

# Rank regularization and Bayesian inference for tensor completion and extrapolation<sup>†</sup>

Juan Andrés Bazerque, Gonzalo Mateos, and Georgios B. Giannakis (contact author)\*

**Abstract**—A novel regularizer of the PARAFAC decomposition factors capturing the tensor's rank is proposed in this paper, as the key enabler for completion of three-way data arrays with missing entries. Set in a Bayesian framework, the tensor completion method incorporates prior information to enhance its smoothing and prediction capabilities. This probabilistic approach can naturally accommodate general models for the data distribution, lending itself to various fitting criteria that yield optimum estimates in the maximum-a-posteriori sense. In particular, two algorithms are devised for Gaussian- and Poisson-distributed data, that minimize the rank-regularized least-squares error and Kullback-Leibler divergence, respectively. The proposed technique is able to recover the “ground-truth” tensor rank when tested on synthetic data, and to complete brain imaging and yeast gene expression datasets with 50% and 15% of missing entries respectively, resulting in recovery errors at  $-10\text{dB}$  and  $-15\text{dB}$ .

**Index Terms**—Tensor, low-rank, missing data, Bayesian inference, Poisson process.

## I. INTRODUCTION

Imputation of missing data is a basic task arising in various Big Data applications as diverse as medical imaging [12], bioinformatics [3], as well as social and computer networking [10], [17]. The key idea rendering recovery feasible is the “regularity” present among missing and available data. Low rank is an attribute capturing this regularity, and can be readily exploited when data are organized in a matrix. A natural approach to *low-rank matrix completion* problem is minimizing the rank of a target matrix, subject to a constraint on the error in fitting the observed entries [4]. Since rank minimization is generally NP-hard [26], the nuclear norm has been advocated recently as a convex surrogate to the rank [11]. Beyond tractability, nuclear-norm minimization enjoys good performance both in theory as well as in practice [4].

The goal of this paper is imputation of missing entries of tensors (also known as multi-way arrays), which are high-order generalizations of matrices frequently encountered in chemometrics, medical imaging, and networking [8], [16]. Leveraging the low-rank structure for tensor completion is challenging, since even computing the tensor rank is NP-hard [14]. Defining a nuclear norm surrogate is not obvious either, since singular values as defined by the Tucker decomposition are not generally related with the rank. Traditional approaches to finding low-dimensional representations of tensors include unfolding the multi-way data and applying matrix factorizations such

as the singular-value decomposition (SVD) [3], [7], [25] or, employing the parallel factor (PARAFAC) decomposition [9], [24]. In the context of tensor completion, an approach falling under the first category can be found in [12], while imputation using PARAFAC was dealt with in [2].

The imputation approach presented in this paper builds on a novel regularizer accounting for the tensor rank, that relies on redefining the matrix nuclear norm in terms of its low-rank factors. The contribution is two-fold. First, it is established that the low-rank inducing property of the regularizer carries over to tensors by promoting sparsity in the factors of the tensor's PARAFAC decomposition. In passing, this analysis allows for drawing a neat connection with the atomic-norm in [5]. The second contribution is the incorporation of prior information, with a Bayesian approach that endows tensor completion with extra smoothing and prediction capabilities. A parallel analysis in the context of reproducing kernel Hilbert spaces (RKHS) further explains these acquired capabilities, provides an alternative means of obtaining the prior information, and establishes a useful connection with collaborative filtering approaches [1] when reduced to the matrix case.

While least-squares (LS) is typically utilized as the fitting criterion for matrix and tensor completion, implicitly assuming Gaussian data, the adopted probabilistic framework supports the incorporation of alternative data models. Targeting count processes available in the form of network traffic data, genome sequencing, and social media interactions, which are modeled as Poisson distributed, the maximum a posteriori (MAP) estimator is expressed in terms of the Kullback-Leibler (K-L) divergence [10].

The remainder of the paper is organized as follows. Section II offers the necessary background on nuclear-norm regularization for matrices, the PARAFAC decomposition, and the definition of tensor rank. Section III presents the tensor completion problem, establishing the low-rank inducing property of the proposed regularization. Prior information is incorporated in Section IV, with Bayesian and RKHS formulations of the tensor imputation method, leading to the low-rank tensor-imputation (LRTI) algorithm. Section V develops the method for Poisson tensor data, and redesigns the algorithm to minimize the rank-regularized K-L divergence. Finally, Section VI presents numerical tests carried out on synthetic and real data, including expression levels in yeast, and brain magnetic resonance images (MRI). Conclusions are drawn in Section VII, while most technical details are deferred to the Appendix.

The notation adopted throughout includes bold lowercase and capital letters for vectors  $\mathbf{a}$  and matrices  $\mathbf{A}$ , respectively, with superscript  $T$  denoting transposition. Tensors are underlined as e.g.,  $\underline{\mathbf{X}}$ , and their slices carry a subscript as in  $\mathbf{X}_p$ ; see also Fig. 1. Both the matrix and tensor Frobenius norms

<sup>†</sup> This work was supported by MURI (AFOSR FA9550-10-1-0567) grant. Part of the paper appeared in the *Proc. of IEEE Workshop on Statistical Signal Processing*, Ann Arbor, USA, August 5-8, 2012.

\* The authors are with the Dept. of ECE and the Digital Technology Center, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455. Tel/fax: (612)626-7781/625-2002; Emails: {bazer002, mate0058, georgios}@umn.edu

are represented by  $\|\cdot\|_F$ . Symbols  $\otimes$ ,  $\odot$ ,  $\circledast$ , and  $\circ$ , denote the Kronecker, Kathri-Rao, Hadamard (entry-wise), and outer product, respectively.

## II. PRELIMINARIES

### A. Nuclear-norm minimization for matrix completion

Low-rank approximation is a popular method for estimating missing values of a matrix  $\mathbf{Z} \in \mathbb{R}^{N \times M}$ , which capitalizes on “regularities” across the data [11]. For the imputation to be feasible, a binding assumption that relates the available entries with the missing ones is required. An alternative is to postulate that  $\mathbf{Z}$  has low rank  $R \ll \min(N, M)$ . The problem of finding matrix  $\hat{\mathbf{Z}}$  with rank not exceeding  $R$ , which approximates  $\mathbf{Z}$  in the given entries specified by a binary matrix  $\Delta \in \{0, 1\}^{N \times M}$ , can be formulated as

$$\hat{\mathbf{Z}} = \arg \min_{\mathbf{X}} \|(\mathbf{Z} - \mathbf{X}) \odot \Delta\|_F^2 \quad \text{s. to } \text{rank}(\mathbf{X}) \leq R. \quad (1)$$

The low-rank property of matrix  $\mathbf{X}$  implies that the vector  $\mathbf{s}(\mathbf{X})$  of its singular values is sparse. Hence, the rank constraint is equivalent to  $\|\mathbf{s}(\mathbf{X})\|_0 \leq R$ , where the  $\ell_0$ -(pseudo)norm  $\|\cdot\|_0$  equals the number of nonzero entries of its vector argument.

Aiming at a convex relaxation of the NP-hard problem (1), one can leverage recent advances in compressive sampling [11] and surrogate the  $\ell_0$ -norm with the  $\ell_1$ -norm, which here equals the nuclear norm of  $\mathbf{X}$  defined as  $\|\mathbf{X}\|_* := \|\mathbf{s}(\mathbf{X})\|_1$ . With this relaxation, the Lagrangian counterpart of (1) is

$$\hat{\mathbf{Z}} = \arg \min_{\mathbf{X}} \frac{1}{2} \|(\mathbf{Z} - \mathbf{X}) \odot \Delta\|_F^2 + \mu \|\mathbf{X}\|_* \quad (2)$$

where  $\mu \geq 0$  is a rank-controlling parameter. Problem (2) can be further transformed by considering the following characterization of the nuclear norm [23]

$$\|\mathbf{X}\|_* = \min_{\{\mathbf{B}, \mathbf{C}\}} \frac{1}{2} (\|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2) \quad \text{s. to } \mathbf{X} = \mathbf{B}\mathbf{C}^T. \quad (3)$$

For an arbitrary matrix  $\mathbf{X}$  with SVD  $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ , the minimum in (3) is attained for  $\mathbf{B} = \mathbf{U}\Sigma^{1/2}$  and  $\mathbf{C} = \mathbf{V}\Sigma^{1/2}$ . The optimization in (3) is over all possible bilinear factorizations of  $\mathbf{X}$ , so that the number of columns of  $\mathbf{B}$  and  $\mathbf{C}$  is also a variable. Building on (3), one can arrive at the following equivalent reformulation of (2) [17]

$$\begin{aligned} \hat{\mathbf{Z}}' = \arg \min_{\{\mathbf{X}, \mathbf{B}, \mathbf{C}\}} & \frac{1}{2} \|(\mathbf{Z} - \mathbf{X}) \odot \Delta\|_F^2 + \frac{\mu}{2} (\|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2) \\ \text{s. to } & \mathbf{X} = \mathbf{B}\mathbf{C}^T. \end{aligned} \quad (4)$$

The equivalence implies that by finding the global minimum of (4), one can recover the optimal solution of (2). However, since (4) is *nonconvex*, it may have multiple stationary points. Interestingly, the next result provides conditions for these stationary points to be globally optimal (parts a) and b) are proved in the Appendix, while the proof for c) can be found in [17].)

**Proposition 1:** Problems (2) and (4) are equivalent, in the sense that:

- a) global minima coincide:  $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}'$ ;
- b) all local minima of (4) are globally optimal; and,
- c) stationary points  $\mathbf{X}$  of (4) satisfying  $\|(\mathbf{X} - \mathbf{Z}) \odot \Delta\|_2 \leq \mu$  are globally optimal.

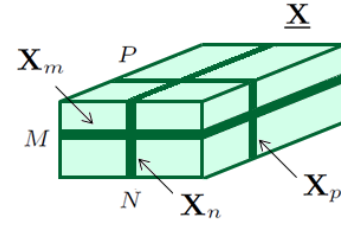


Fig. 1. Tensor slices along the row, column, and tube dimensions.

This result plays a critical role in this paper, as the Frobenius-norm regularization for controlling the rank in (4), will be useful to obtain its tensor counterparts in Section III.

### B. PARAFAC decomposition

The PARAFAC decomposition of a tensor  $\underline{\mathbf{X}} \in \mathbb{R}^{M \times N \times P}$  is at the heart of the proposed imputation method, since it offers a means to define its rank [9], [24]. Given  $R \in \mathbb{N}$ , consider matrices  $\mathbf{A} \in \mathbb{R}^{N \times R}$ ,  $\mathbf{B} \in \mathbb{R}^{M \times R}$ , and  $\mathbf{C} \in \mathbb{R}^{P \times R}$ , such that

$$\underline{\mathbf{X}}(m, n, p) = \sum_{r=1}^R \mathbf{A}(m, r) \mathbf{B}(n, r) \mathbf{C}(p, r). \quad (5)$$

The rank of  $\underline{\mathbf{X}}$  is the minimum value of  $R$  for which this decomposition is possible. For  $R^* := \text{rank}(\underline{\mathbf{X}})$ , the PARAFAC decomposition is given by the corresponding factor matrices  $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$  (all with  $R^*$  columns), so that (5) holds with  $R = R^*$ .

To appreciate why the aforementioned rank definition is natural, rewrite (5) as  $\underline{\mathbf{X}} = \sum_{r=1}^R \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$ , where  $\mathbf{a}_r$ ,  $\mathbf{b}_r$ , and  $\mathbf{c}_r$  represent the  $r$ -th columns of  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$ , respectively; and the outer products  $\underline{\mathbf{Q}}_r := \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \in \mathbb{R}^{M \times N \times P}$  have entries  $\underline{\mathbf{Q}}_r(m, n, p) := \mathbf{A}(m, r) \mathbf{B}(n, r) \mathbf{C}(p, r)$ . The rank of a tensor is thus the minimum number of outer products (rank one factors) required to represent the tensor. It is not uncommon to adopt an equivalent normalized representation

$$\underline{\mathbf{X}} = \sum_{r=1}^R \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = \sum_{r=1}^R \gamma_r (\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r) \quad (6)$$

by defining unit-norm vectors  $\mathbf{u}_r := \mathbf{a}_r / \|\mathbf{a}_r\|$ ,  $\mathbf{v}_r := \mathbf{b}_r / \|\mathbf{b}_r\|$ ,  $\mathbf{w}_r := \mathbf{c}_r / \|\mathbf{c}_r\|$ , and weights  $\gamma_r := \|\mathbf{a}_r\| \|\mathbf{b}_r\| \|\mathbf{c}_r\|$ ,  $r = 1, \dots, R$ .

