} \,\middle|\, \sum_{p=1}^{d+1} \delta_p = 1, \quad \delta_1, \ldots, \delta_{d+1} \geq 0 \right\}$$
$$= \{\mathbf{x} \in \mathbb{R}_+^{d+1} \mid \|\mathbf{x}\|_1 = 1\}$$

For nonnegative valued tensors, the E-norm has the advantage that Equation (8) reduces to a simple sum of all entries. This simple observation leads to the following proposition stating that the decomposition in Equation (11) may be realized over unit simplices if we normalize A by its E-norm.

Proposition 4.1. *Let $A \in \mathbb{R}_+^{d_1\times\cdots\times d_k}$ be a nonnegative tensor with $\operatorname{rank}_+(A) = r$. Then there exist $\boldsymbol{\delta} = [\delta_1,\ldots,\delta_r]^\top \in \mathbb{R}_+^r$, $\mathbf{u}_p \in \mathbb{R}_+^{d_1-1}$, $\mathbf{v}_p \in \mathbb{R}_+^{d_2-1}, \ldots, \mathbf{z}_p \in \mathbb{R}_+^{d_k-1}$, $p = 1,\ldots,r$, where*

$$\|\boldsymbol{\delta}\|_1 = \|A\|_E$$

and

$$\|\mathbf{u}_p\|_1 = \|\mathbf{v}_p\|_1 = \cdots = \|\mathbf{z}_p\|_1 = 1$$

such that

$$A = \sum_{p=1}^{r} \delta_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p \tag{12}$$

Proof. If $A = 0$, this is obvious. So we will suppose that $A \neq 0$. By the minimality of $r = \operatorname{rank}_+(A)$, we know that $\mathbf{u}_p,\mathbf{v}_p,\ldots,\mathbf{z}_p$ in Equation (12) are all nonzero and we may assume that

$$\|\mathbf{u}_p\|_1 = \|\mathbf{v}_p\|_1 = \cdots = \|\mathbf{z}_p\|_1 = 1$$

since otherwise we may normalize

$$\hat{\mathbf{u}}_p = \mathbf{u}_p/\|\mathbf{u}_p\|_1,\, \hat{\mathbf{v}}_p = \mathbf{v}_p/\|\mathbf{v}_p\|_1, \ldots, \hat{\mathbf{z}}_p = \mathbf{z}_p/\|\mathbf{z}_p\|_1$$

and set

$$\hat{\delta}_p = \delta_p \|\mathbf{u}_p\|_1 \|\mathbf{v}_p\|_1 \cdots \|\mathbf{z}_p\|_1$$

and still have an equation of the form in (12). It remains to show that

$$\|\boldsymbol{\delta}\|_1 = \|A\|_E$$

Note that since all quantities involved are nonnegative,

$$\|A\|_E = \left\| \sum_{p=1}^{r} \delta_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p \right\|_E = \sum_{p=1}^{r} \delta_p \|\mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p\|_E$$

By Equation (9), the RHS can be further simplified to

$$\sum_{p=1}^{r} \delta_p \|\mathbf{u}_p\|_1 \|\mathbf{v}_p\|_1 \cdots \|\mathbf{z}_p\|_1 = \sum_{p=1}^{r} \delta_p = \|\boldsymbol{\delta}\|_1$$

as required. ■

Note that the conditions on the vectors imply that they lie in unit simplices of various dimensions:

$$\mathbf{u}_1, \ldots, \mathbf{u}_r \in \Delta^{d_1-1}, \quad \mathbf{v}_1, \ldots, \mathbf{v}_r \in \Delta^{d_2-1}, \quad \ldots, \quad \mathbf{z}_1, \ldots, \mathbf{z}_r \in \Delta^{d_k-1} \tag{13}$$

For $k = 2$, the decomposition in (12) is best viewed as a parallel to the singular value decomposition of a matrix $A \in \mathbb{R}^{m\times n}$, which is in particular an expression of the form

$$A = \sum_{p=1}^{r} \sigma_p\, \mathbf{u}_p \otimes \mathbf{v}_p \tag{14}$$

where $r = \operatorname{rank}(A)$,

$$\|\boldsymbol{\sigma}\|_2 = \left[\sum_{p=1}^{r} |\sigma_p|^2\right]^{\frac{1}{2}} = \|A\|_F, \quad \text{and} \quad \|\mathbf{u}_p\|_2 = \|\mathbf{v}_p\|_2 = 1$$

for all $p = 1,\ldots,r$. Here $\boldsymbol{\sigma} = [\sigma_1,\ldots,\sigma_r]^\top \in \mathbb{R}^r$ is the vector of nonzero singular values of A. If A is normalized to have unit F-norm, then all quantities in Equation (14) may be viewed as living in unit spheres of various dimensions: $A \in \mathbb{S}^{mn-1}$, $\boldsymbol{\sigma} \in \mathbb{S}^{r-1}$, $\mathbf{u}_1,\ldots,\mathbf{u}_r \in \mathbb{S}^{m-1}$, $\mathbf{v}_1,\ldots,\mathbf{v}_r \in \mathbb{S}^{n-1}$ where $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^d$. For $k = 2$, the nonnegative matrix decomposition in Proposition 4.1 is one where the unit spheres are replaced by unit simplices and the l^2- and F-norms replaced by the l^1- and E-norms. An obvious departure from the case of SVD is that the vectors in Equation (13) are not orthogonal.

Henceforth when we use the terms NTF and NMF, we will mean a decomposition of the type in Proposition 4.1. For a nonnegative tensor with unit E-norm, $A \in \Delta^{d_1\cdots d_k-1}$, the decomposition in Proposition 4.1 has a probabilistic interpretation.

