Rank Regularization and Bayesian Inference for Tensor Completion and Extrapolation

Juan Andrés Bazerque, *Member, IEEE*, Gonzalo Mateos, *Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***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 -11 dB and -15 dB.**

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

I. Introduction

IMPUTATION of missing data is a basic task arising in various Big Data applications as diverse as medical imaging [15], bioinformatics [3], as well as social and computer networking [11], [24]. The key idea rendering recovery feasible is the "regularity" present in 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 the *low-rank matrix completion* problem is to minimize the rank of a target matrix, subject to a constraint on the error in fitting the observed entries [7]. Since rank minimization is generally NP-hard [39], the nuclear norm has been advocated recently as a convex surrogate to the rank [14]. Beyond tractability, nuclear-norm minimization offers desirable merits both in theory as well as in practice [7]. Several iterative solvers have been proposed in this context, and are effective for low- to medium-size matrix completion problems; see e.g., [6]. The corresponding algorithms however, require computation of a singular-value decomposition (SVD) per iteration and become prohibitively expensive when dealing with high-dimensional data. Scalable *distributed* algorithms for matrix completion were developed in [30] and [24], while real-time *online* algorithms for imputation of streaming data are also available; see e.g., [4], [12], [25].

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 [13], [21]. Leveraging the low-rank structure for tensor completion is challenging, since even computing the tensor rank is NP-hard [18]. 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 SVD [3], [10], [38], or, employing the parallel factor (PARAFAC) decomposition [22], [33], [34], [37]. In the context of tensor completion, approaches falling under the first category can be found in [15] and [38], 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 an alternative characterization of the nuclear norm based on a low-rank factorization of its matrix argument. 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 [8]. 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 [11]. Approaches to non-negative matrix and tensor factorization based on KL-divergence minimization include those in [23]

Manuscript received December 31, 2012; revised May 11, 2013; accepted August 06, 2013. Date of publication August 15, 2013; date of current version October 14, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ana Perez-Neira. This work was supported by NSF EARS Grant No. 1343248, MURI Grant No. AFOSR FA9550-10-1-0567, and NIH Grant No. 1R01GM104975-01. Parts of the paper appeared in the IEEE Workshop on Statistical Signal Processing, Ann Arbor, MI, USA, August 5–8, 2012, and in the International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, May 26–31, 2013.

The authors are with the Department of ECE and the Digital Technology Center, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: bazer002@umn.edu; mate0058@umn.edu; georgios@umn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2013.2278516

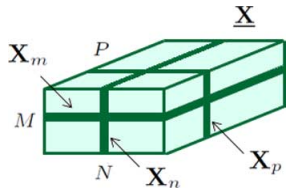

Fig. 1. Tensor slices along the row, column, and tube dimensions.

and [41]. Different from the PARAFAC decomposition algorithm for sparse count data in [11], the novel algorithm of this paper focuses on the imputation of missing count data, and allows one to incorporate prior information through rank regularization.

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 are represented by $\|\cdot\|_F$. Symbols $\otimes$, $\odot$, $\circledast$, and $\circ$, denote the Kronecker, Khatri-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 [14]. 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 $\mathbf{\Delta} \in \{0,1\}^{N\times M}$, can be formulated as

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

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 ℓ_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 [14] and surrogate the ℓ_0-norm with the ℓ_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

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

where $\mu \geq 0$ is a rank-controlling parameter. Problem (2) can be further transformed by considering the following characterization of the nuclear norm [35]

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

For arbitrary $\mathbf{X}$ with SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, the minimum in (3) is attained for $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}^{1/2}$ and $\mathbf{C} = \mathbf{V}\mathbf{\Sigma}^{1/2}$. 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.

For given R, note that the factorization $\mathbf{X} = \mathbf{B}\mathbf{C}^T$ with $\mathbf{B} \in \mathbb{R}^{N\times R}$ and $\mathbf{C} \in \mathbb{R}^{M\times R}$ implies $\text{rank}(\mathbf{X}) \leq R$. Introducing the aforementioned bilinear factorization of $\mathbf{X}$, and replacing $\|\mathbf{X}\|_*$ in (2) with the Frobenius-norm regularization dictated by (3), one arrives at the following reformulation of (2) [24]

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

Problems (2) and (4) can be readily proved equivalent [cf. Proposition 1-a)], in the sense 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 that need not be globally optimal. Interestingly, the next result provides global optimality conditions for these stationary points [parts a) and b) are proved in Appendix A, while the proof for c) can be found in [24].]