Let  $\mathbf{X}_p$ ,  $p = 1, \dots, P$  denote the  $p$ -th slice of  $\underline{\mathbf{X}}$  along its third (tube) dimension, such that  $\mathbf{X}_p(m, n) := \underline{\mathbf{X}}(m, n, p)$ ; see Fig. 1. The following compact form of the PARAFAC decomposition in terms of slice factorizations will be used in the sequel

$$\mathbf{X}_p = \text{Adiag}[\mathbf{e}_p^T \mathbf{C}] \mathbf{B}, \quad p = 1, \dots, P \quad (7)$$

where the diagonal matrix  $\text{diag}[\mathbf{u}]$  has the vector  $\mathbf{u}$  on its diagonal, and  $\mathbf{e}_p^T$  is the  $p$ -th row of the  $P \times P$  identity matrix. The PARAFAC decomposition is symmetric [cf. (5)], and one can also write  $\mathbf{X}_m = \mathbf{B} \text{diag}[\mathbf{e}_m^T \mathbf{A}] \mathbf{C}$ , or,  $\mathbf{X}_n = \mathbf{C} \text{diag}[\mathbf{e}_n^T \mathbf{B}] \mathbf{A}$  in terms of slices along the first (row), or, second (column) dimensions.

### III. RANK REGULARIZATION FOR TENSORS

Generalizing the nuclear-norm regularization technique (2) from low-rank matrix to tensor completion is not straightforward, since singular values of a tensor (given by the Tucker decomposition) are not related to the rank [16]. Fortunately, the Frobenius-norm regularization outlined in Section II-A offers a viable option for low-rank tensor completion under the PARAFAC model, by solving

$$\begin{aligned} \hat{\underline{\mathbf{Z}}} := \arg \min_{\{\underline{\mathbf{X}}, \underline{\mathbf{A}}, \underline{\mathbf{B}}, \underline{\mathbf{C}}\}} & \frac{1}{2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 + \frac{\mu}{2} (\|\underline{\mathbf{A}}\|_F^2 + \|\underline{\mathbf{B}}\|_F^2 + \|\underline{\mathbf{C}}\|_F^2) \\ \text{s. to } & \underline{\mathbf{X}}_p = \underline{\mathbf{A}} \text{diag}[\mathbf{e}_p^T \underline{\mathbf{C}}] \underline{\mathbf{B}}, \quad p = 1, \dots, P \end{aligned} \quad (8)$$

where the Frobenius norm of a tensor is defined as  $\|\underline{\mathbf{X}}\|_F^2 := \sum_m \sum_n \sum_p \underline{\mathbf{X}}^2(m, n, p)$ , and the Hadamard product as  $(\underline{\mathbf{X}} \circledast \underline{\Delta})(m, n, p) := \underline{\mathbf{X}}(m, n, p) \underline{\Delta}(m, n, p)$ .

Different from the matrix case, it is unclear whether the regularization in (8) bears any relation with the tensor rank. Interestingly, the following analysis corroborates the capability of (8) to produce a low-rank tensor  $\hat{\underline{\mathbf{Z}}}$ , for sufficiently large  $\mu$ . In this direction, consider an alternative completion problem stated in terms of the normalized tensor representation (6)

$$\begin{aligned} \hat{\underline{\mathbf{Z}}} := \arg \min_{\{\underline{\mathbf{X}}, \gamma, \{\mathbf{u}_r\}, \{\mathbf{v}_r\}, \{\mathbf{w}_r\}\}} & \frac{1}{2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 + \frac{\mu}{2} \|\gamma\|_{2/3}^{2/3} \\ \text{s. to } & \underline{\mathbf{X}} = \sum_{r=1}^R \gamma_r (\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r) \end{aligned} \quad (9)$$

where  $\gamma := [\gamma_1, \dots, \gamma_R]^T$ ; the nonconvex  $\ell_{2/3}$  (pseudo)-norm is given by  $\|\gamma\|_{2/3} := (\sum_{r=1}^R |\gamma_r|^{2/3})^{3/2}$ ; and the unit-norm constraint on the factors' columns is left implicit. Problems (8) and (9) are equivalent as established by the following proposition (its proof is provided in the Appendix.)

**Proposition 2:** *The solutions of (8) and (9) coincide, i.e.,  $\hat{\underline{\mathbf{Z}}} = \hat{\underline{\mathbf{Z}}}$ , with optimal factors related by  $\hat{\mathbf{a}}_r = \sqrt[3]{\gamma_r} \hat{\mathbf{u}}_r$ ,  $\hat{\mathbf{b}}_r = \sqrt[3]{\gamma_r} \hat{\mathbf{v}}_r$ , and  $\hat{\mathbf{c}}_r = \sqrt[3]{\gamma_r} \hat{\mathbf{w}}_r$ ,  $r = 1, \dots, R$ .*

To further stress the capability of (8) to produce a low-rank approximant tensor  $\underline{\mathbf{X}}$ , consider transforming (9) once more by rewriting it in the constrained-error form

$$\begin{aligned} \hat{\underline{\mathbf{Z}}}'' := \arg \min_{\{\underline{\mathbf{X}}, \gamma, \{\mathbf{u}_r\}, \{\mathbf{v}_r\}, \{\mathbf{w}_r\}\}} & \|\gamma\|_{2/3} \\ \text{s. to } & \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 \leq \sigma^2, \quad \underline{\mathbf{X}} = \sum_{r=1}^R \gamma_r (\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r). \end{aligned} \quad (10)$$

For any value of  $\sigma^2$  there exists a corresponding Lagrange multiplier  $\lambda$  such that (9) and (10) yield the same solution, under the identity  $\mu = 2/\lambda$ . [Since  $f(x) = x^{2/3}$  is an increasing function, the exponent of  $\|\gamma\|_{2/3}$  can be safely eliminated without affecting the minimizer of (10).] The  $\ell_{2/3}$ -norm  $\|\gamma\|_{2/3}$  in (10) produces a sparse vector  $\gamma$  when minimized [6], sharing this well-documented property of the  $\ell_1$ -norm as their norm-one balls, depicted in Fig. 2, share the ‘‘pointy geometry’’ which is responsible for inducing sparsity.

With (8) equivalently rewritten as in (10), its low-rank inducing property is now revealed. As  $\gamma$  in (10) becomes sparse, some of its entries  $\gamma_r$  are zeroed, and the corresponding outer-products  $\gamma_r (\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$  drop from the sum in (6), thus lowering the rank of  $\underline{\mathbf{X}}$ .

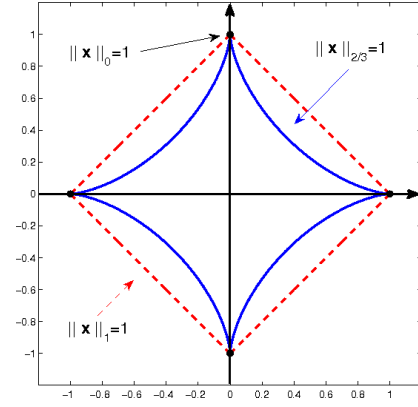


Fig. 2. The  $\ell_{2/3}$ -norm ball compared to the  $\ell_0$ - and  $\ell_1$ -norm balls

The next property is a direct consequence of the low-rank promoting property of (8) as established in Proposition 2.

**Corollary 1:** *If  $\hat{\underline{\mathbf{Z}}}$  denotes the solution to problem (8), and  $\mu \geq \mu_{\max} := \|\underline{\Delta} \circledast \underline{\mathbf{Z}}\|_F^{4/3}$ , then  $\hat{\underline{\mathbf{Z}}} = \mathbf{0}_{M \times N \times P}$ .*

Corollary 1 asserts that if the penalty parameter is chosen large enough, the rank is reduced to the extreme case  $\text{rank}(\hat{\underline{\mathbf{Z}}}) = 0$ . To see why this is a non-trivial property, it is prudent to think of ridge-regression estimates where similar quadratic regularizers are adopted, but an analogous property does not hold. In ridge regression one needs to let  $\mu \rightarrow \infty$  in order to obtain an all-zero solution. Characterization of  $\mu_{\max}$  is also of practical relevance as it provides a frame of reference for tuning the regularization parameter.

Using (10), it is also possible to relate (8) with the atomic norm in [5]. Indeed, the infimum  $\ell_1$ -norm of  $\gamma$  is a proper norm for  $\underline{\mathbf{X}}$ , named atomic norm, and denoted by  $\|\underline{\mathbf{X}}\|_{\mathcal{A}} := \|\gamma\|_1$  [5]. Thus, by replacing  $\|\gamma\|_{2/3}$  with  $\|\underline{\mathbf{X}}\|_{\mathcal{A}}$ , (10) becomes convex in  $\underline{\mathbf{X}}$ . Still, the complexity of solving such a variant of (10) resides in that  $\|\underline{\mathbf{X}}\|_{\mathcal{A}}$  is generally intractable to compute [5]. In this regard, it is remarkable that arriving to (10) had the sole purpose of demonstrating the low-rank inducing property, and that (8) is to be solved by the algorithm developed in the ensuing section. Such an algorithm will neither require computing the atomic norm or PARAFAC decomposition of  $\underline{\mathbf{X}}$ , nor knowing its rank. The number of columns in  $\underline{\mathbf{A}}$ ,  $\underline{\mathbf{B}}$ , and  $\underline{\mathbf{C}}$  can be set to an overestimate of the rank of  $\underline{\mathbf{Z}}$ , such as the upper bound  $\bar{R} := \min\{MN, NP, PM\} \geq \text{rank}(\underline{\mathbf{Z}})$ , and the low-rank of  $\underline{\mathbf{X}}$  will be induced by regularization as argued earlier. To carry out a fair comparison, only convergence to a stationary point of (8) will be guaranteed in this paper.

**Remark 1:** These insights foster future research directions for the design of a convex regularizer of the tensor rank. Specifically, substituting  $\rho(\underline{\mathbf{A}}, \underline{\mathbf{B}}, \underline{\mathbf{C}}) := \sum_{r=1}^R (\|\mathbf{a}_r\|^3 + \|\mathbf{b}_r\|^3 + \|\mathbf{c}_r\|^3)$  for the regularization term in (8), turns  $\|\gamma\|_{2/3}$  into  $\|\gamma\|_1 = \|\underline{\mathbf{X}}\|_{\mathcal{A}}$  in the equivalent (10). It is envisioned that with such a modification in place, the acquired convexity of (10) would enable a reformulation of Proposition 1, providing conditions for global optimality of the stationary points of (8).

Still, a limitation of (8) is that it does not allow for incorporating side information that could be available in addition to the given entries  $\underline{\Delta} \circledast \underline{\mathbf{Z}}$ .

**Remark 2:** In the context of recommender systems, a description of the users and/or products through attributes (e.g.,



gender, age) or measures of similarity, is typically available. It is thus meaningful to exploit both known preferences and descriptions to model the preferences of users [1]. In three-way (samples, genes, conditions) microarray data analysis, the relative position of single-nucleotide polymorphisms in the DNA molecule implies degrees of correlation among genotypes [22]. These correlations could be available either through a prescribed model, or, through estimates obtained using a reference tensor  $\underline{\mathbf{Z}}$ . A probabilistic approach to tensor completion capable of incorporating such types of extra information is the subject of the ensuing section.

#### IV. BAYESIAN LOW-RANK TENSOR APPROXIMATION

##### A. Bayesian PARAFAC model

A probabilistic approach is developed in this section in order to integrate the available statistical information into the tensor imputation setup. To this end, suppose that the observation noise is zero-mean, white, Gaussian; that is

$$Z_{mnp} = X_{mnp} + e_{mnp}; \text{ such that } e_{mnp} \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d..} \quad (11)$$

Since vectors  $\mathbf{a}_r$  in (6) are interchangeable, identical distributions are assigned across  $r = 1, \dots, R$ , and they are modeled as independent from each other, zero-mean Gaussian distributed with covariance matrix  $\mathbf{R}_A \in \mathbb{R}^{M \times M}$ . Similarly, vectors  $\mathbf{b}_r$  and  $\mathbf{c}_r$  are uncorrelated and zero-mean, Gaussian, with covariance matrix  $\mathbf{R}_B$  and  $\mathbf{R}_C$ , respectively. In addition  $\mathbf{a}_r$ ,  $\mathbf{b}_r$ , and  $\mathbf{c}_r$  are assumed mutually uncorrelated. And since scale ambiguity is inherently present in the PARAFAC model, vectors  $\mathbf{a}_r$ ,  $\mathbf{b}_r$ , and  $\mathbf{c}_r$  are set to have equal power; that is,

$$\theta := \text{Tr}(\mathbf{R}_A) = \text{Tr}(\mathbf{R}_B) = \text{Tr}(\mathbf{R}_C). \quad (12)$$

Under these assumptions, the negative of the posterior distribution can be readily written as  $\exp(-L(\underline{\mathbf{X}}))$ , with

$$\begin{aligned} L(\underline{\mathbf{X}}) &= \frac{1}{2\sigma^2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 \\ &+ \frac{1}{2} \sum_{r=1}^R (\mathbf{a}_r^T \mathbf{R}_A^{-1} \mathbf{a}_r + \mathbf{b}_r^T \mathbf{R}_B^{-1} \mathbf{b}_r + \mathbf{c}_r^T \mathbf{R}_C^{-1} \mathbf{c}_r) \\ &= \frac{1}{2\sigma^2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 + \frac{1}{2} [\text{Tr}(\mathbf{A}^T \mathbf{R}_A^{-1} \mathbf{A}) \\ &+ \text{Tr}(\mathbf{B}^T \mathbf{R}_B^{-1} \mathbf{B}) + \text{Tr}(\mathbf{C}^T \mathbf{R}_C^{-1} \mathbf{C})]. \end{aligned}$$

Correspondingly, the MAP estimator of  $\underline{\mathbf{X}}$  is

$$\begin{aligned} \hat{\underline{\mathbf{X}}} &:= \arg \min_{\{\underline{\mathbf{X}}, \mathbf{A}, \mathbf{B}, \mathbf{C}\}} \frac{1}{2\sigma^2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 + \frac{1}{2} [\text{Tr}(\mathbf{A}^T \mathbf{R}_A^{-1} \mathbf{A}) \\ &+ \text{Tr}(\mathbf{B}^T \mathbf{R}_B^{-1} \mathbf{B}) + \text{Tr}(\mathbf{C}^T \mathbf{R}_C^{-1} \mathbf{C})] \\ \text{s. to } \mathbf{X}_p &= \text{A} \text{diag}[\mathbf{e}_p^T \mathbf{C}] \mathbf{B}^T, \quad p = 1, \dots, P \end{aligned} \quad (13)$$

reducing to (8) when  $\mathbf{R}_A = \mathbf{I}_M$ ,  $\mathbf{R}_B = \mathbf{I}_N$ , and  $\mathbf{R}_C = \mathbf{I}_P$ . This Bayesian approach interprets the regularization parameter  $\mu$  [cf. (8)] as the noise variance, which is useful in practice to select  $\mu$ . The ensuing section explores the advantages of incorporating prior information to the imputation method.