Let $U,V,\ldots,Z$ be discrete random variables and $q(u,v,\ldots,z) = \Pr(U = u,V = v,\ldots,Z = z)$ be their joint probability distribution. Suppose $U,V,\ldots,Z$ satisfy the *naïve Bayes hypothesis*, i.e., they are conditionally independent upon a single hidden discrete random variable Θ. Let $q_1(u \mid \theta),q_2(v \mid \theta),\ldots,q_k(z \mid \theta)$ denote, respectively the marginal probability distributions of $U,V,\ldots,Z$ conditional on the event $\Theta = \theta$. Then the probability distributions must satisfy the relation

$$q(u, v, \dots, z) = \sum_{\theta=1}^{r} \delta(\theta)\, q_1(u \mid \theta) q_2(v \mid \theta) \cdots q_k(z \mid \theta) \qquad (15)$$

where $\delta(\theta) = \Pr(\Theta = \theta)$. Since the discrete random variables $U, V, \dots, Z$ may take $d_1, d_2, \dots, d_k$ possible values, respectively, the Bayes rule in Equation (15) can be rewritten as the tensor decomposition in Equation (12), provided we "store" the marginal distributions $q_1(u \mid \theta), q_2(v \mid \theta), \dots, q_k(z \mid \theta)$ in the vectors $\mathbf{u}_\theta, \mathbf{v}_\theta, \dots, \mathbf{z}_\theta$, respectively. The requirement that $r = \operatorname{rank}_+(A)$ corresponds to the *Ockham principle of parsimony*: that the model (15) be the simplest possible, i.e., the hidden variable Θ be minimally supported.

For the case $k = 2$, Equation (15) is Hofmann's PLSI [22]. While it is known [24] that the multiplicative updating rule for NMF with KL divergence in Reference [17] is equivalent to the use of EM algorithm for maximum likelihood estimation of PLSI in Reference [22], this is about the equivalence of two algorithms (EM and multiplicative updating) applied to two approximation problems (maximum likelihood of PLSI and minimum KL divergence of NMF). Since the EM algorithm and the NMF multiplicative updating rules are first-order methods that can at best converge to a stationary point, saying that these two algorithms are equivalent for their respective approximation problems does not imply that the respective models are equivalent. The preceding paragraph states that the probabilistic relational models behind PLSI and NTF (and therefore NMF) are one and the same—a collection of random variables satisfying the naïve Bayes assumption with respect to a parsimonious hidden variable. This is a statement independent of approximation or computation.

5. NONEXISTENCE OF GLOBALLY OPTIMAL SOLUTION FOR REAL AND COMPLEX TENSOR APPROXIMATIONS

A major difficulty that one should be aware of is that the problem of finding a best rank-r approximation for tensors of order 3 or higher has no solution in general. There exists $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ such that

$$\inf \left\| A - \sum_{p=1}^{r} \lambda_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p \right\| \qquad (16)$$

is not attained by *any* choice of $\lambda_p, \mathbf{u}_p, \mathbf{v}_p, \dots, \mathbf{z}_p$, $p = 1, \dots, r$. It is also in general not possible to determine a priori if a given $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ will fail to have a best rank-r approximation. This problem is more widespread than one might imagine. It has been shown in Reference [21] that examples of this failure happens over a wide range of dimensions, orders, ranks, and for any continuous measure of proximity (thus including all norms and Brègman divergence). Moreover such failures can occur with positive probability and in some cases with certainty, i.e., where the infimum in Equation (16) is *never* attained. This phenomenon also extends to symmetric tensors [25].

This poses some serious conceptual difficulties—if one cannot guarantee a solution a priori, then what is one trying to compute in instances where there are no solutions? We often get the answer "an approximate solution." But how could one approximate a solution that does not even exist in the first place? Conceptual issues aside, this also causes computational difficulties in practice. Forcing a solution in finite precision for a problem that does not have a solution is an ill-advised strategy since a well-posed problem near to an ill-posed one is, by definition, ill-conditioned and therefore hard to compute. This ill-conditioning manifests itself in iterative algorithms as summands that grew unbounded in magnitude but with the peculiar property that the sum remains bounded. This was first observed by Bini, Lotti, and Romani [26] in the context of arbitrary precision approximations (where this phenomenon is desirable). Independently Harshman and his collaborators Kruskal and Lundy [4] also investigated this phenomenon, which they called PARAFAC degeneracy, from the perspective of model fitting (where it is undesirable). In Theorem 6.1, we will prove the cheerful fact that one does not need to worry about PARAFAC degeneracy when fitting a nonnegative PARAFAC model.

The first published account of an explicitly constructed example of PARAFAC degeneracy appeared in a study of the complexity of matrix multiplication by Bini, Capovani, Lotti, and Romani [3]. However their discussion was for a context entirely different from data analysis/model fitting and was presented in notations somewhat unusual. Until today, many remain unconvinced that the construction in Reference [3] indeed provides an explicit example of PARAFAC degeneracy and continue to credit the much later work of Paatero [27]. The truth is that such constructions are well-known in algebraic computational complexity; in addition to Reference [3], one may also find them in References [7,26,28], all predating [27]. As a small public service[§], we will translate the original construction of Bini, Capovani, Lotti, and Romani into notations more familiar to the technometrics communities.

In Reference [3], Bini, Capovani, Lotti, and Romani gave an algorithm that can approximate to arbitrary precision the product of two $n \times n$ matrices and requires only $O(n^{2.7799})$ scalar multiplications. The key to their construction is the following triplet of matrices which at first glance seem somewhat mysterious:

$$U = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & \varepsilon & \varepsilon \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} \varepsilon & 0 & 0 & -\varepsilon & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \varepsilon \\ 1 & -1 & 1 & 0 & 1 \end{bmatrix},$$

$$W = \begin{bmatrix} \varepsilon^{-1} & \varepsilon^{-1} & -\varepsilon^{-1} & \varepsilon^{-1} & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -\varepsilon^{-1} & 0 & \varepsilon^{-1} \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix}$$

We will show that these matrices may be used to construct a sequence of tensors exhibiting PARAFAC degeneracy.