Proposition 1: If $R \geq \text{rank}(\hat{\mathbf{Z}})$*, then 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* $\|(\mathbf{X}-\mathbf{Z})\circledast\mathbf{\Delta}\|_2 \leq \mu$ *are globally optimal.*

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.

Remark 1: Without missing data, all entries of $\mathbf{\Delta}$ are equal to one, and (1) boils down to principal component analysis. In this case, (1) can be solved by truncating the SVD of $\mathbf{Z}$, so that only its R largest singular values are retained. The presence of missing entries changes the problem profoundly, as (1) becomes NP-hard [39]. This highlights the importance of the nuclear norm regularizer (2) as a clever alternative to rank minimization in the presence of missing data. Reduced complexity alternatives to SVD are also available; e.g., the truncated multi-stage Wiener filter (MSWF) [16]. MSWF offers an attractive alternative to (1) for matrix (and even tensor) dimensionality reduction. This approach is not pursued here however, since redesigning the MSWF to cope with missing data may prove challenging [cf. (1) with and without missing data.] Conversely, exploring variants of (2) for reduced-rank Wiener filtering in the presence of missing data, constitutes an interesting direction for future research.

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 [22], [37]. 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

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

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 $\underline{\mathbf{O}}_r := \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \in \mathbb{R}^{M\times N\times P}$ have entries $\underline{\mathbf{O}}_r(m,n,p) := \mathbf{A}(m,r)\mathbf{B}(n,r)\mathbf{C}(p,r)$. 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

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

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 $\gamma_r := \|\mathbf{a}_r\|\|\mathbf{b}_r\|\|\mathbf{c}_r\|$, $r = 1,\ldots,R$.

Let $\mathbf{X}_p$, $p = 1,\ldots,P$ 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

$$\mathbf{X}_p = \mathbf{A}\text{diag}\left[\mathbf{e}_p^T\mathbf{C}\right]\mathbf{B}^T, \quad p = 1,\ldots,P \tag{7}$$

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 $\mathbf{X}_m = \mathbf{B}\text{diag}\left[\mathbf{e}_m^T\mathbf{A}\right]\mathbf{C}^T$, or, $\mathbf{X}_n = \mathbf{C}\text{diag}\left[\mathbf{e}_n^T\mathbf{B}\right]\mathbf{A}^T$ in terms of slices along the first (row), or, second (column) dimensions. Given $\underline{\mathbf{X}}$, under some technical conditions then $\{\mathbf{A},\mathbf{B},\mathbf{C}\}$ are unique up to a common column permutation and scaling (meaning PARAFAC is identifiable); see e.g., [22], [32], [36], [37].

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 [21]. 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

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

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 $(\underline{\mathbf{X}}\circledast\underline{\mathbf{\Delta}})(m,n,p) := \underline{\mathbf{X}}(m,n,p)\underline{\mathbf{\Delta}}(m,n,p)$.

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 μ. In this direction, consider an alternative completion problem stated in terms of the normalized tensor representation (6)

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

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

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

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

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

For any value of σ^2 there exists a corresponding Lagrange multiplier λ 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 $\|\boldsymbol{\gamma}\|_{2/3}$ can be safely eliminated without affecting the minimizer of (10).] The key observation is that minimizing $\|\boldsymbol{\gamma}\|_{2/3}$ in (10) yields a sparse vector $\boldsymbol{\gamma}$ [9]. As with the well-known sparsity-promoting ℓ_1-norm, the unit $\ell_{2/3}$-norm ball exhibits a "pointy geometry" at the axes responsible for inducing sparsity; see Fig. 2.

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

The next property is a direct consequence of the low-rank promoting property of (8) as established in Proposition 2; see Appendix C for a proof.