##### B. Nonparametric tensor decomposition

Incorporating the information conveyed by  $\mathbf{R}_A$ ,  $\mathbf{R}_B$ , and  $\mathbf{R}_C$ , together with a practical means of finding these matrices can be facilitated by interpreting (13) in the context of RKHS [27]. In particular, the analysis presented next will use the Representer Theorem, interpreted as an instrument for finding the best interpolating function in a Hilbert space spanned by kernels, just as interpolation with sinc-kernels is carried out in the space of bandlimited functions for the purpose of reconstructing a signal from its samples [19].

In this context, it is instructive to look at a tensor  $f : \mathcal{M} \times \mathcal{N} \times \mathcal{P} \rightarrow \mathbb{R}$  as a function of three variables  $m, n$ , and  $p$ , living in measurable spaces  $\mathcal{M}$ ,  $\mathcal{N}$ , and  $\mathcal{P}$ , respectively. Generalizing (8) to this nonparametric framework, low-rank functions  $f$  are formally defined to belong to the following family

$$\mathcal{F}_R := \{f : \mathcal{M} \times \mathcal{N} \times \mathcal{P} \rightarrow \mathbb{R} : f(m, n, p) = \sum_{r=1}^R a_r(m) b_r(n) c_r(p) \text{ such that } a_r(m) \in \mathcal{H}_{\mathcal{M}}, b_r(n) \in \mathcal{H}_{\mathcal{N}}, c_r(p) \in \mathcal{H}_{\mathcal{P}}\}$$

where  $\mathcal{H}_{\mathcal{M}}$ ,  $\mathcal{H}_{\mathcal{N}}$ , and  $\mathcal{H}_{\mathcal{P}}$  are Hilbert spaces constructed from specified kernels  $k_{\mathcal{M}}$ ,  $k_{\mathcal{N}}$  and  $k_{\mathcal{P}}$ , defined over  $\mathcal{M}$ ,  $\mathcal{N}$ , and  $\mathcal{P}$ , while  $R$  is an initial overestimate of the rank of  $f$ .

The following nonparametric fitting criterion is adopted for finding the best  $\hat{f}_R$  interpolating data  $\{z_{mnp} : \delta_{mnp} = 1\}$

$$\begin{aligned} \hat{f}_R &:= \arg \min_{f \in \mathcal{F}_R} \sum_{m=1}^M \sum_{n=1}^N \sum_{p=1}^P \delta_{mnp} (z_{mnp} - f(m, n, p))^2 \\ &+ \frac{\mu}{2} \sum_{r=1}^R (\|a_r\|_{\mathcal{H}_{\mathcal{M}}}^2 + \|b_r\|_{\mathcal{H}_{\mathcal{N}}}^2 + \|c_r\|_{\mathcal{H}_{\mathcal{P}}}^2). \end{aligned} \quad (14)$$

It is shown in the Appendix that leveraging the Representer Theorem, the minimizer of (14) admits a finite dimensional representation in terms of  $k_{\mathcal{M}}$ ,  $k_{\mathcal{N}}$  and  $k_{\mathcal{P}}$ ,

$$\hat{f}_R(m, n, p) = \mathbf{k}_{\mathcal{M}}^T(m) \mathbf{K}_{\mathcal{M}}^{-1} \mathbf{A} \text{diag}[\mathbf{k}_{\mathcal{P}}^T(p) \mathbf{K}_{\mathcal{P}}^{-1} \mathbf{C}] \mathbf{B}^T \mathbf{K}_{\mathcal{N}}^{-1} \mathbf{k}_{\mathcal{N}}(n) \quad (15)$$

where vector  $\mathbf{k}_{\mathcal{M}}(m)$  and matrix  $\mathbf{K}_{\mathcal{M}}$  have entries  $k_{\mathcal{M}}(m, m')$ ,  $m, m' = 1, \dots, M$ ; and where  $\mathbf{k}_{\mathcal{N}}(n)$ ,  $\mathbf{K}_{\mathcal{N}}$ ,  $\mathbf{k}_{\mathcal{P}}(p)$ , and  $\mathbf{K}_{\mathcal{P}}$  are correspondingly defined in terms of  $k_{\mathcal{N}}$  and  $k_{\mathcal{P}}$ . It is also shown in the Appendix that the coefficient matrices  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  can be found by solving

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}, \mathbf{C}} \sum_{p=1}^P \|(\mathbf{Z}_p - \mathbf{A} \text{diag}[\mathbf{e}_p^T \mathbf{C}] \mathbf{B}^T) \circledast \underline{\Delta}_p\|_F^2 \\ + \frac{\mu}{2} (\text{Tr}(\mathbf{A}^T \mathbf{K}_{\mathcal{M}}^{-1} \mathbf{A}) + \text{Tr}(\mathbf{B}^T \mathbf{K}_{\mathcal{N}}^{-1} \mathbf{B}) + \text{Tr}(\mathbf{C}^T \mathbf{K}_{\mathcal{P}}^{-1} \mathbf{C})) \\ \text{s. to } \mathbf{A} \in \mathbb{R}^{M \times R}, \mathbf{B} \in \mathbb{R}^{N \times R}, \mathbf{C} \in \mathbb{R}^{P \times R}. \end{aligned} \quad (16)$$

Problem (16) reduces to (8) when the side information is discarded by selecting  $k_{\mathcal{M}}$ ,  $k_{\mathcal{N}}$  and  $k_{\mathcal{P}}$  as Kronecker deltas, in which case  $\mathbf{K}_{\mathcal{M}}$ ,  $\mathbf{K}_{\mathcal{N}}$ , and  $\mathbf{K}_{\mathcal{P}}$  are identity matrices. In the general case, (16) yields the sought nonlinear low-rank approximation method for  $f(m, n, p)$  when combined with (15), evidencing the equivalence between (14) and (13).

Interpreting (14) as an interpolator renders (13) a natural choice for tensor imputation, where in general, missing entries are to be inserted by connecting them to surrounding points on

the three-dimensional arrangement. Relative to (8), this RKHS perspective also highlights (13)'s extra smoothing and extrapolation capabilities. Indeed, by capitalizing on the similarities captured by  $\mathbf{K}_M$ ,  $\mathbf{K}_N$  and  $\mathbf{K}_P$ , (16) can recover completely missing slices. This feature is not shared by imputation methods that leverage low-rank only, since these require at least one point in the slice to build on colinearities. Extrapolation is also possible in this sense. If for instance  $\mathbf{K}_M$  can be expanded to capture a further point  $M + 1$  not in the original set, then a new slice of data can be predicted by (15) based on its correlation  $k_M(M + 1)$  with the available entries. These extra capabilities will be exploited in Section VI, where correlations are leveraged for the imputation of MRI data. The method described by (13) and (16) can be applied to matrix completion by just setting entries of  $\mathbf{C}$  to one, and can be extended to higher-order dimensions with a straightforward alteration of the algorithms and theorems throughout this paper.

Identification of covariance matrices  $\mathbf{R}_A$ ,  $\mathbf{R}_B$ , and  $\mathbf{R}_C$  with kernel matrices  $\mathbf{K}_M$ ,  $\mathbf{K}_N$  and  $\mathbf{K}_P$  is the remaining aspect to clarify in the connection between (13) and (16). It is apparent from (13) and (16) that correlations between columns of the factors are reflected in similarities between the tensor slices, giving rise to the opportunity of obtaining one from the other. This aspect is explored next.

### C. Covariance estimation

To implement (13), matrices  $\mathbf{R}_A$ ,  $\mathbf{R}_B$ , and  $\mathbf{R}_C$  must be postulated a priori, or alternatively replaced by their sample estimates. Such estimates need a training set of vectors  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $\mathbf{c}$  abiding to the Bayesian model just described, and this requires PARAFAC decomposition of training data. In order to abridge this procedure, it is convenient to inspect how  $\mathbf{R}_A$ ,  $\mathbf{R}_B$ , and  $\mathbf{R}_C$  are related to their kernel counterparts.

Based on the equivalence between the standard RKHS interpolator and the linear mean-square error estimator [21], it is useful to re-visit the probabilistic framework and identify kernel similarities between slices of  $\underline{\mathbf{X}}$  with their mutual covariances. Focusing on the tube dimension of  $\underline{\mathbf{X}}$ , one can write  $\mathbf{K}_P(p', p) := \mathbb{E}(\text{Tr}(\mathbf{X}_{p'}^T \mathbf{X}_p))$ , that is, the covariance between slices  $\mathbf{X}_{p'}$  and  $\mathbf{X}_p$  taking  $\langle \mathbf{X}, \mathbf{Y} \rangle := \text{Tr}(\mathbf{X}^T \mathbf{Y})$  as the standard inner product in the matrix space. Under this alternative definition for  $\mathbf{K}_P$ , and corresponding definitions for  $\mathbf{K}_N$ , and  $\mathbf{K}_M$ , it is shown in the Appendix that

$$\mathbf{K}_M = \theta^2 \mathbf{R}_A, \quad \mathbf{K}_N = \theta^2 \mathbf{R}_B, \quad \mathbf{K}_P = \theta^2 \mathbf{R}_C \quad (17)$$

and that  $\theta$  is related to the second-order moment of  $\underline{\mathbf{X}}$  by

$$\mathbb{E} \|\underline{\mathbf{X}}\|_F^2 = R\theta^3. \quad (18)$$

Since sample estimates for  $\mathbf{K}_M$ ,  $\mathbf{K}_N$ ,  $\mathbf{K}_P$ , and  $\mathbb{E} \|\underline{\mathbf{X}}\|_F$  can be readily obtained from the tensor data, (17) and (18) provide an agile means of estimating  $\mathbf{R}_A$ ,  $\mathbf{R}_B$ , and  $\mathbf{R}_C$  without requiring PARAFAC decompositions over the set of training tensors.

This strategy remains valid when kernels are not estimated from data. One such case emerges in collaborative filtering of user preferences [1], where the similarity of two users is modeled as a function of attributes; such age or income.

### D. Block successive upper-bound minimization algorithm

An iterative algorithm is developed here for solving (13), by cyclically minimizing the cost over  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$ . In the first step of the cycle the cost in (13) is minimized with respect to (w.r.t.)  $\mathbf{A}$  considering  $\mathbf{B}$  and  $\mathbf{C}$  as parameters. Accordingly, the partial cost to minimize reduces to

$$f(\mathbf{A}) := \frac{1}{2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 + \frac{\mu}{2} \text{Tr}(\mathbf{A}^T \mathbf{R}_A^{-1} \mathbf{A}) \quad (19)$$

where  $\mu$  was identified with and substituted for  $\sigma^2$ . Function (19) is quadratic in  $\mathbf{A}$  and can be readily minimized after re-writing it in terms of  $\mathbf{a} := \text{vec}(\mathbf{A})$  [see (47) in the Appendix]. However, such an approach becomes computationally infeasible for other than small datasets, since it involves storing  $P$  matrices of dimensions  $NM \times MR$ , and solving a linear system of  $MR \times MR$  equations. The alternative pursued here to overcome this obstacle relies on the so-called block successive upper-bound minimization (BSUM) algorithm [20].

In BSUM one minimizes a judiciously chosen upper-bound  $g(\mathbf{A}, \bar{\mathbf{A}})$  of  $f(\mathbf{A})$ , which: i) depends on the current iterate  $\bar{\mathbf{A}}$ ; ii) should be simpler to optimize; and iii) satisfies certain local-tightness conditions; see also [20] and properties i)-iii) below.