We will assume that U has a 4th row of zeros and so U, V, $W \in \mathbb{R}^{4\times 5}$. As usual, u_{ij}, v_{ij}, w_{ij} will denote the (i,j)th entry of the respective matrices. Let $n \geq 4$ and $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4 \in \mathbb{R}^n$ (or $\mathbb{C}^n$) be

[§] And also to fulfill, belatedly, an overdued promise made to Harshman when he was preparing his bibliography on PARAFAC degeneracy.

linearly independent vectors. For $\varepsilon > 0$, define

$$A_\varepsilon := \sum_{j=1}^{4}\left[\left(\sum_{i=1}^{4} u_{ij}\mathbf{x}_i\right)\otimes\left(\sum_{i=1}^{4} v_{ij}\mathbf{x}_i\right)\otimes\left(\sum_{i=1}^{4} w_{ij}\mathbf{x}_i\right)\right]$$

Observe that

$$\begin{aligned} A_\varepsilon = {} & (\mathbf{x}_1+\mathbf{x}_3)\otimes(\varepsilon\mathbf{x}_1+\mathbf{x}_4)\otimes(\varepsilon^{-1}\mathbf{x}_1+\mathbf{x}_4)\\ & +\mathbf{x}_3\otimes(-\mathbf{x}_2-\mathbf{x}_4)\otimes\varepsilon^{-1}\mathbf{x}_1+\mathbf{x}_1\otimes\mathbf{x}_4\otimes(-\varepsilon^{-1}\mathbf{x}_1-\varepsilon^{-1}\mathbf{x}_3)\\ & +(\varepsilon\mathbf{x}_2+\mathbf{x}_3)\otimes(-\varepsilon\mathbf{x}_1+\mathbf{x}_2)\otimes(\varepsilon^{-1}\mathbf{x}_1+\mathbf{x}_2)\\ & +(\mathbf{x}_1+\varepsilon\mathbf{x}_2)\otimes(\varepsilon\mathbf{x}_3+\mathbf{x}_4)\otimes(\varepsilon^{-1}\mathbf{x}_3-\mathbf{x}_4)\end{aligned}$$

It is straight forward to verify that

$$\lim_{\varepsilon\to 0} A_\varepsilon = A$$

where

$$\begin{aligned} A = {} & \mathbf{x}_1\otimes\mathbf{x}_1\otimes\mathbf{x}_1+\mathbf{x}_1\otimes\mathbf{x}_3\otimes\mathbf{x}_3+\mathbf{x}_2\otimes\mathbf{x}_2\otimes\mathbf{x}_1+\mathbf{x}_2\otimes\mathbf{x}_4\otimes\mathbf{x}_3\\ & +\mathbf{x}_3\otimes\mathbf{x}_2\otimes\mathbf{x}_2+\mathbf{x}_3\otimes\mathbf{x}_4\otimes\mathbf{x}_4\end{aligned}$$

Note that the sequence A_ε exhibits PARAFAC degeneracy: as $\varepsilon\to 0$, each of the summands becomes unbounded in magnitude but A_ε remains bounded (and in fact converges to A).

Regardless of whether $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 \in \mathbb{R}^n$ or $\mathbb{C}^n$, it is clear that for all $\varepsilon > 0$,

$$\operatorname{rank}(A_\varepsilon)\le 5$$

Furthermore, one may show that $\operatorname{rank}(A) = 6$ over $\mathbb{C}$ and therefore $\operatorname{rank}(A) \ge 6$ over $\mathbb{R}$ (cf. remarks at the end of Section 3). In either case, A is an instance in $\mathbb{R}^{n\times n\times n}$ or $\mathbb{C}^{n\times n\times n}$ where the approximation problem in Equation (1) has no solution for $r = 5$—since $\inf_{\operatorname{rank}(X)\le 5}\|A-X\| = 0$ and $\operatorname{rank}(A)\ge 6$ together imply that

$$\operatorname*{argmin}_{\operatorname{rank}(X)\le 5}\|A-X\| = \emptyset$$

Hence the construction in Reference [3] also yields an explicit example of a best rank-r approximation problem (over $\mathbb{R}$ and $\mathbb{C}$) that has no solution.

6. EXISTENCE OF GLOBALLY OPTIMAL SOLUTION FOR NONNEGATIVE TENSOR APPROXIMATIONS

As we have mentioned in Section 2, nonnegativity constraints are often natural in the use of PARAFAC. Empirical evidence from Bro's chemometrics studies revealed that PARAFAC degeneracy was never observed when fitting nonnegative-valued data with a nonnegative PARAFAC model. This then led Harshman to conjecture that this is always the case. The text of his e-mail had been reproduced in Reference [19].

The conjectured result involves demonstrating the existence of global minimum over a non-compact feasible region and is thus not immediate. Nevertheless the proof is still straightforward by the following observation: If a continuous real-valued function has a non-empty compact sublevel set, then it has to attain its infimum—a consequence of the extreme value theorem. This is essentially what we will show in the following proof for the nonnegative PARAFAC loss function (in fact, we will show that all sublevel sets of the function are compact). We will use the E-norm in our proof for simplicity, the result for other norms then follows from the equivalence of all norms on finite-dimensional spaces. Essentially the same proof, but in terms of the more customary F-norm, appeared in Reference [19]. We will follow the notations in Section 4.

Theorem 6.1. *Let $A\in\mathbb{R}^{d_1\times\cdots\times d_k}$ be nonnegative. Then*

$$\inf\left\{\left\|A-\sum_{p=1}^{r}\delta_p\,\mathbf{u}_p\otimes\mathbf{v}_p\otimes\cdots\otimes\mathbf{z}_p\right\|_E \;\middle|\; \delta\in\mathbb{R}^r_+,\ \mathbf{u}_p\in\Delta^{d_1-1},\ldots,\mathbf{z}_p\in\Delta^{d_k-1},\ p=1,\ldots,r\right\}$$

is attained.