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

Corollary 1 asserts that if μ 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 linear models and ridge-regression estimates which entail similar quadratic regularizers, but an analogous property does not hold. In ridge regression one needs to let $\mu \to \infty$ in order to obtain an all-zero

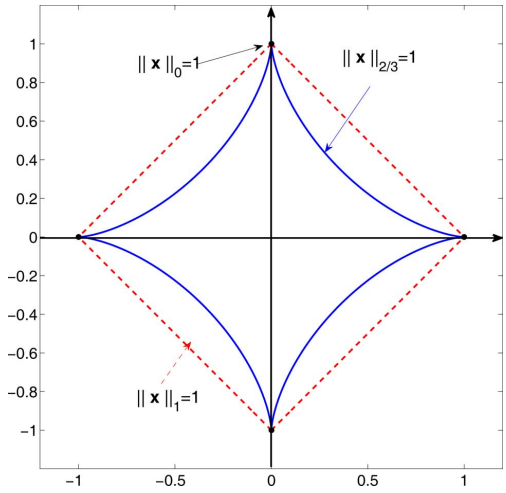

Fig. 2. The unit $\ell_{2/3}$-norm ball compared to its ℓ_0- and ℓ_1-norm counterparts.

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 [8]. Indeed, the infimum ℓ_1-norm of $\boldsymbol{\gamma}$ is a proper norm for $\underline{\mathbf{X}}$, named atomic norm, and denoted by $\|\underline{\mathbf{X}}\|_{\mathcal{A}} := \|\boldsymbol{\gamma}\|_1$ [8]. Thus, by replacing $\|\boldsymbol{\gamma}\|_{2/3}$ 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 [8]. 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 $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ 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. It is also fair to say that only convergence to a stationary point of (8) will be established in this paper.

Remark 2: These insights foster future research directions for the design of a convex regularizer of the tensor rank. Specifically, substituting $\rho(\mathbf{A},\mathbf{B},\mathbf{C}) := \sum_{r=1}^{R}(\|\mathbf{a}_r\|^3 + \|\mathbf{b}_r\|^3 + \|\mathbf{c}_r\|^3)$ for the regularization term in (8), turns $\|\boldsymbol{\gamma}\|_{2/3}$ into $\|\boldsymbol{\gamma}\|_1 = \|\underline{\mathbf{X}}\|_{\mathcal{A}}$ 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 for the tensor case, providing conditions for global optimality of the stationary points of (8).

Remark 3: Feasibility of the imputation task relies fundamentally on assuming a low-dimensional data model, to couple the available and missing entries as well as reduce the effective degrees of freedom in the problem. Under the low-rank assumption $rank(\bar{\mathbf{X}}) = R < \bar{R}$ for instance, a rough idea on the fraction p_m of missing data that can be afforded is obtained by comparing the number of unknowns $R(M+N+P)$ with the number of available data samples (equations) $(1-p_m)MNP$. Ensuring that $(1-p_m)MNP \geq R(M+N+P)$, implies that the tensor can be recovered even if a fraction $p_m \leq 1 - R(M+N+P)/(MNP)$ of entries is missing. Different low-dimensional models would lead to alternative imputation methods, as the unfolded tensor regularization in [15], or the truncated MSWF [16] discussed in Remark 1. The comparative performance of these methods would depend on the accuracy of their modeling assumptions. This paper focuses on low-rank tensors, hence (8) is expected to outperform its competitors. This intuition is corroborated by numerical tests in Section VI.

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 $\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}$.

Remark 4: 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 [31]. These correlations could be available either through a prescribed model, or, through estimates obtained using a reference tensor $\check{\underline{\mathbf{Z}}}$. 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 the noisy tensor measurements $z_{mnp} := \underline{\mathbf{Z}}(m,n,p)$ are given by

$$z_{mnp} = x_{mnp} + e_{mnp}, \quad e_{mnp} \sim \mathcal{N}(0,\sigma^2), \text{ i.i.d..} \tag{11}$$

Since vectors $\mathbf{a}_r$ in (6) are interchangeable, identical distributions are assigned across $r = 1,\ldots,R$, 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. 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,

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

where $\text{Tr}(\cdot)$ denotes the matrix trace operator.

Under these assumptions, the posterior distribution $p(\mathbf{A},\mathbf{B},\mathbf{C}|\underline{\mathbf{Z}})$ can be factorized as $p(\underline{\mathbf{Z}}|\mathbf{A},\mathbf{B},\mathbf{C})p(\mathbf{A})p(\mathbf{B})p(\mathbf{C})/P(\underline{\mathbf{Z}})$ and is thus proportional to $\exp(-L(\underline{\mathbf{X}},\mathbf{A},\mathbf{B},\mathbf{C}))$, where

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

and with $\underline{\mathbf{X}} := \sum_{r=1}^{R}\mathbf{a}_r\circ\mathbf{b}_r\circ\mathbf{c}_r$ as in (5).