For  $\bar{\mathbf{A}}$  given, consider the function

$$g(\mathbf{A}, \bar{\mathbf{A}}) := \frac{1}{2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \circledast \underline{\Delta}\|_F^2 + \mu \left( \frac{\lambda}{2} \text{Tr}(\mathbf{A}^T \mathbf{A}) - \text{Tr}(\Theta^T \mathbf{A}) + \frac{1}{2} \text{Tr}(\Theta^T \bar{\mathbf{A}}) \right) \quad (20)$$

where  $\lambda := \lambda_{\max}(\mathbf{R}_A^{-1})$  is the maximum eigenvalue of  $\mathbf{R}_A^{-1}$ , and  $\Theta := \lambda \mathbf{I} - \mathbf{R}_A^{-1}$ . The following properties of  $g(\mathbf{A}, \bar{\mathbf{A}})$  imply that it majorizes  $f(\mathbf{A})$  at  $\bar{\mathbf{A}}$ , satisfying the technical conditions required for the convergence of BSUM (properties i)-iii) are established in the the proof of Lemma 1 in the Appendix).

- i)  $f(\bar{\mathbf{A}}) = g(\bar{\mathbf{A}}, \bar{\mathbf{A}})$ ;
- ii)  $\frac{d}{d\mathbf{A}} f(\mathbf{A})|_{\mathbf{A}=\bar{\mathbf{A}}} = \frac{d}{d\mathbf{A}} g(\mathbf{A}, \bar{\mathbf{A}})|_{\mathbf{A}=\bar{\mathbf{A}}}$ ; and,
- iii)  $f(\mathbf{A}) \leq g(\mathbf{A}, \bar{\mathbf{A}})$ ,  $\forall \mathbf{A}$ .

The computational advantage of minimizing  $g(\mathbf{A}, \bar{\mathbf{A}})$  in place of  $f(\mathbf{A})$  comes from  $g(\mathbf{A}, \bar{\mathbf{A}})$  being separable across rows of  $\mathbf{A}$ . To see this, consider the Kathri-Rao product  $\Pi := \mathbf{C} \odot \mathbf{B} := [\mathbf{c}_1 \otimes \mathbf{b}_1, \dots, \mathbf{c}_R \otimes \mathbf{b}_R]$ , defined by the column-wise Kronecker products  $\mathbf{c}_r \otimes \mathbf{b}_r$ . Let also matrix  $\mathbf{Z} := [\mathbf{Z}_1, \dots, \mathbf{Z}_P] \in \mathbb{N}^{M \times NP}$  denote the unfolding of  $\underline{\mathbf{Z}}$  along its tube dimension, and likewise for  $\Delta := [\Delta_1, \dots, \Delta_P] \in \{0, 1\}^{M \times NP}$  and  $\mathbf{X} := [\mathbf{X}_1, \dots, \mathbf{X}_P] \in \mathbb{R}_+^{M \times NP}$ . Then, using the following identity [10]

$$\mathbf{X} := [\mathbf{X}_1, \dots, \mathbf{X}_P] = \mathbf{A} \Pi^T. \quad (21)$$

it is possible to rewrite (20) as

$$g(\mathbf{A}, \bar{\mathbf{A}}) := \frac{1}{2} \|(\mathbf{Z} - \mathbf{A} \Pi^T) \circledast \Delta\|_F^2 + \mu \left( \frac{\lambda}{2} \text{Tr}(\mathbf{A}^T \mathbf{A}) - \text{Tr}(\Theta^T \mathbf{A}) + \frac{1}{2} \text{Tr}(\Theta^T \bar{\mathbf{A}}) \right)$$

**Algorithm 1** : Low-rank tensor imputation (LRTI)

---

```

1: function UPDATE_FACTOR( $\mathbf{A}, \mathbf{R}, \mathbf{\Pi}, \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu$ )
2:   Set  $\lambda = \lambda_{\max}(\mathbf{R}^{-1})$ 
3:   Unfold  $\underline{\mathbf{A}}$  and  $\underline{\mathbf{Z}}$  over dimension of  $\mathbf{A}$  into  $\mathbf{\Delta}$  and  $\mathbf{Z}$ 
4:   Set  $\mathbf{\Theta} = (\lambda \mathbf{I} - \mathbf{R}^{-1})\mathbf{A}$ 
5:   for  $m = 1, \dots, M$  do
6:     Select rows  $\mathbf{z}_m^T$ ,  $\delta_m^T$ , and  $\theta_m^T$ , and set  $\mathbf{D}_m = \text{diag}(\delta_m)$ 
7:     Compute  $\mathbf{a}_m = (\mathbf{\Pi}^T \mathbf{D}_m \mathbf{\Pi} + \lambda \mu \mathbf{I})^{-1} (\mathbf{\Pi}^T \mathbf{D}_m \mathbf{z}_m + \mu \theta_m)$ 
8:     Update  $\mathbf{A}$  with row  $\mathbf{a}_m^T$ 
9:   end for
10:  return  $\mathbf{A}$ 
11: end function
12: Initialize  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  randomly.
13: while  $|\text{cost} - \text{cost\_old}| < \epsilon$  do
14:    $\mathbf{A} = \text{UPDATE\_FACTOR}(\mathbf{A}, \mathbf{R}_A, (\mathbf{C} \odot \mathbf{B}), \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu)$ 
15:    $\mathbf{B} = \text{UPDATE\_FACTOR}(\mathbf{B}, \mathbf{R}_B, (\mathbf{A} \odot \mathbf{C}), \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu)$ 
16:    $\mathbf{C} = \text{UPDATE\_FACTOR}(\mathbf{C}, \mathbf{R}_C, (\mathbf{B} \odot \mathbf{A}), \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu)$ 
17:   Recalculate cost in (13)
18: end while
19: return  $\underline{\mathbf{X}}$  with slices  $\hat{\mathbf{X}}_p = \mathbf{A} \text{diag}(\mathbf{e}_p^T \mathbf{C}) \mathbf{B}^T$ 

```

---

which can be decomposed as

$$g(\mathbf{A}, \bar{\mathbf{A}}) = \sum_{m=1}^M \left[ \frac{1}{2} \|\delta_m \otimes \mathbf{z}_m - \text{diag}(\delta_m) \mathbf{\Pi} \mathbf{a}_m\|_2^2 + \frac{\mu}{2} (\lambda \|\mathbf{a}_m\|^2 + \theta_m^T \mathbf{a}_m + \theta_m^T \bar{\mathbf{a}}_m) \right] \quad (22)$$

where  $\mathbf{z}_m^T$ ,  $\mathbf{a}_m^T$ ,  $\delta_m^T$ ,  $\theta_m^T$ , and  $\bar{\mathbf{a}}_m^T$ , represent the  $m$ -th rows of matrices  $\mathbf{Z}$ ,  $\mathbf{A}$ ,  $\mathbf{\Delta}$ ,  $\mathbf{\Theta}$ , and  $\bar{\mathbf{A}}$ , respectively. Not only (22) evidences the separability of (20) across rows of  $\mathbf{A}$ , but it also presents each of its summands in a standardized quadratic form that can be readily minimized by equating its gradient to zero. Accordingly, the majorization strategy reduces the computational load to  $R$  systems of  $M$  equations that can be solved in parallel. Collecting the solution of such quadratic programs into the rows of a matrix  $\mathbf{A}^*$  yields the minimizer of (20), and the update  $\mathbf{A} \leftarrow \mathbf{A}^*$  for the BSUM cycle. Such a procedure is presented in Algorithm 1, where analogous updates for  $\mathbf{B}$  and  $\mathbf{C}$  are carried out cyclically.

By virtue of properties i)-iii) in Lemma 1, convergence of Algorithm 1 follows readily from that of the BSUM algorithm [20].

**Proposition 3:** *The iterates for  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  generated by Algorithm 1 converge to a stationary point of (13).*

## V. INFERENCE FOR LOW-RANK POISSON TENSORS

Adoption of the LS criterion in (8) assumes in a Bayesian setting that the random  $\underline{\mathbf{Z}}$  is Gaussian distributed. This section deals with a Poisson-distributed tensor  $\underline{\mathbf{Z}}$ , a natural alternative to the Gaussian model when integer-valued data are obtained by counting independent events [10]. Suppose that the entries  $z_{mnp}$  of  $\underline{\mathbf{Z}}$  are Poisson distributed, with probability mass function

$$P(z_{mnp} = k) = \frac{x_{mnp}^k e^{-x_{mnp}}}{k!} \quad (23)$$

and means given by the corresponding entries in tensor  $\underline{\mathbf{X}}$ . For mutually-independent  $\{z_{mnp}\}$ , the log-likelihood  $l_{\underline{\mathbf{A}}}(\underline{\mathbf{Z}}; \underline{\mathbf{X}})$  of

$\underline{\mathbf{X}}$  given data  $\underline{\mathbf{Z}}$  only on the entries specified by  $\underline{\mathbf{A}}$ , takes the form

$$l_{\underline{\mathbf{A}}}(\underline{\mathbf{Z}}; \underline{\mathbf{X}}) = \sum_{m=1}^M \sum_{n=1}^N \sum_{p=1}^P \delta_{mnp} [z_{mnp} \log(x_{mnp}) - x_{mnp}] \quad (24)$$

after dropping terms  $\log(z_{mnp}!)$  that do not depend on  $\underline{\mathbf{X}}$ .

The choice of the Poisson distribution in (23) over a Gaussian one for counting data, prompts minimization of the K-L divergence (24) instead of LS as a more suitable criterion [10]. Still, the entries of  $\underline{\mathbf{X}}$  are not coupled in (24), and a binding PARAFAC modeling assumption is natural for feasibility of the tensor approximation task under missing data. Mimicking the method for Gaussian data, (nonnegative) Gaussian priors are assumed for the factors of the PARAFAC decomposition. Accordingly, the MAP estimator of  $\underline{\mathbf{X}}$  given Poisson-distributed data (entries of  $\underline{\mathbf{Z}}$  indexed by  $\underline{\mathbf{A}}$ ) becomes

$$\hat{\underline{\mathbf{Z}}} := \underset{\{\underline{\mathbf{X}}, \mathbf{A}, \mathbf{B}, \mathbf{C}\} \in \mathcal{T}}{\text{argmin}} \sum_{m=1}^M \sum_{n=1}^N \sum_{p=1}^P \delta_{mnp} (x_{mnp} - z_{mnp} \log(x_{mnp})) + \frac{\mu}{2} [\text{Tr}(\mathbf{A}^T \mathbf{R}_A^{-1} \mathbf{A}) + \text{Tr}(\mathbf{B}^T \mathbf{R}_B^{-1} \mathbf{B}) + \text{Tr}(\mathbf{C}^T \mathbf{R}_C^{-1} \mathbf{C})] \quad (25)$$

over the feasible set  $\mathcal{T} := \{\underline{\mathbf{X}}, \mathbf{A}, \mathbf{B}, \mathbf{C} : \mathbf{A} \geq \mathbf{0}, \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \mathbf{X}_p = \mathbf{A} \text{diag}[\mathbf{e}_p^T \mathbf{C}] \mathbf{B}^T, p = 1, \dots, P\}$ , where the symbol  $\geq$  should be understood to imply entry-wise nonnegativity.

With the aid of Representer's Theorem, it is also possible to interpret (25) as a variational estimator in RKHS, with K-L analogues to (14)-(16), so that the conclusions thereby regarding smoothing, prediction and prior covariance estimation carry over to the low-rank Poisson imputation method (25).

### A. Block successive upper-bound minimization algorithm

A K-L counterpart of the LRTI algorithm is developed in this section, that provably converges to a stationary point of (25), via an alternating-minimization iteration which optimizes (25) sequentially w.r.t. one factor matrix, while holding the others fixed.