Proof. Recall that $\mathbb{R}^n_+ = \{\mathbf{x}\in\mathbb{R}^n \mid \mathbf{x}\ge 0\}$ and $\Delta^{n-1} = \{\mathbf{x}\in\mathbb{R}^n_+ \mid \|\mathbf{x}\|_1 = 1\}$. We define the function $f:\mathbb{R}^r\times(\mathbb{R}^{d_1}\times\cdots\times\mathbb{R}^{d_k})^r\to\mathbb{R}$ by

$$f(T) := \left\|A-\sum_{p=1}^{r}\delta_p\,\mathbf{u}_p\otimes\mathbf{v}_p\otimes\cdots\otimes\mathbf{z}_p\right\|_E \qquad (17)$$

where we let $T = (\delta_1,\ldots,\delta_r;\mathbf{u}_1,\mathbf{v}_1,\ldots,\mathbf{z}_1;\cdots;\mathbf{u}_r,\mathbf{v}_r,\ldots,\mathbf{z}_r)$ denote the argument of f. Let $\mathcal{D}$ be the following subset of $\mathbb{R}^r\times(\mathbb{R}^{d_1}\times\cdots\times\mathbb{R}^{d_k})^r = \mathbb{R}^{r(1+d_1+\cdots+d_k)}$,

$$\mathcal{D} := \mathbb{R}^r_+\times\left(\Delta^{d_1-1}\times\cdots\times\Delta^{d_k-1}\right)^r$$

Note that $\mathcal{D}$ is closed but unbounded. Let the infimum in question be $\mu := \inf\{f(T)\mid T\in\mathcal{D}\}$. We will show that the sublevel set of f restricted to $\mathcal{D}$,

$$\mathcal{E}_\alpha = \{T\in\mathcal{D}\mid f(T)\le\alpha\}$$

is compact for all $\alpha > \mu$ and thus the infimum of f on $\mathcal{D}$ must be attained. The set $\mathcal{E}_\alpha = \mathcal{D}\cap f^{-1}(-\infty,\alpha]$ is closed since f is continuous (by the continuity of norm). It remains to show that $\mathcal{E}_\alpha$ is bounded. Suppose the contrary. Then there exists a sequence $(T_n)_{n=1}^{\infty}\subset\mathcal{D}$ with $\|T_n\|_1\to\infty$ but $f(T_n)\le\alpha$ for all n. Clearly, $\|T_n\|_1\to\infty$ implies that $\delta_q^{(n)}\to\infty$ for at least one $q\in\{1,\ldots,r\}$. Note that

$$f(T)\ge\left|\,\|A\|_E-\left\|\sum_{p=1}^{r}\delta_p\,\mathbf{u}_p\otimes\mathbf{v}_p\otimes\cdots\otimes\mathbf{z}_p\right\|_E\right|$$

Since all terms involved in the approximant are nonnegative, we have

$$\left\|\sum_{p=1}^{r}\delta_p\,\mathbf{u}_p\otimes\mathbf{v}_p\otimes\cdots\otimes\mathbf{z}_p\right\|_E = \sum_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k}\sum_{p=1}^{r}\delta_p u_{pj_1}v_{pj_2}\cdots z_{pj_k}$$

$$
\begin{aligned}
&\geq \sum_{j_1,\dots,j_k=1}^{d_1,\dots,d_k} \delta_q u_{qj_1} v_{qj_2}\cdots z_{qj_k}\\
&= \delta_q \sum_{j_1,\dots,j_k=1}^{d_1,\dots,d_k} u_{qj_1} v_{qj_2}\cdots z_{qj_k}\\
&= \delta_q \|\mathbf{u}_q \otimes \mathbf{v}_q \otimes \cdots \otimes \mathbf{z}_q\|_E\\
&= \delta_q \|\mathbf{u}_q\|_1 \|\mathbf{v}_q\|_1 \cdots \|\mathbf{z}_q\|_1\\
&= \delta_q
\end{aligned}
$$

where the last two equalities follow from Equation (9) and $\|\mathbf{u}_q\|_1 = \|\mathbf{v}_q\|_1 = \cdots = \|\mathbf{z}_q\|_1 = 1$. Hence, as $\delta_q^{(n)} \to \infty$, $f(T_n) \to \infty$—contradicting the assumption that $f(T_n) \leq \alpha$ for all n. ■

The proof essentially shows that the function f is *coercive*—a real-valued function f is said to be coercive if $\lim_{\|\mathbf{x}\|\to+\infty} f(\mathbf{x}) = +\infty$ [29]. This is a standard condition often used to guarantee that a continuous function on a noncompact domain attains its global minimum and is equivalent to saying that f has bounded sublevel sets. A minor point to note is that had we instead optimized over a sum of rank-1 terms $\mathbf{u} \otimes \mathbf{v} \otimes \cdots \otimes \mathbf{z}$, the proof would fail because the vectors $\mathbf{u}, \mathbf{v}, \dots, \mathbf{z}$ may be scaled by non-zero positive scalars that product to 1, i.e.,

$$\alpha\mathbf{u} \otimes \beta\mathbf{v} \otimes \cdots \otimes \zeta\mathbf{z} = \mathbf{u} \otimes \mathbf{v} \otimes \cdots \otimes \mathbf{z}, \qquad \alpha\beta\cdots\zeta = 1$$

So for example $(n\mathbf{x}) \otimes \mathbf{y} \otimes (\mathbf{z}/n)$ can have diverging loading factors even while the outer-product remains fixed. We avoided this by requiring that $\mathbf{u}, \mathbf{v}, \dots, \mathbf{z}$ be unit vectors and having a δ that records the magnitude.

The following proposition provides four useful characterizations of the statement that the function $\operatorname{rank}_+ : \mathbb{R}_+^{d_1\times\cdots\times d_k} \to \mathbb{R}$ is upper semicontinuous. This is the nonnegative rank equivalent of a similar result in Reference [21].

Proposition 6.2. *Let $r, k \in \mathbb{N}$ and let the topology on $\mathbb{R}_+^{d_1\times\cdots\times d_k}$ be induced by the E-norm. The following statements are equivalent; and since the last statement is true by Theorem 6.1, so are the others.*

(a) *The set $\mathcal{S}_r := \{X \in \mathbb{R}_+^{d_1\times\cdots\times d_k} \mid \operatorname{rank}_+(X) \leq r\}$ is closed.*

(b) *Every $A \in \mathbb{R}_+^{d_1\times\cdots\times d_k}$, $\operatorname{rank}_+(A) > r$, has a best nonnegative rank-r approximation, i.e.,*

$$\inf\{\|A - X\|_E \mid \operatorname{rank}_+(X) \leq r\}$$

is attained (by some X_r with $\operatorname{rank}_+(X_r) \leq r$).

(c) *No $A \in \mathbb{R}_+^{d_1\times\cdots\times d_k}$, $\operatorname{rank}_+(A) > r$, can be approximated arbitrarily closely by nonnegative tensors of strictly lower nonnegative rank, i.e.,*

$$\inf\{\|A - X\|_E \mid \operatorname{rank}_+(X) \leq r\} > 0$$

(d) *No sequence $(X_n)_{n=1}^{\infty} \subset \mathbb{R}_+^{d_1\times\cdots\times d_k}$, $\operatorname{rank}_+(X_n) \leq r$, can converge to $A \in \mathbb{R}_+^{d_1\times\cdots\times d_k}$ with $\operatorname{rank}_+(A) > r$.*

Proof. (a) ⇒ (b): Suppose $\mathcal{S}_r$ is closed. Since the set $\{X \in \mathbb{R}_+^{d_1\times\cdots\times d_k} \mid \|A - X\| \leq \|A\|\}$ intersects $\mathcal{S}_r$ non-trivially (e.g., 0 is in both sets). Their intersection $\mathcal{T} = \{X \in \mathbb{R}_+^{d_1\times\cdots\times d_k} \mid \operatorname{rank}_+(X) \leq r, \|A - X\| \leq \|A\|\}$ is a non-empty compact set. Now observe that

$$\delta := \inf\{\|A - X\| \mid X \in \mathcal{S}_r\} = \inf\{\|A - X\| \mid X \in \mathcal{T}\}$$

since any $X' \in \mathcal{S}_r \backslash \mathcal{T}$ must have $\|A - X'\| > \|A\|$ while we know that $\delta \leq \|A\|$. By the compactness of $\mathcal{T}$, there exists $X_* \in \mathcal{T}$ such that $\|A - X_*\| = \delta$. So the required infimum is attained by $X_* \in \mathcal{T} \subset \mathcal{S}$. The remaining implications (b) ⇒ (c) ⇒ (d) ⇒ (a) are obvious. ■

In the language of References [7,26], Proposition 6.2 implies that "nonnegative border rank" coincides with nonnegative rank. An immediate corollary is that the *E*-norm in Theorem 6.1 and Proposition 6.2 may be replaced by any other norm. In fact we will see later that we may replace norms with more general measures of proximity.

Corollary 6.3. *Let $A = [\![a_{j_1\cdots j_k}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_k}$ be nonnegative and $\|\cdot\| : \mathbb{R}^{d_1\times\cdots\times d_k} \to [0,\infty)$ be an arbitrary norm. Then*

$$\inf\left\{\left\|A - \sum_{p=1}^{r} \delta_p\, \mathbf{u}_p \otimes \mathbf{v}_p \otimes \cdots \otimes \mathbf{z}_p\right\| \;\middle|\; \delta \in \mathbb{R}_+^r, \mathbf{u}_p \in \Delta^{d_1-1}, \dots, \mathbf{z}_p \in \Delta^{d_k-1}, p = 1, \dots, r\right\}$$

is attained.

Proof This simply follows from the fact that all norms on finite dimensional spaces are equivalent and therefore induce the same topology on $\mathbb{R}_+^{d_1\times\cdots\times d_k}$. So Proposition 6.2 holds for any norms. In particular, the statement (b) in Proposition 6.2 for an arbitrary norm $\|\cdot\|$ is exactly the result desired here. ■

Corollary 6.3 implies that the PARAFAC degeneracy discussed in Section 5 does not happen for nonnegative approximations of nonnegative tensors. There is often a simplistic view of PARAFAC degeneracy as being synonymous to "between component cancellation" and thus cannot happen for nonnegative tensor approximation since it is "purely additive with no cancellation between parts" [17,18]. While it provides an approximate intuitive picture, this point of view is flawed since PARAFAC degeneracy is not the same as "between component cancellation." There is cancellation in

$$n\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} - \left(n + \frac{1}{n}\right)\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} \tag{18}$$

but the sequence exhibits no PARAFAC degeneracy. Conversely, the sequence of nonnegative tensors

$$A_n = \left[\begin{array}{cc|cc} 0 & 1 & 1 & 1/n \\ 1 & 1/n & 1/n & 1/n^2 \end{array}\right] \in \mathbb{R}^{2\times2\times2}$$

may each be decomposed nonnegatively as

$$A_n = A + \frac{1}{n}B + \frac{1}{n^2}C \tag{19}$$

with $A, B, C \in \mathbb{R}^{2\times2\times2}$ given by

$$A = \left[\begin{array}{cc|cc} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right], \quad B = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right],$$
$$C = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

and each may in turn be decomposed into a sum of rank-1 terms. While there is no "between component cancellation" among these rank-1 summands, it is known [21] that the convergence

$$\lim_{n\to\infty} A_n = A$$

exhibits PARAFAC degeneracy over $\mathbb{R}^{2\times2\times2}$ or $\mathbb{C}^{2\times2\times2}$, where there are decompositions of A_n different from the one given in Equation (19) exhibiting PARAFAC degeneracy. That such decompositions cannot happen over $\mathbb{R}_+^{2\times2\times2}$ is precisely the statement of Proposition 6.2, which we proved by way of Theorem 6.1.