In the sequel, the goal is to arrive at a suitable expression for the cost in (25), when viewed only as a function of e.g.,  $\mathbf{A}$ . To this end, let matrix  $\mathbf{Z} := [\mathbf{Z}_1, \dots, \mathbf{Z}_P] \in \mathbb{N}^{M \times NP}$  denote the unfolding of  $\underline{\mathbf{Z}}$  along its tube dimension, and likewise for  $\mathbf{\Delta} := [\mathbf{\Delta}_1, \dots, \mathbf{\Delta}_P] \in \{0, 1\}^{M \times NP}$  and  $\mathbf{X} := [\mathbf{X}_1, \dots, \mathbf{X}_P] \in \mathbb{R}_+^{M \times NP}$ . Based on these definitions, (24) can be written as

$$l_{\underline{\mathbf{A}}}(\underline{\mathbf{Z}}; \underline{\mathbf{X}}) = \mathbf{1}_M^T (\mathbf{\Delta} \otimes [\mathbf{X} - \mathbf{Z} \otimes \log(\mathbf{X})]) \mathbf{1}_{NP} \quad (26)$$

where  $\mathbf{1}_M$ ,  $\mathbf{1}_{NP}$  are all-one vectors of dimensions  $M$  and  $NP$  respectively, and  $\log(\cdot)$  should be understood entry-wise. The log-likelihood in (26) can be expressed in terms of  $\mathbf{A}$ , and the Kathri-Rao product  $\mathbf{\Pi} := \mathbf{B} \odot \mathbf{C}$  by resorting again to (21). Substituting (21) into (26) one arrives at the desired expression for the cost in (25) as a function of  $\mathbf{A}$ , namely

$$f(\mathbf{A}) := \mathbf{1}_M^T (\mathbf{\Delta} \otimes [\mathbf{A} \mathbf{\Pi} - \mathbf{Z} \otimes \log(\mathbf{A} \mathbf{\Pi}^T)]) \mathbf{1}_{NP} + \frac{\mu}{2} \text{Tr}(\mathbf{A}^T \mathbf{R}_A^{-1} \mathbf{A}).$$

A closed-form minimizer  $\mathbf{A}^*$  for  $f(\mathbf{A})$  is not available, but since  $f(\mathbf{A})$  is convex one could in principle resort to an

**Algorithm 2** : Low-rank Poisson-tensor imputation (LRPTI)

---

```

1: function UPDATE_FACTOR( $\mathbf{A}, \mathbf{R}, \mathbf{\Pi}, \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu$ )
2:   Set  $\lambda = \lambda_{\max}(\mathbf{R}^{-1})$ 
3:   Unfold  $\underline{\mathbf{A}}$  and  $\underline{\mathbf{Z}}$  over dimension of  $\mathbf{A}$  into  $\Delta$  and  $\mathbf{Z}$ 
4:   Compute  $\mathbf{S} = \frac{\Delta}{\lambda \mu} \circledast (\frac{\Delta \otimes \mathbf{Z}}{\mathbf{A} \mathbf{\Pi}^T \mathbf{\Pi}})$  (element-wise division)
5:   Compute  $\mathbf{T} = \frac{1}{2\lambda \mu} (\mu(\lambda \mathbf{I} - \mathbf{R}^{-1})\mathbf{A} - \Delta \mathbf{\Pi})$ 
6:   Update  $\mathbf{A}$  with entries  $a_{mr} = t_{mr} + \sqrt{t_{mr}^2 + s_{mr}}$ 
7:   return  $\mathbf{A}$ 
8: end function
9: Initialize  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  randomly.
10: while  $|\text{cost} - \text{cost\_old}| < \epsilon$  do
11:    $\mathbf{A} = \text{UPDATE\_FACTOR}(\mathbf{A}, \mathbf{R}_A, (\mathbf{C} \odot \mathbf{B}), \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu)$ 
12:    $\mathbf{B} = \text{UPDATE\_FACTOR}(\mathbf{B}, \mathbf{R}_B, (\mathbf{A} \odot \mathbf{C}), \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu)$ 
13:    $\mathbf{C} = \text{UPDATE\_FACTOR}(\mathbf{C}, \mathbf{R}_C, (\mathbf{B} \odot \mathbf{A}), \underline{\mathbf{A}}, \underline{\mathbf{Z}}, \mu)$ 
14:   Recalculate cost in (25)
15: end while
16: return  $\underline{\mathbf{X}}$  with slices  $\hat{\mathbf{X}}_p = \mathbf{A} \text{diag}(\mathbf{e}_p^T \mathbf{C}) \mathbf{B}^T$ 

```

---

iterative procedure to obtain  $\mathbf{A}^*$ . To avoid extra inner iterations, the approach here relies again on the BSUM algorithm [20].

For  $\bar{\mathbf{A}}$  given, consider the separable function

$$g(\mathbf{A}, \bar{\mathbf{A}}) := \mu \lambda \sum_{m,r=1}^{M,R} \left( \frac{a_{mr}^2}{2} - 2t_{mr}a_{mr} - s_{mr} \log(a_{mr}) + u_{mr} \right) \quad (27)$$

where  $\lambda := \lambda_{\max}(\mathbf{R}_A^{-1})$  is the largest eigenvalue of  $\mathbf{R}_A^{-1}$ , and the parameters  $s_{mr}$ ,  $t_{mr}$ , and  $u_{mr}$  are defined in terms of  $\bar{\mathbf{A}}$ ,  $\mathbf{Z}$ ,  $\Delta$ ,  $\mathbf{\Pi}$ , and  $\Theta := (\lambda \mathbf{I} - \mathbf{R}_A^{-1}) \bar{\mathbf{A}}$  by

$$s_{mr} := \frac{1}{\lambda \mu} \sum_{k=1}^{NP} \frac{\delta_{mk} z_{mk} \bar{a}_{mr} \pi_{kr}}{\sum_{r'=1}^R \bar{a}_{mr'} \pi_{kr'}},$$

$$t_{mr} := \frac{1}{2\lambda \mu} \left( \mu \theta_{mr} - \sum_{k=1}^{NP} \pi_{kr} \delta_{mk} \right)$$

and  $u_{mr} := \frac{1}{\lambda \mu} \left( \theta_{mr} \bar{a}_{mr} + \sum_{k=1}^{NP} \delta_{mk} z_{mk} \bar{a}_{mr} \pi_{kr} v_{mrk} \right)$ , with  $v_{mrk} := \log(\bar{a}_{mr} \pi_{kr} / \sum_{r'=1}^R \bar{a}_{mr'} \pi_{kr'}) / \sum_{r'=1}^R \bar{a}_{mr'} \pi_{kr'}$ . As asserted in the following lemma,  $g(\mathbf{A}, \bar{\mathbf{A}})$  majorizes  $f(\mathbf{A})$  at  $\bar{\mathbf{A}}$  and satisfies the technical conditions required for the convergence of BSUM (see the Appendix for a proof.)

**Lemma 1:** Function  $g(\mathbf{A}, \bar{\mathbf{A}})$  satisfies the following properties

- i)  $f(\bar{\mathbf{A}}) = g(\bar{\mathbf{A}}, \bar{\mathbf{A}})$ ;
- ii)  $\frac{d}{d\mathbf{A}} f(\mathbf{A})|_{\mathbf{A}=\bar{\mathbf{A}}} = \frac{d}{d\mathbf{A}} g(\mathbf{A}, \bar{\mathbf{A}})|_{\mathbf{A}=\bar{\mathbf{A}}}$ ; and,
- iii)  $f(\mathbf{A}) \leq g(\mathbf{A}, \bar{\mathbf{A}})$ ,  $\forall \mathbf{A}$ .

Moreover,  $g(\mathbf{A}, \bar{\mathbf{A}})$  is minimized at  $\mathbf{A} = \mathbf{A}_g^*$  with entries  $a_{g,mr}^* := t_{mr} + \sqrt{t_{mr}^2 + s_{mr}}$ .

Lemma 1 highlights the reason behind adopting  $g(\mathbf{A}, \bar{\mathbf{A}})$  in the proposed block-coordinate descent algorithm: it is separable across the entries of its matrix argument [cf. (27)], and hence it admits a closed-form minimizer given by the  $MR$  scalars  $a_{g,mr}^*$ . The updates  $\mathbf{A} \leftarrow \mathbf{A}_g^*$  are tabulated under Algorithm 2 for solving (25), where analogous updates for  $\mathbf{B}$  and  $\mathbf{C}$  are carried out cyclically.

By virtue of properties i)-iii) in Lemma 1, convergence of Algorithm 2 follows readily from the general convergence theory available for the BSUM algorithm [20].

**Proposition 4:** The iterates for  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  generated by Algorithm 2 converge to a stationary point of (25).

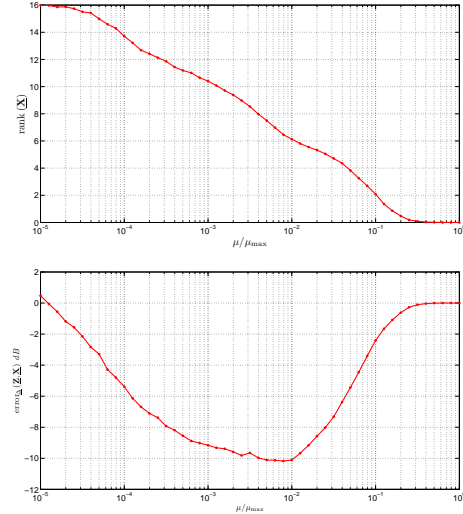


Fig. 3. Performance of the low-rank tensor imputation method as function of the regularizing parameter  $\mu$ ; (top) rank of the tensor as recovered by (8) averaged over 100 test repetitions, (bottom) relative recovery error.

A related algorithm, abbreviated as CP-APR can be found in [10], where the objective is to find the tensor's low-rank factors per se. The LRPTI algorithm here generalizes CP-APR by focusing on recovering missing data, and incorporating prior information through rank regularization. In terms of convergence to a stationary point, the added regularization allows for lifting the assumption on the linear independence of the rows of  $\mathbf{\Pi}$ , as required by CP-APR [10] - an assumption without a straightforward validation since iterates  $\mathbf{\Pi}$  are not accessible beforehand.

## VI. NUMERICAL TESTS

### A. Simulated Gaussian data

Synthetic tensor-data of dimensions  $M \times N \times P = 16 \times 4 \times 4$  were generated according to the Bayesian tensor model described in Section IV. Specifically, entries of  $\underline{\mathbf{Z}}$  consist of realizations of Gaussian random variables generated according to (11), with means specified by entries of  $\underline{\mathbf{X}}$  and variance scaled to yield an SNR of  $-20\text{dB}$ . Tensor  $\underline{\mathbf{X}}$  is constructed from factors  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$ , as in (7). Matrices  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  have  $R = 6$  columns containing realizations of independent zero-mean, unit-variance, Gaussian random variables.

A quarter of the entries of  $\underline{\mathbf{Z}}$  were removed at random and reserved to evaluate performance. The remaining seventy five percent of the data were used to recover  $\underline{\mathbf{Z}}$  considering the removed data as missing entries. Method (8) was employed for recovery, as implemented by the LRTI Algorithm, with regularization  $\frac{\mu}{2} (\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2)$  resulting from setting  $\mathbf{R}_A = \mathbf{I}_M$ ,  $\mathbf{R}_B = \mathbf{I}_N$ , and  $\mathbf{R}_C = \mathbf{I}_P$ .

The relative recovery error between  $\hat{\underline{\mathbf{Z}}}$  and data  $\underline{\mathbf{Z}}$  was computed, along with the rank of the recovered tensor, as a measure of performance. Fig. 3 depicts these figures of merit averaged over 100 repetitions of the experiment, across values of  $\mu$  varying on the interval  $10^{-5} \mu_{\max}$  to  $\mu_{\max}$ , which is computed as in Corollary 1. Fig 3 (bottom) shows that the LRTI algorithm is successful in recovering the missing entries



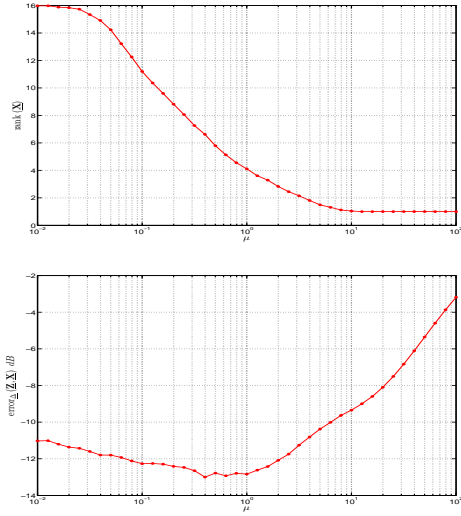


Fig. 4. Performance of the low-rank Poisson imputation method as function of the regularizing parameter  $\mu$ ; (top) rank of the recovered tensor averaged over 100 test repetitions, (bottom) relative recovery error.

of  $\underline{Z}$  up to  $-10$ dB for a wide range of values of  $\mu$ , presenting a minimum at  $\mu = 10^{-2}\mu_{\max}$ . This result is consistent with Fig. 3 (top), which shows that rank  $R^* = 6$  is approximately recovered at the minimum error. Fig. 3 (top) also corroborates the low-rank inducing effect of (8), with the recovered rank varying from the upper bound  $\bar{R} = NP = 16$  to  $R = 0$ , as  $\mu$  is increased, and confirms that the recovered tensor is null at  $\mu_{\max}$  as asserted by Corollary 1.

### B. Simulated Poisson data

The synthetic example just described was repeated for the low-rank Poisson-tensor model described in Section V. Specifically, tensor data of dimensions  $M \times N \times P = 16 \times 4 \times 4$  were generated according to the low-rank Poisson-tensor model of Section V. Entries of  $\underline{Z}$  consist of realizations of Poisson random variables generated according to (23), with means specified by entries of  $\underline{X}$ . Tensor  $\underline{X}$  is again constructed as in (7) from factors  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  having  $R = 6$  columns, containing the absolute value of realizations of independent Gaussian random variables scaled to yield  $\mathbb{E}[x_{mnp}] = 100$ . Half of the entries of  $\underline{Z}$  were considered missing to be recovered from the remaining half. Method (25) was employed for recovery, as implemented by the LRPTI Algorithm, with regularization  $\frac{\mu}{2}(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2)$ .

Fig. 4 shows the estimated rank and recovery error over 100 realizations of the experiment, for  $\mu$  in the interval 0.01 to 100. The recovery error in Fig. 4 (bottom) exhibits a minimum of  $-15$ dB at  $\mu = 1$ , where the rank  $R^* = 6$  is recovered [cf. Fig. 4 (top).] The low-rank inducing effect of (8) is again corroborated by the decreasing trend in Fig. 4 (top), but in this case the rank is lower bounded by  $R = 1$ , because the K-L fitting criterion prevents (25) from yielding a null estimate  $\hat{\underline{Z}}$ .

### C. MRI data

Estimator (14) was tested against a corrupted version of the MRI brain data set 657 from the Internet brain segmentation

repository [15]. The tensor  $\underline{Z}$  to be estimated corresponds to a three-dimensional MRI scan of the brain comprising a set of  $P = 18$  images, each of  $M \times N = 256 \times 196$  pixels. Fifty percent of the data is removed uniformly at random together with the whole slice  $\mathbf{Z}_n$ ,  $n = 50$ . Fig. 5 depicts the results of applying estimator (14) to the remaining data, which yields a reconstruction error of  $-10.54$ dB. The original slice  $\mathbf{Z}_p$ ,  $p = 5$ , its corrupted counterpart, and the resulting estimate are shown on top and center left. Covariance matrices  $\mathbf{K}_M$ ,  $\mathbf{K}_N$  and  $\mathbf{K}_P$  are estimated from six additional tensor samples containing complementary scans of the brain also available at [15]. Fig. 5 (center right) represents the covariance matrix  $\mathbf{K}_N$  for column slices perpendicular to  $\mathbf{Z}_p$ , showing a structure that reflects symmetries of the brain. This correlation is the key enabler for the method to recover the missing slice up to  $-9.60$ dB (see Fig. 5 (bottom)) by interpolating its a priori similar parallel counterparts.

All in all, the experiment evidences the merits of low-rank PARAFAC decomposition for modeling a tensor, the ability of the Bayesian estimator (13) in recovering missing data, and the usefulness of incorporating correlations as side information.

On account of the comprehensive analysis of three-way MRI data arrays in [8], and the nonnegative PARAFAC decomposition advanced thereby, inference of tensors with nonnegative continuous entries will be pursued as future research, combining methods and algorithms in sections IV and V of this paper.

### D. RNA sequencing data

The RNA-Seq method described in [18] exhaustively counts the number of RNA transcripts from yeast cells. The reverse transcription of RNA molecules into cDNA is achieved by  $P = 2$  alternative methods, differentiated by the use of oligo-dT, or random-hexonucleotide primers. These cDNA molecules are sequenced to obtain counts of RNA molecules across  $M = 6,604$  genes on the yeast genome. The experiment was repeated in [18] for a biological and a technological replicate of the original sample totalling  $N = 3$  instances per primer selection. The data are thus organized in a tensor of dimensions  $6,604 \times 3 \times 2$  as shown in Fig. 6 (top), with integer data that are modeled as Poisson counts. Fifteen percent of the data is removed and reserved for assessing performance. The missing data are represented in white in Fig. 6 (center).

The LRPTI algorithm is run with the data available in Fig. 6 (center) producing the recovered tensor depicted in Fig. 6 (bottom). The recovery error for this experiment was  $-15$ dB.

## VII. CONCLUDING SUMMARY

It was shown in this paper that regularizing with the Frobenius-norm square of the PARAFAC decomposition factors, controls the tensor's rank by inducing sparsity in the vector of amplitudes of its rank-one components. A Bayesian method for tensor completion was developed based on this property, introducing priors on the tensor factors. It was argued, and corroborated numerically, that this prior information endows the completion method with extra capabilities in terms of smoothing and extrapolation. It was also suggested through



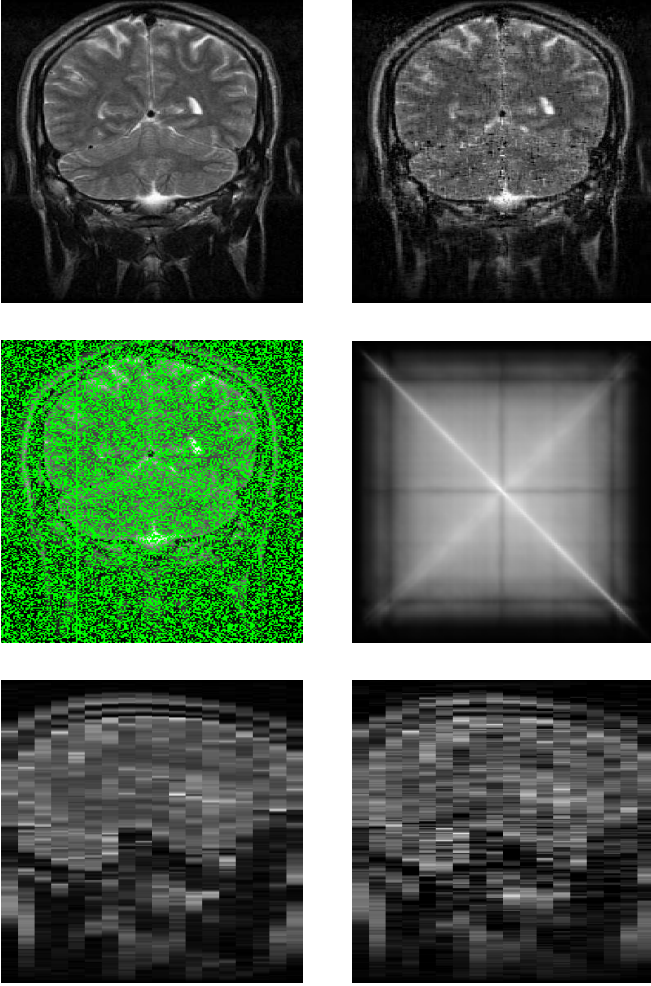


Fig. 5. Results of applying (14) to the MRI brain data set 657. (top) original and recovered fibers  $\mathbf{Z}_p$  and  $\hat{\mathbf{Z}}_p$  for  $p = 5$ . (center) input fiber  $\mathbf{Z}_p$ ,  $p = 5$  with missing data, and covariance matrix  $\mathbf{K}_{\mathcal{N}}$ . (bottom) original and recovered columns  $\mathbf{Z}_n$  and  $\hat{\mathbf{Z}}_n$  for the position  $n = 50$  in which the whole input slice is missing )

a parallelism between Bayesian and RKHS inference, that the prior covariance matrices can be obtained from (sample) correlations among the tensor's slices. In such a probabilistic context, generic distribution models for the data lead to multiple fitting criteria. Gaussian and Poisson processes were especially considered by developing algorithms that minimize the regularized LS and K-L divergence, respectively.

Numerical tests on synthetic data corroborated the low-rank inducing property, and the ability of the completion method to recover the “ground-truth” rank, while experiments with brain images and gene expression levels in yeast served to evaluate the method's performance on real datasets.

Although the results and algorithms in this paper were presented for three-way arrays, they are readily extendible to higher-order tensors or reducible to the matrix case.

## APPENDIX

### I. Proof of Proposition 1

*Proof:* a) The equivalence of (2) and (4) results immedi-

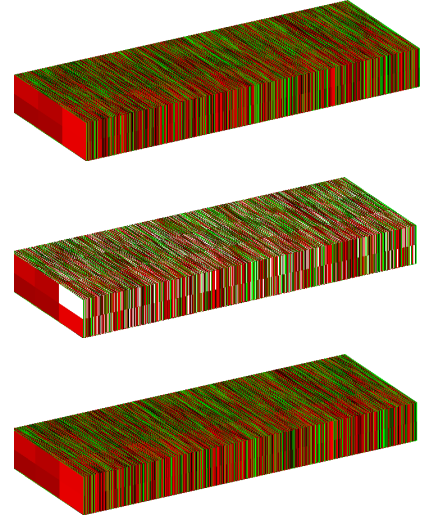


Fig. 6. Imputation of sequencing count data via LRPTI; (top) original data; (center) data with missing entries; (bottom) recovered tensor.

ately from (3). Indeed, if (4) is minimized in two steps

$$\min_{\mathbf{X}} \frac{1}{2} \min_{\mathbf{B}, \mathbf{C}} \|(\mathbf{Z} - \mathbf{X}) \otimes \Delta\|_F^2 + \frac{\mu}{2} (\|\mathbf{C}\|_F^2 + \|\mathbf{B}\|_F^2) \quad \text{s. to } \mathbf{CB}^T = \mathbf{X} \quad (28)$$

it is apparent that the LS part of the cost does not depend on the inner minimization variables. Hence, (28) can be rewritten as

$$\min_{\mathbf{X}} \frac{1}{2} \|(\mathbf{Z} - \mathbf{X}) \otimes \Delta\|_F^2 + \min_{\substack{\mathbf{B}, \mathbf{C} \\ \text{s. to } \mathbf{CB}^T = \mathbf{X}}} \frac{\mu}{2} (\|\mathbf{C}\|_F^2 + \|\mathbf{B}\|_F^2) \quad (29)$$

and by recognizing (3) as the inner problem in (29), the equivalence follows.

b) Consider the cost in (4) at the local minimum  $(\bar{\mathbf{B}}, \bar{\mathbf{C}})$

$$U(\bar{\mathbf{B}}, \bar{\mathbf{C}}) := \frac{1}{2} \|(\mathbf{Z} - \bar{\mathbf{X}}) \otimes \Delta\|_F^2 + \frac{\mu}{2} (\|\bar{\mathbf{C}}\|_F^2 + \|\bar{\mathbf{B}}\|_F^2)$$

where  $\bar{\mathbf{X}} := \bar{\mathbf{B}}\bar{\mathbf{C}}^T$ . Arguing by contradiction, suppose that there is a different local minimum  $(\mathbf{B}, \mathbf{C})$  such that  $U(\mathbf{B}, \mathbf{C}) \neq U(\bar{\mathbf{B}}, \bar{\mathbf{C}})$ , and without loss of generality set  $U(\mathbf{B}, \mathbf{C}) < U(\bar{\mathbf{B}}, \bar{\mathbf{C}})$ , so that  $dU := U(\mathbf{B}, \mathbf{C}) - U(\bar{\mathbf{B}}, \bar{\mathbf{C}}) < 0$ , which can be expanded to

$$dU = \text{Tr}[(\Delta \otimes (\mathbf{Z} - \bar{\mathbf{X}}))(\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}))] + \|\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X})\|_F^2 + \frac{\mu}{2} (\|\mathbf{C}\|_F^2 - \|\bar{\mathbf{C}}\|_F^2 + \|\mathbf{B}\|_F^2 - \|\bar{\mathbf{B}}\|_F^2) < 0. \quad (30)$$

Setting this inequality aside for now, consider the augmented matrix  $\mathbf{Q}$  in terms of generic  $\mathbf{B}$  and  $\mathbf{C}$  matrices:

$$\mathbf{Q} := \begin{bmatrix} \mathbf{B} \\ \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{B}^T & \mathbf{C}^T \end{bmatrix} = \begin{pmatrix} \mathbf{B}\mathbf{B}^T & \mathbf{X} \\ \mathbf{X}^T & \mathbf{C}\mathbf{C}^T \end{pmatrix} \quad (31)$$

and the corresponding  $\bar{\mathbf{Q}}$  defined in terms of  $\bar{\mathbf{B}}$  and  $\bar{\mathbf{C}}$ .

For each value of  $\theta \in (0, 1)$  consider the convex combination

$$\mathbf{Q}_\theta := \bar{\mathbf{Q}} + \theta(\mathbf{Q} - \bar{\mathbf{Q}}). \quad (32)$$

As both  $\mathbf{Q}$  and  $\bar{\mathbf{Q}}$  are positive semi-definite, so is  $\mathbf{Q}_\theta$  and by means of the Choleski factorization one obtains

$$\mathbf{Q}_\theta := \begin{bmatrix} \mathbf{B}_\theta \\ \mathbf{C}_\theta \end{bmatrix} \begin{bmatrix} \mathbf{B}_\theta' & \mathbf{C}_\theta' \end{bmatrix} = \begin{pmatrix} \mathbf{B}_\theta\mathbf{B}_\theta' & \mathbf{X}_\theta \\ \mathbf{X}_\theta' & \mathbf{C}_\theta\mathbf{C}_\theta' \end{pmatrix}. \quad (33)$$

which defines  $\mathbf{B}_\theta$ ,  $\mathbf{C}_\theta$  and  $\mathbf{X}_\theta$ .