As one may surmise from the discussion above, the popular view of PARAFAC degeneracy in technometrics as "cancellation between diverging components in a way that the sum stays bounded" is mathematically imprecise and includes many unintended non-examples like (18). A more careful formulation refining the "between component cancellation" view may be found in Reference [30]. In Reference [21, Proposition 4.1], several mathematically precise formulations of PARAFAC degeneracy were given as equivalent statements about the insolvability of the best rank-r arpproximation problem, with the "diverging components" phenomenon deduced as consequences [21, Propositions 4.8 and 4.9].

7. BRÈGMAN DIVERGENCES

In many applications, a norm may not be the most suitable measure of proximity. Other measures based on entropy, margin, spectral separation, volume, etc, are often used as loss functions in matrix and tensor approximations. Such measures may not even be a metric, an example being the Brègman divergence [31–33], a class of proximity measures that often have information theoretic or probabilistic interpretations. In the definition below, $\mathrm{ri}(\Omega)$ denotes the relative interior of Ω, i.e., the interior of Ω regarded as a subset of its affine hull; $\|\cdot\|$ is any arbitrary norm on $\mathbb{R}^{d_1\times\cdots\times d_k}$—again the choice of which is immaterial since all norms induce the same topology on Ω.

Definition 7.1. *Let $\emptyset \neq \Omega \subseteq \mathbb{R}^{d_1\times\cdots\times d_k}$ be a closed convex set. Let $\varphi : \Omega \to \mathbb{R}$ be continuously differentiable on $\mathrm{ri}(\Omega)$ and strictly convex and continuous on Ω. The function $D_\varphi : \Omega \times \mathrm{ri}(\Omega) \to \mathbb{R}$ defined by*

$$D_\varphi(A, B) = \varphi(A) - \varphi(B) - \langle \nabla\varphi(B), A - B\rangle$$

*is a **Brègman divergence** if*

(i) *For any fixed $A \in \Omega$, the sublevel set*
$$\mathcal{L}_\alpha(A) = \{X \in \mathrm{ri}(\Omega) \mid D_\varphi(A, X) \leq \alpha\}$$
is bounded for all $\alpha \in \mathbb{R}$.

(ii) *Let $(X_n)_{n=1}^{\infty} \subset \mathrm{ri}(\Omega)$ and $A \in \Omega$. If*
$$\lim_{n\to\infty} \|A - X_n\| = 0$$
then
$$\lim_{n\to\infty} D_\varphi(A, X_n) = 0$$

(iii) *Let $(X_n)_{n=1}^{\infty} \subset \mathrm{ri}(\Omega)$, $A \in \Omega$, and $(A_n)_{n=1}^{\infty} \subset \Omega$. If*
$$\lim_{n\to\infty} \|A - X_n\| = 0, \quad \limsup_{n\to\infty} \|A_n\| < \infty, \quad \lim_{n\to\infty} D_\varphi(A_n, X_n) = 0$$
then
$$\lim_{n\to\infty} \|A_n - X_n\| = 0$$

Note that $D_\varphi(A,B) \geq 0$ and that $D_\varphi(A,B) = 0$ iff $A = B$ by the strict convexity of φ. However D_φ need not satisfy the triangle inequality nor must it be symmetric in its two arguments. So a Brègman divergence is not a metric in general.

Brègman divergences are particularly important in nonnegative matrix and tensor decompositions [17,18]. In fact, one of the main novelty of NMF as introduced by Lee and Seung [17] over the earlier studies in technometrics [13–16] is their use of the *Kullback-Leibler divergence* [34] as a proximity measure[‖]. The KL divergence is defined for nonnegative matrices in Reference [17] but it is straightforward to extend the definition to nonnegative tensors. For $A \in \mathbb{R}_+^{d_1\times\cdots\times d_k}$ and $B \in \mathrm{ri}(\mathbb{R}_+^{d_1\times\cdots\times d_k})$, this is

$$D_{\mathrm{KL}}(A, B) = \sum_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} \left[a_{j_1\cdots j_k} \log\left(\frac{a_{j_1\cdots j_k}}{b_{j_1\cdots j_k}}\right) - a_{j_1\cdots j_k} + b_{j_1\cdots j_k} \right]$$

where $0 \log 0$ is taken to be 0, the limiting value. It comes from the following choice of φ,

$$\varphi_{\mathrm{KL}}(A) = \sum_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} a_{j_1\cdots j_k} \log a_{j_1\cdots j_k}$$

We note that Kullback and Liebler's original definition [34] was in terms of probability distributions. The version that we introduced here is a slight generalization. When A and B are probability distributions as in Section 4, then $\|A\|_E = \|B\|_E = 1$ and our definition reduces to the original one in Reference [34],

$$D_{\mathrm{KL}}(A, B) = \sum_{j_1,\ldots,j_k=1}^{d_1,\ldots,d_k} a_{j_1\cdots j_k} \log\left(\frac{a_{j_1\cdots j_k}}{b_{j_1\cdots j_k}}\right)$$

In this case $D_{\mathrm{KL}}(A,B)$ may also be interpreted as the relative entropy of the respective distributions.

[‖] This brought back memories of the many intense e-mail exchanges with Harshman, of which one was about the novelty of NMF. His fervently argued messages will be missed.

It is natural to ask if the following analogous nonnegative tensor approximation problem for Brègman divergence will always have a solution:

$$X_r \in \operatorname{argmin}\{D_\varphi(A, X) \mid X \in \operatorname{ri}(\Omega), \operatorname{rank}_+(X) \le r\} \tag{20}$$

Clearly, the problem cannot be expected to have a solution in general since $\operatorname{ri}(\Omega)$ is not closed. For example let $A = \mathbf{e} \otimes \mathbf{e} \otimes \mathbf{e} \in \mathbb{R}_+^{2\times2\times2}$ where $\mathbf{e} = [1,0]^\top$, then

$$\inf\{D_{\mathrm{KL}}(A, X) \mid X \in \operatorname{ri}(\mathbb{R}_+^{2\times2\times2}), \operatorname{rank}_+(X) \le 1\} = 0$$

cannot be attained by any $\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} \in \operatorname{ri}(\mathbb{R}_+^{2\times2\times2})$ since if we set $\mathbf{x}_n = \mathbf{y}_n = \mathbf{z}_n = [1,n^{-1}]^\top$, then as $n \to \infty$,

$$D_{\mathrm{KL}}\left(A, \mathbf{x}_n \otimes \mathbf{y}_n \otimes \mathbf{z}_n\right) = \frac{1}{n^3} \to 0$$

This is simply a consequence of the way a Brègman divergence is defined and has nothing to do with any peculiarities of tensor rank, unlike the example discussed in Section 5. This difficulty may be avoided by posing the problem for any closed (but not necessarily compact) subset of $\operatorname{ri}(\Omega)$.