Expanding the cost difference  $dU_\theta$  as in (30) results in

$$\begin{aligned} dU_\theta &:= U(\mathbf{B}_\theta, \mathbf{C}_\theta) - U(\bar{\mathbf{B}}, \bar{\mathbf{C}}) \\ &= \text{Tr} [(\Delta \otimes (\mathbf{Z} - \bar{\mathbf{X}})) (\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}_\theta))] \\ &\quad + \frac{\mu}{2} (\|\mathbf{C}_\theta\|_F^2 - \|\bar{\mathbf{C}}\|_F^2 + \|\mathbf{B}_\theta\|_F^2 - \|\bar{\mathbf{B}}\|_F^2) \\ &\quad + \|\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}_\theta)\|_F^2. \end{aligned}$$

From the definitions (31)-(33) it follows that  $\bar{\mathbf{X}} - \mathbf{X}_\theta = \theta(\bar{\mathbf{X}} - \mathbf{X})$ ,  $\|\mathbf{B}_\theta\|_F^2 - \|\bar{\mathbf{B}}\|_F^2 = \theta(\|\mathbf{B}\|_F^2 - \|\bar{\mathbf{B}}\|_F^2)$ , and  $\|\mathbf{C}_\theta\|_F^2 - \|\bar{\mathbf{C}}\|_F^2 = \theta(\|\mathbf{C}\|_F^2 - \|\bar{\mathbf{C}}\|_F^2)$ , so that

$$\begin{aligned} dU_\theta &:= \theta \text{Tr} [(\Delta \otimes (\mathbf{Z} - \bar{\mathbf{X}})) (\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}))] \\ &\quad + \frac{\mu\theta}{2} (\|\mathbf{C}\|_F^2 - \|\bar{\mathbf{C}}\|_F^2 + \|\mathbf{B}\|_F^2 - \|\bar{\mathbf{B}}\|_F^2) \\ &\quad + \theta^2 \|\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}_\theta)\|_F^2 \end{aligned}$$

and thus, it can be put in terms of (30) as in

$$dU_\theta := \theta (dU - \|\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}_\theta)\|_F^2) + \theta^2 \|\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}_\theta)\|_F^2.$$

If  $dU$  were strictly negative, so would  $dU - \|\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}_\theta)\|_F^2$ , and hence

$$\lim_{\theta \rightarrow 0} \frac{1}{\theta} dU_\theta = (dU - \|\Delta \otimes (\bar{\mathbf{X}} - \mathbf{X}_\theta)\|_F^2) < 0.$$

but then there is in every neighborhood of  $(\bar{\mathbf{B}}, \bar{\mathbf{C}})$  a point  $(\mathbf{B}_\theta, \mathbf{C}_\theta)$  such that  $U(\mathbf{B}_\theta, \mathbf{C}_\theta) < U(\bar{\mathbf{B}}, \bar{\mathbf{C}})$ ,  $\bar{\mathbf{B}}, \bar{\mathbf{C}}$  cannot be a local minimum. This contradiction implies that  $U(\mathbf{B}, \mathbf{C}) = U(\bar{\mathbf{B}}, \bar{\mathbf{C}})$  for any pair of local minima, which proves the statement in part b) of Proposition 1. ■

## II-Equivalence of tensor completion problems

*Proof:* The Frobenius square-norms of  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  are separable across columns; hence, the penalty in (8) can be rewritten as

$$\begin{aligned} \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2 &= \sum_{r=1}^R \|\mathbf{a}_r\|^2 + \|\mathbf{b}_r\|^2 + \|\mathbf{c}_r\|^2 \\ &= \sum_{r=1}^R a_r^2 + b_r^2 + c_r^2 \end{aligned} \quad (34)$$

by defining  $a_r := \|\mathbf{a}_r\|$ ,  $b_r := \|\mathbf{b}_r\|$ ,  $c_r := \|\mathbf{c}_r\|$ ,  $r = 1, \dots, R$ .

On the other hand,  $\underline{\mathbf{X}}$  can be expressed w.l.o.g. in terms of the normalized outer products (6) with  $\gamma_r := a_r b_r c_r$ . Substituting (6) and (34) for the tensor and the penalty respectively, (8) reduces to

$$\begin{aligned} \min_{\{\hat{\mathbf{u}}\}, \{\hat{\mathbf{v}}\}, \{\hat{\mathbf{w}}\}} \min_{\gamma} \min_{\{a_r\}, \{b_r\}, \{c_r\}} &\frac{1}{2} \|(\underline{\mathbf{Z}} - \underline{\mathbf{X}}) \otimes \underline{\Delta}\|_F^2 \\ &+ \frac{\mu}{2} \sum_{r=1}^R a_r^2 + b_r^2 + c_r^2 \\ \text{s. to } \underline{\mathbf{X}} &= \sum_{r=1}^R \gamma_r (\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r) \\ \gamma_r &= a_r b_r c_r. \end{aligned} \quad (35)$$

Focusing on the inner minimization w.r.t. norms  $a_r$ ,  $b_r$ , and  $c_r$  for arbitrary fixed directions  $\{\mathbf{u}_r\}$ ,  $\{\mathbf{v}_r\}$ , and  $\{\mathbf{w}_r\}$ , and fixed products  $\gamma_r := a_r b_r c_r$ . The constraints and hence

the LS part of the cost depend on  $\gamma_r$  only, and not on their particular factorizations  $a_r b_r c_r$ . Thus, the penalty is the only term that varies when  $\gamma_r$  is constant, rendering the inner-most minimization in (35) equivalent to

$$\begin{aligned} \min_{a_r, b_r, c_r} a_r^2 + b_r^2 + c_r^2 \\ \gamma_r = a_r b_r c_r. \end{aligned} \quad (36)$$

The arithmetic geometric-mean inequality gives the solution to (36), as it states that for scalars  $a_r^2$ ,  $b_r^2$  and  $c_r^2$ , it holds that

$$\sqrt[3]{a_r^2 b_r^2 c_r^2} \leq (1/3)(a_r^2 + b_r^2 + c_r^2)$$

with equality when  $a_r^2 = b_r^2 = c_r^2$ , so that the minimum of (36) is attained at  $a_r^2 = b_r^2 = c_r^2 = \gamma_r^{2/3}$ .

Substituting the corresponding  $\sum_{r=1}^R (a_r^2 + b_r^2 + c_r^2) = 3 \sum_{r=1}^R \gamma_r^{2/3} = 3 \|\gamma\|_{2/3}^{2/3}$  into (35) yields (9). Equivalence of the optimization problems is transitive; hence, by showing that both (9) and (8) equivalent to (35) proves them equivalent to each other, as desired. ■

## III. Proof of Corollary 1

*Proof:* The following result on the norm of the matrix inverse will be used in the proof of the corollary.

**Lemma 2:** [13, p.58] If  $\mathbf{E} \in \mathbb{R}^{m \times m}$  satisfies  $\|\mathbf{E}\|_F \leq 1$ , then  $\mathbf{I} + \mathbf{E}$  is invertible, and  $\|(\mathbf{I} + \mathbf{E})^{-1}\|_F \leq (1 - \|\mathbf{E}\|_F)^{-1}$ .

Another useful inequality holds for any value of  $\mu$ , and for  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  being the minimizers of (8)

$$\mu (\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2) \leq \|\Delta \otimes \underline{\mathbf{Z}}\|_F^2 \quad (37)$$

as it follows from comparing the cost at such a minimum, and at the feasible point  $(\mathbf{A}, \mathbf{B}, \mathbf{C}) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$ .

A second characterization of the minimum of (8) will be obtained by equating the gradient to zero. By vectorizing matrix  $\mathbf{A}$ , the cost in (8) can be rewritten as

$$\sum_{p=1}^P \frac{1}{2} \left\| \text{diag}[\delta_p] (\mathbf{z}_p - (\mathbf{B} \text{diag}[\mathbf{e}_p^T \mathbf{C}] \otimes \mathbf{I})) \mathbf{a} \right\|_2^2 + \frac{\mu}{2} \|\mathbf{a}\|_2^2 \quad (38)$$

where  $\mathbf{z}_p$ ,  $\delta_p$ , and  $\mathbf{a}$  denote the vector rearrangements of matrices  $\mathbf{Z}_p$ ,  $\mathbf{D}_p$ , and  $\mathbf{A}$ , respectively. Additional regularization that vanishes when taking derivatives w.r.t.  $\mathbf{A}$  were removed from (38). Setting the gradient of (38) w.r.t.  $\mathbf{a}$  to zero, yields

$$\mathbf{a} = (\mathbf{I} + \mathbf{E})^{-1} \boldsymbol{\zeta}$$

with

$$\begin{aligned} \mathbf{E} &:= \frac{1}{\mu} \sum_{p=1}^P (\mathbf{B}^T \text{diag}[\mathbf{e}_p^T \mathbf{C}] \otimes \mathbf{I}) \text{diag}[\delta_p] (\mathbf{B} \text{diag}[\mathbf{e}_p^T \mathbf{C}] \otimes \mathbf{I}) \\ \boldsymbol{\zeta} &:= \frac{1}{\mu} \sum_{p=1}^P (\mathbf{B}^T \text{diag}[\mathbf{e}_p^T \mathbf{C}] \otimes \mathbf{I}) \text{diag}[\delta_p] \mathbf{z}_p. \end{aligned}$$

The norms of  $\mathbf{E}$  and  $\boldsymbol{\zeta}$  can be bounded by using the sub-multiplicative property of the norm, and the Cauchy-Schwarz inequality, which results in

$$\begin{aligned} \|\mathbf{E}\|_F &\leq \frac{1}{\mu} \|\mathbf{B}\|_F^2 \|\mathbf{C}\|_F^2 \\ \|\boldsymbol{\zeta}\|_F &\leq \frac{1}{\mu} \|\Delta \otimes \underline{\mathbf{Z}}\|_F \|\mathbf{B}\|_F \|\mathbf{C}\|_F. \end{aligned}$$

Then according to the previous lemma, if  $\mu$  is chosen large enough so that  $\|\mathbf{E}\|_F \leq 1$  then the norm of  $\mathbf{A}$  is bounded by

$$\|\mathbf{A}\|_F = \|\mathbf{a}\|_2 \leq (\mu - \|\mathbf{B}\|_F^2 \|\mathbf{C}\|_F^2)^{-1} \|\mathbf{B}\|_F \|\mathbf{C}\|_F \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F \quad (39)$$

which constitutes the sought second characterization of the minimum of (8).

Yet a third characterization was obtained during the proof of Proposition 2, in which the norm of the factor columns were shown equal to each other, so that

$$\|\mathbf{A}\|_F = \|\mathbf{B}\|_F = \|\mathbf{C}\|_F. \quad (40)$$

Substituting (40) into (37) and (39) yields

$$\|\mathbf{A}\|_F^2 \leq \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F^2 / 3\mu \quad (41)$$

$$\|\mathbf{A}\|_F \leq (\mu - \|\mathbf{A}\|_F^4)^{-1} \|\mathbf{A}\|_F^2 \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F. \quad (42)$$

Form (42), two cases are found possible:

**case 1:**  $\|\mathbf{A}\|_F = 0$ ; and

**case 2:**  $1 \leq (1 - \|\mathbf{A}\|_F^4 / \mu)^{-1} \|\mathbf{A}\|_F \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F / \mu. \quad (43)$

To argue that the second case is impossible, substitute (41) into (43) and square the result to obtain

$$1 \leq (1 - \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F^4 / 9\mu^3)^{-2} \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F^4 / 3\mu^3 \quad (44)$$

But by hypothesis  $\mu \geq \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F^{4/3}$  so that  $\|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F^4 / \mu^3 \leq 1$ , and the right-hand side of (44) is bounded by 0.43, so that the inequality does not hold. This implies that the first case in (43); i.e.,  $\|\mathbf{A}\|_F = 0$ , must hold, which in accordance with (40), further implies a null solution of (8). That was the object of this proof. Still, the bound at 0.43 can be pushed to one by further reducing  $\mu$ , and the proof remains valid under the slightly relaxed condition  $\mu > (18/(5 + \sqrt{21}))^{-1/3} \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F^{4/3} \simeq 0.81 \|\underline{\Delta} \otimes \underline{\mathbf{Z}}\|_F^{4/3}$ . ■

#### IV-RKHS imputation

Recursive application of Representer's Theorem yields finite dimensional representations for the minimizers  $a_r$ ,  $b_r$ , and  $c_r$  of (14), given by

$$\hat{a}_r(m) = \sum_{m'=1}^M \alpha_{rm'} k_{\mathcal{M}}(m', m)$$

$$\hat{b}_r(n) = \sum_{n'=1}^N \beta_{rn'} k_{\mathcal{N}}(n', n)$$

$$\hat{c}_r(p) = \sum_{p'=1}^P \gamma_{rp'} k_{\mathcal{P}}(p', p).$$

Defining vectors  $\mathbf{k}_{\mathcal{M}}^T(m) := [k_{\mathcal{M}}(1, m), \dots, k_{\mathcal{M}}(M, m)]$ , and correspondingly  $\mathbf{k}_{\mathcal{N}}^T(n) := [k_{\mathcal{N}}(1, n), \dots, k_{\mathcal{N}}(N, n)]$ , and  $\mathbf{k}_{\mathcal{P}}^T(p) := [k_{\mathcal{P}}(1, p), \dots, k_{\mathcal{P}}(P, p)]$ , along with matrices  $\hat{\mathbf{A}} \in \mathbb{R}^{M \times R}$ :  $\hat{A}(m, r) := \alpha_{mr}$ ,  $\hat{\mathbf{B}} \in \mathbb{R}^{N \times R}$ :  $\hat{B}(n, r) := \beta_{nr}$ , and  $\hat{\mathbf{C}} \in \mathbb{R}^{P \times R}$ :  $\hat{C}(p, r) := \gamma_{pr}$ , it follows that

$$\begin{aligned} \hat{f}_R(m, n, p) &= \sum_{r=1}^R \hat{a}_r(m) \hat{b}_r(n) \hat{c}_r(p) \\ &= \mathbf{k}_{\mathcal{M}}^T(m) \hat{\mathbf{A}} \text{diag}[\mathbf{k}_{\mathcal{P}}^T(p) \hat{\mathbf{C}}] \hat{\mathbf{B}}^T \mathbf{k}_{\mathcal{N}}(n). \end{aligned} \quad (45)$$

Matrices  $\hat{\mathbf{A}}$ ,  $\hat{\mathbf{B}}$ , and  $\hat{\mathbf{C}}$  are further obtained by solving

$$\begin{aligned} \min_{\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}} \sum_{p=1}^P &\|(\mathbf{Z}_p - \mathbf{K}_{\mathcal{M}} \mathbf{A} \text{diag}[\mathbf{e}_p^T \mathbf{K}_{\mathcal{P}} \mathbf{C}] \mathbf{B}^T \mathbf{K}_{\mathcal{N}}) \otimes \underline{\Delta}_p\|_F^2 \\ &+ \frac{\mu}{2} (\text{trace}(\mathbf{A}^T \mathbf{K}_{\mathcal{M}} \mathbf{A}) + \text{trace}(\mathbf{B}^T \mathbf{K}_{\mathcal{N}} \mathbf{B}) + \text{trace}(\mathbf{C}^T \mathbf{K}_{\mathcal{P}} \mathbf{C})) \\ \text{s. to } &\mathbf{A} \in \mathbb{R}^{M \times R}, \mathbf{B} \in \mathbb{R}^{N \times R}, \mathbf{C} \in \mathbb{R}^{P \times R} \end{aligned}$$

which is transformed into (16) by changing variables  $\mathbf{A} = \mathbf{K}_{\mathcal{M}} \hat{\mathbf{A}}$ ,  $\mathbf{B} = \mathbf{K}_{\mathcal{N}} \hat{\mathbf{B}}$ , and  $\mathbf{C} = \mathbf{K}_{\mathcal{P}} \hat{\mathbf{C}}$ , just as (45) becomes (15).