Proposition 7.2. *Let Ω be a closed convex subset of $\mathbb{R}_+^{d_1\times\cdots\times d_k}$ and $A \in \Omega$. Let $D_\varphi : \Omega \times \operatorname{ri}(\Omega) \to \mathbb{R}$ be a Brègman divergence. Then*

$$\inf\{D_\varphi(A, X) \mid X \in K, \operatorname{rank}_+(X) \le r\} \tag{21}$$

is attained for any closed subset $K \subseteq \operatorname{ri}(\Omega)$.

Proof. Recall that $\mathcal{S}_r := \{X \in \mathbb{R}_+^{d_1\times\cdots\times d_k} \mid \operatorname{rank}_+(X) \le r\}$. The statement is trivial if $\operatorname{rank}_+(A) \le r$. So we will also assume that $\operatorname{rank}_+(A) \ge r + 1$. Let μ be the infimum in Equation (21) and let $\alpha > \mu$. By (i) in Definition 7.1, the sublevel set $\mathcal{L}_\alpha(A)$ is bounded and so its subset

$$K \cap \mathcal{S}_r \cap \mathcal{L}_\alpha(A) = \{X \in K \cap \mathcal{S}_r \mid D_\varphi(A, X) \le \alpha\}$$

must also be bounded. Note that $K \cap \mathcal{S}_r$ is closed. Since φ is continuously differentiable on $\operatorname{ri}(\Omega)$, the function $X \mapsto D_\varphi(A,X)$ is continuous and so $K \cap \mathcal{S}_r \cap \mathcal{L}_\alpha(A)$ is also closed. Hence $D_\varphi(A,X)$ must attain μ on the compact set $K \cap \mathcal{S}_r \cap \mathcal{L}_\alpha(A)$. ■

As one can see from the proof, Proposition 7.2 extends to any other measure of proximity $d(A, X)$ where the function $X \mapsto d(A, X)$ is continuous and coercive. Of course this is just a restatement of the problem, the bulk of the work involved is usually to show that the proximity function in question has those required properties.

8. ASIDE: NORM-REGULARIZED AND ORTHOGONAL APPROXIMATIONS

We have often been asked about norm-regularized and orthogonal approximations of tensors that are not necessarily nonnegative. These approximation problems are useful in practice [27,35,36]. Nevertheless these always have optimal solutions for a much simpler reason—they are continuous optimization problems over compact feasible set, so the existence of global minima is immediate from the extreme value theorem (note that this is not the case for nonnegative tensor approximation). In the following, we will let $A \in \mathbb{R}^{d_1\times\cdots\times d_r}$, not necessarily nonnegative.

Recall that $\mathrm{O}(n,r)$, the set of $n \times r$ matrices ($r \le n$) with orthonormal columns, is compact in $\mathbb{R}^{n\times r}$. If we impose orthonormality constraints on the normalized loading factors in Equation (2), i.e., $[\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathrm{O}(d_1,r), \ldots, [\mathbf{z}_1, \ldots, \mathbf{z}_r] \in \mathrm{O}(d_k, r)$, then it follows that $|\lambda_p| \le \|A\|_F^2$ for $p = 1, \ldots, r$, i.e., $\boldsymbol{\lambda} \in [-\|A\|_F, \|A\|_F]^r \subset \mathbb{R}^r$. Since the PARAFAC objective is continuous and we are effectively minimizing over the compact feasible region

$$[-\|A\|_F, \|A\|_F]^r \times \mathrm{O}(d_1, r) \times \cdots \times \mathrm{O}(d_k, r)$$

this shows that orthogonal PARAFAC always has a globally optimal solution.

Next, the regularization proposed in Reference [27] is to add to the PARAFAC objective terms proportional to the 2-norm of each loading factor, i.e.,

$$\left\| A - \sum_{p=1}^{r} \mathbf{a}_p \otimes \mathbf{b}_p \otimes \cdots \otimes \mathbf{c}_p \right\|_F^2 + \rho \sum_{p=1}^{r} \left(\|\mathbf{a}_p\|_2^2 + \|\mathbf{b}_p\|_2^2 + \cdots + \|\mathbf{c}_p\|_2^2 \right) \tag{22}$$

From constrained optimization theory, we know that, under some regularity conditions, minimizing a continuous function $f(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ under constraints $\|\mathbf{x}_i\|_2 = r_i$, $i = 1, \ldots, k$, is *equivalent* to minimizing the functional $f(\mathbf{x}_1, \ldots, \mathbf{x}_k) + \sum_{i=1}^{k} \rho_i \|\mathbf{x}_i\|_2^2$ for appropriate $\rho_1, \ldots, \rho_k \in \mathbb{R}$. In a finite dimensional space, the sphere of radius r_i is compact, and so is the feasible set defined by $\|\mathbf{x}_i\|_2 = r_i$, $i = 1, \ldots, k$, and thus f must attain its extrema. In the same vein, Equation (22) is equivalent to an equality constrained optimization problem and so norm-regularized PARAFAC always has a globally optimal solution. This approach may also be applied to regularizations other than the one discussed here.

Acknowledgements

We thank the reviewers for their many helpful comments. This work has been partially supported by contract ANR-06-BLAN-0074 'Decotes'.