#### V-Covariance estimation

Inspection of the entries of  $\mathbf{K}_{\mathcal{P}}(p, p') := \mathbb{E}[\text{Tr}(\mathbf{X}_p^T \mathbf{X}_{p'})]$  under the PARAFAC model, yields

$$\begin{aligned} \mathbf{K}_{\mathcal{P}}(p, p') &:= \mathbb{E} \left[ \text{Tr} \left( \sum_{r=1}^R \mathbf{b}_r \mathbf{c}_r(p) \mathbf{a}_r^T \sum_{r'=1}^R \mathbf{a}_{r'} \mathbf{c}_{r'}(p') \mathbf{b}_{r'}^T \right) \right] \\ &= \sum_{r=1}^R \sum_{r'=1}^R \mathbb{E}(\mathbf{c}_r^T(p) \mathbf{c}_{r'}(p')) \mathbb{E}(\mathbf{b}_{r'}^T \mathbf{b}_r) \mathbb{E}(\mathbf{a}_r^T \mathbf{a}_{r'}) \\ &= \sum_{r=1}^R \mathbb{E}(\mathbf{c}_r(p) \mathbf{c}_r(p')) \mathbb{E}\|\mathbf{b}_r\|^2 \mathbb{E}\|\mathbf{a}_r\|^2 \\ &= \sum_{r=1}^R \mathbf{R}_C(p, p') \text{Tr}(\mathbf{R}_B) \text{Tr}(\mathbf{R}_A) \\ &= R \mathbf{R}_C(p, p') \text{Tr}(\mathbf{R}_B) \text{Tr}(\mathbf{R}_A) \end{aligned}$$

which, after summing over  $p' = p$ , yields

$$\begin{aligned} \mathbb{E}\|\underline{\mathbf{X}}\|_F^2 &= \sum_{p=1}^P \mathbb{E}\|\mathbf{X}_p\|_F^2 = \sum_{p=1}^P \mathbf{R}_{\mathcal{P}}(p, p) \\ &= R \text{Tr}(\mathbf{R}_C) \text{Tr}(\mathbf{R}_B) \text{Tr}(\mathbf{R}_A). \end{aligned} \quad (46)$$

In addition, by incorporating the equal power assumption (12), equation (46) further simplifies to

$$\mathbb{E}\|\underline{\mathbf{X}}\|_F^2 = R\theta^3$$

as stated in (18).

#### VI - Vector form of (19)

The vec operator can be combined with the Kronecker product to factorize  $\text{vec}(\mathbf{A} \mathbf{Q}^T) = (\mathbf{Q} \otimes \mathbf{I}) \text{vec}(\mathbf{A})$ , and with the Hadamard product to convert it to a standard matrix product  $\text{vec}(\underline{\Delta} \otimes \mathbf{A}) = \text{diag}(\text{vec}(\underline{\Delta})) \text{vec}(\mathbf{A})$ . Using these two properties, (19) can be put in terms of  $\mathbf{a} := \text{vec}(\mathbf{A})$  as in

$$\begin{aligned} f(\mathbf{a}) &:= \frac{1}{2} \sum_{p=1}^P \|\text{diag}(\text{vec}(\underline{\Delta}_p)) (\text{vec}(\mathbf{Z}_p) - \mathbf{B} \text{diag}(\mathbf{e}_p^T \mathbf{C}) \mathbf{a})\|_2^2 \\ &\quad + \frac{\mu}{2} \mathbf{a} (\mathbf{I} \otimes \mathbf{R}_A^{-1}) \mathbf{a} \end{aligned} \quad (47)$$

#### VII - Proof of Lemma 1

*Proof:* Function  $g(\mathbf{A}, \bar{\mathbf{A}})$  in (27) is formed from  $f(\mathbf{A})$  after substituting  $g_1(\mathbf{A}, \bar{\mathbf{A}})$  for  $f_1(\mathbf{A})$ , and  $g_2(\mathbf{A}, \bar{\mathbf{A}})$  for  $f_2(\mathbf{A})$ , respectively, as defined by

$$f_1(\mathbf{A}) := \text{Tr}(\mathbf{A}^T \mathbf{R}_A^{-1} \mathbf{A}) \quad (48)$$

$$g_1(\mathbf{A}, \bar{\mathbf{A}}) := \lambda \text{Tr}(\mathbf{A}^T \mathbf{A}) - 2 \text{Tr}(\mathbf{\Theta}^T \mathbf{A}) + \text{Tr}(\mathbf{\Theta}^T \bar{\mathbf{A}}) \quad (49)$$



where  $\lambda := \lambda_{\max}(\mathbf{R}_A^{-1})$  and  $\Theta := \lambda \mathbf{I} - \mathbf{R}_A^{-1}$ , and

$$f_2(\mathbf{A}) := -\mathbf{1}_M \Delta \otimes \mathbf{Z} \log(\mathbf{A} \Pi^T) \mathbf{1}_{NP} \quad (50)$$

$$g_2(\mathbf{A}, \bar{\mathbf{A}}) := - \sum_{m=1}^M \sum_{k=1}^{NP} \delta_{mk} z_{mk} \alpha_{mkr} \log \left( \frac{a_{mr} \pi_{kr}}{\alpha_{mkr}} \right) \quad (51)$$

with  $\alpha_{mkr} := \bar{a}_{mr} \pi_{kr} / \sum_{r'=1}^R \bar{a}_{mr} \pi_{kr'}$ .

Hence, properties i)-iii) will be satisfied by the pair of functions  $g(\mathbf{A}, \bar{\mathbf{A}})$  and  $f(\mathbf{A})$  in Lemma 1, as long as they are satisfied both by the pair in (48)-(49) and that in (50)-(51).

Focusing on the first pair, both functions are separable per column of  $\mathbf{A}$  and  $\bar{\mathbf{A}}$ , and their difference takes the form

$$\begin{aligned} g_1(\mathbf{A}, \bar{\mathbf{A}}) - f_1(\mathbf{A}) &= \sum_{r=1}^R [\lambda \mathbf{a}_r^T \mathbf{a}_r - 2\theta_r^T \mathbf{a}_r + \theta_r^T \bar{\mathbf{a}}_r - \bar{\mathbf{a}}_r^T \mathbf{R}_A^{-1} \bar{\mathbf{a}}_r] \\ &= \sum_{r=1}^R (\mathbf{a}_r - \bar{\mathbf{a}}_r)^T (\lambda \mathbf{I} - \mathbf{R}_A^{-1}) (\mathbf{a}_r - \bar{\mathbf{a}}_r) \end{aligned}$$

which is positive and, together with its gradient, vanish at  $\bar{\mathbf{A}}$ . This establishes that properties i)-iii) are satisfied by  $g_1(\mathbf{A}, \bar{\mathbf{A}})$  and  $f_1(\mathbf{A})$ , and thus they are so for functions  $g(\mathbf{A}, \bar{\mathbf{A}})$  and  $f(\mathbf{A})$  in (20) and (19).

Considering the second pair, and expanding  $f_2(\mathbf{A})$  yields

$$f_2(\mathbf{A}) = - \sum_{m=1}^M \sum_{k=1}^{NP} \delta_{mk} z_{mk} \log \left( \sum_{r'=1}^R a_{mr'} \pi_{kr'} \right) \quad (52)$$

where the logarithm can be rewritten as (see also [10])

$$\log \left( \sum_{r'=1}^R a_{mr'} \pi_{kr'} \right) = \log \left( \sum_{r'=1}^R \alpha_{mkr'} \frac{a_{mr'} \pi_{kr'}}{\alpha_{mkr'}} \right) \quad (53)$$

$$\geq \sum_{r=1}^R \alpha_{mkr} \log \left( \sum_{r'=1}^R \frac{a_{mr'} \pi_{kr'}}{\alpha_{mkr}} \right) \quad (54)$$

and the inequality holds because of the concavity of the logarithm with an argument being a convex combination with coefficients  $\{\alpha_{mkr}\}_{r=1}^R$  summing up to one.

Since substituting (54) for (53) in (52) results in (51), it follows that  $g_2(\mathbf{A}, \bar{\mathbf{A}})$  and  $f_2(\mathbf{A})$  satisfy property iii). The proof is complete after evaluating at the pair of functions and their derivatives at  $\mathbf{A}$  to confirm that properties i) and ii) hold too.

The minimum  $a_{g,mr}^* := t_{mr} + \sqrt{t_{mr}^2 + s_{mr}}$  is obtained readily after equating to zero the derivative of the corresponding summand in (22), and selecting the nonnegative root. ■

## REFERENCES

- [1] J. Abernethy, F. Bach, T. Evgeniou, and J.P. Vert. "A new approach to collaborative filtering: Operator estimation with spectral regularization," *J. of Machine Learning Research*, vol. 10, pp.803-826, 2009.
- [2] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mrup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41-56, 2011.
- [3] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc. of the Natl. Academy of Science*, vol. 97, no. 18, pp. 10101-10106, 2000.
- [4] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, pp. 925-936, June 2010.
- [5] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, Dec. 2012.
- [6] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707-710, Oct. 2007.
- [7] J. Chen, and Y. Saad, "On the tensor SVD and the optimal low-rank orthogonal approximation of tensors," *SIAM Journal on Matrix Analysis and Applications (SIMAX)*, vol. 30, no. 4, pp. 1709-1734, 2009.
- [8] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis*, John Wiley, 2009.
- [9] J. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Lin. Alg. Applicat.*, vol. 18, no. 2, pp. 95138, 1977.
- [10] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," *SIAM Journal on Matrix Analysis and Applications*, Dec. 2012 (to appear; see also arXiv:1112.2414v3 [math.NA]).
- [11] M. Fazel, "Matrix rank minimization with applications" *PhD Thesis*, Electrical Engineering Dept., Stanford University, vol. 54, pp. 1-130, 2002.
- [12] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, pp. 1-19, 2011.
- [13] G. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 3rd Edition, Oct. 1996.
- [14] J. Håstad, "Tensor rank is NP-complete," *J. Algorithms*, vol. 11, no. 4, pp. 644-654, 1990.
- [15] Internet brain segmentation repository, "MR brain data set 657," *Center for Morphometric Analysis at Massachusetts General Hospital*, available at <http://www.cma.mgh.harvard.edu/ibsr/>.
- [16] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455-500, 2009.
- [17] M. Mardani, G. Mateos, and G. B. Giannakis, "In-network sparsity-regularized rank minimization: Algorithms and applications," *IEEE Trans. Signal Process.*, 2012 arXiv:1203.1570v1 [cs.MA].
- [18] U. Nagalakshmi et al., "The transcriptional landscape of the yeast genome defined by RNA sequencing" *Science*, vol. 320, no. 5881, pp. 1344-1349, June 2008.
- [19] M. Z. Nashed and Q. Sun, "Function spaces for sampling expansions," *Multiscale Signal Analysis and Modelling*, edited by X. Shen and A. Zayed, Lecture Notes in EE, Springer, pp. 81-104, 2012.
- [20] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Opt.*, 2012; see also arXiv:1209.2385v1 [math.OC].
- [21] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT Press, 2006.
- [22] P. Scheet, and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase," *American Journal of Human Genetics*, vol. 78, pp. 629-644, 2006.
- [23] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," *Advances in Neural Information Processing Systems*, vol. 17, pp. = 1329-1336, 2005.
- [24] J. M. F. ten Berge and N. D. Sidiropoulos, "On uniqueness in CANDECOMP/PARAFAC," *Psychometrika*, vol. 67, no. 3, pp. 399-409, 2002.
- [25] R. Tomioka, K. Hayashi, and H. Kashima, "Estimation of low-rank tensors via convex optimization," submitted 2011, also available at [ArXiv:1010.0789v2 \[stat.ML\]](http://arxiv.org/abs/1010.0789v2).
- [26] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49-95, 1996.
- [27] G. Wahba, *Spline Models for Observational Data*, SIAM, PA 1990.