REFERENCES

1. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an explanatory multi-modal factor analysis. *UCLA Working Papers in Phonetics* 1970; **16**: 1–84.
2. Deerwester S, Dumais S, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.* 1990; **41**(6): 391–407.
3. Bini D, Capovani M, Lotti G, Romani F. $O(n^{2.7799})$ complexity for $n \times n$ approximate matrix multiplication. *Inform. Process. Lett.* 1979; **8**(5): 234–235.
4. Kruskal JB, Harshman RA, Lundy ME. How 3-MFA data can cause degenerate PARAFAC solutions, among other relationships, pp. 115–122, in Reference [37].
5. Hitchcock FL. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.* 1927; **6**(1): 164–189.
6. Hitchcock FL. Multiple invariants and generalized rank of a *p*-way matrix or tensor. *J. Math. Phys.* 1927; **7**(1): 39–79.

7. Bürgisser P, Clausen M, Shokrollahi MA. *Algebraic Complexity Theory (Grundlehren der mathematischen Wissenschaften)*, vol. 315. Springer-Verlag: Berlin, Germany, 1996.
8. Strassen V. Gaussian elimination is not optimal. *Numer. Math.* 1969; **13**: 354–356.
9. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via *n*-way generalization of Eckart–Young decomposition. *Psychometrika* 1970; **35**(3): 283–319.
10. Eckart C, Young G. The approximation of one matrix by another of lower rank. *Psychometrika* 1936; **1**(3): 211–218.
11. Bro R, de Jong S. A fast non-negativity constrained least squares algorithm. *J. Chemometrics* 1997; **11**(5): 393–401.
12. Bro R, Sidiropoulos N. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics* 1998; **12** (4): 223–247.
13. Carroll JD, De Soete G, Pruzansky S. Fitting of the latent class model via iteratively reweighted least squares CANDECOMP with nonnegativity constraints, pp. 463–472, in Reference [37].
14. Krijnen WP, Ten Berge JMF. Contrastvrije oplossingen van het CANDECOMP/PARAFAC-model. *Kwantitatieve Methoden* 1991; **12**(37): 87–96.
15. Paatero P. A weighted non-negative least squares algorithm for three-way PARAFAC factor analysis. *Chemometrics Intell. Lab. Syst.* 1997; **38**(2): 223–242.
16. Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 1994; **5**(2): 111–126.
17. Lee DD, Seung HS. Learning the parts of objects by nonnegative matrix factorization. *Nature* 1999; **401**: 788–791.
18. Shashua A, Hazan T. Non-negative tensor factorization with applications to statistics and computer vision. *Proc. Int. Conf. Mach. Learn. (ICML'05)* 2005; **22**: 792–799.
19. Lim L-H. Optimal solutions to nonnegative PARAFAC /multilinear NMF always exist. *Workshop on Tensor Decompositions and Applications*, Centre International de rencontres Mathématiques, Luminy, France, August 29–September 2, 2005.
20. Gelfand IM, Kapranov MM, Zelevinsky AV. *Discriminants, Resultants, and Multidimensional Determinants*. Birkhäuser Publishing: Boston, MA, 1994.
21. de Silva V, Lim L-H. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.* 2008; **30**(3): 1084–1127.
22. Hofmann T. Probabilistic Latent Semantic Indexing. *Proc. Annual Int. SIGIR Conf. Res. Develop. Inform. Retrieval (SIGIR'99)* 1999; **22**: 50–57.
23. Garcia LD, Stillman M, Sturmfels B. Algebraic geometry of Bayesian networks. *J. Symbolic Comput.* 2005; **39**(3–4): 331–355.
24. Gaussier E, Goutte C. Relation between PLSA and NMF and implications. *Proc. Annual Int. SIGIR Conf. Res. Develop. Inform. Retrieval (SIGIR'05)* 2005; **28**: 601–602.
25. Comon P, Golub G, Lim L-H, Mourrain B. Symmetric tensors and symmetric tensor rank. *SIAM J. Matrix Anal. Appl.* 2008; **30**(3): 1254–1279.
26. Bini D, Lotti G, Romani F. Approximate solutions for the bilinear form computational problem. *SIAM J. Comput.* 1980; **9**(4): 692–697.
27. Paatero P. Construction and analysis of degenerate PARAFAC models. *J. Chemometrics* 2000; **14**(1): 285–299.
28. Knuth DE. *The Art of Computer Programming: Seminumerical Algorithms* (3rd edn), vol. 2. Addision Wesley: Reading, MA, 1998.
29. Brinkhuis J, Tikhomirov V. *Optimization: Insights and Applications*. Princeton University Press: Princeton, NJ, 2005.
30. Stegeman A. Low-rank approximation of generic $p \times q \times 2$ arrays and diverging components in the CANDECOMP/PARAFAC model. *SIAM J. Matrix Anal. Appl.* 2008; **30**(3): 988–1007.
31. Brègman L. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.* 1967; **7**(3): 620–631.
32. Dhillon IS, Tropp JA. Matrix nearness problems using Brègman divergences. *SIAM J. Matrix Anal. Appl.* 2007; **29**(4): 1120–1146.
33. Iusem AN. "Bregman distance" and "Bregman function". In *Encyclopaedia of Mathematics*, Hazewinkel M (ed.). Kluwer: Dordrecht, Netherlands, 1997; 152–154.
34. Kullback S, Leibler RA. On information and sufficiency. *Ann. Math. Statistics* 1951; **22**(1): 79–86.
35. Comon P. Independent component analysis: a new concept? *Signal Process.* 1994; **36**(3): 287–314.
36. Harshman RA Lundy ME. Data preprocessing and the extended PARAFAC model. In *Research Methods for Multimode Data Analysis*, Law HG, Snyder CW, Hattie JA, McDonald RP (eds). Praeger: New York, 1984; 216–281.
37. Coppi R, Bolasco S (eds). *Multiway Data Analysis*. Elsevier Science: Amsterdam, Netherlands, 1989.