The MAP estimator of $(\mathbf{A},\mathbf{B},\mathbf{C})$ is defined as the maximizer of $p(\mathbf{A},\mathbf{B},\mathbf{C}|\underline{\mathbf{Z}})$ [20, p. 350]. Equivalently, the MAP estimator of $\underline{\mathbf{X}}$ follows from minimizing $L(\underline{\mathbf{X}},\mathbf{A},\mathbf{B},\mathbf{C})$ w.r.t. $\underline{\mathbf{X}}$, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, with (5) as a constraint; i.e.,

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

reducing to (8) when $\mathbf{R}_A=\mathbf{I}_M$, $\mathbf{R}_B=\mathbf{I}_N$, and $\mathbf{R}_C=\mathbf{I}_P$.

Remark 5: From this Bayesian vantage point, the regularization parameter μ [cf. (8)] can be interpreted as the noise variance, which is useful in practice to select μ. This parameter choice is complemented by the guidelines to obtain the prior covariances which are outlined in Section IV-C.

First, 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 [40]. 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 [27].

In this context, it is instructive to look at a tensor $f:\mathcal{M}\times\mathcal{N}\times\mathcal{P}\to\mathbb{R}$ 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

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

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 $\{z_{mnp}:\delta_{mnp}=1\}$

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

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

$$\begin{aligned}&\hat{f}_R(m,n,p) \\ &= \boldsymbol{k}_{\mathcal{M}}^T(m)\mathbf{K}_{\mathcal{M}}^{-1}\mathbf{A}\mathrm{diag}\left[\boldsymbol{k}_{\mathcal{P}}^T(p)\mathbf{K}_{\mathcal{P}}^{-1}\mathbf{C}\right]\mathbf{B}^T\mathbf{K}_{\mathcal{N}}^{-1}\boldsymbol{k}_{\mathcal{N}}(n)\end{aligned} \tag{15}$$

where vector $\mathbf{k}_{\mathcal{M}}^T(m) := [k_{\mathcal{M}}(m,1),\ldots,k_{\mathcal{M}}(m,M)]$, $m\in\mathcal{M}$, and matrix $\mathbf{K}_{\mathcal{M}}$ has entries $k_{\mathcal{M}}(m,m')$, $m,m'=1,\ldots,M$. Likewise, $\mathbf{k}_{\mathcal{N}}(n)$, $\mathbf{K}_{\mathcal{N}}$, $\mathbf{k}_{\mathcal{P}}(p)$, and $\mathbf{K}_{\mathcal{P}}$ are correspondingly defined in terms of $k_{\mathcal{N}}$ and $k_{\mathcal{P}}$. It is also shown in Appendix D that the coefficient matrices $\mathbf{A}\in\mathbb{R}^{M\times R}$, $\mathbf{B}\in\mathbb{R}^{N\times R}$, and $\mathbf{C}\in\mathbb{R}^{P\times R}$ in (15) can be found by solving

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{C}} & \sum_{p=1}^{P}\left\|\left(\mathbf{Z}_p-\mathbf{A}\mathrm{diag}\left[\mathbf{e}_p^T\mathbf{C}\right]\mathbf{B}^T\right)\circledast\mathbf{\Delta}_p\right\|_F^2 \\ & +\frac{\mu}{2}\left(\mathrm{Tr}(\mathbf{A}^T\mathbf{K}_{\mathcal{M}}^{-1}\mathbf{A})+\mathrm{Tr}(\mathbf{B}^T\mathbf{K}_{\mathcal{N}}^{-1}\mathbf{B})+\mathrm{Tr}(\mathbf{C}^T\mathbf{K}_{\mathcal{P}}^{-1}\mathbf{C})\right).\end{aligned} \tag{16}$$

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 completion, where in general, missing entries are to be imputed 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 $\mathbf{K}_{\mathcal{M}}$, $\mathbf{K}_{\mathcal{N}}$ and $\mathbf{K}_{\mathcal{P}}$, (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 $\mathbf{K}_{\mathcal{M}}$ 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 $k_{\mathcal{M}}(M+1)$ with the available entries. These extra capabilities will be exploited in Section VI-C, 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 propositions throughout this paper.

Identification of covariance matrices $\mathbf{R}_A$, $\mathbf{R}_B$, and $\mathbf{R}_C$ with kernel matrices $\mathbf{K}_{\mathcal{M}}$, $\mathbf{K}_{\mathcal{N}}$ and $\mathbf{K}_{\mathcal{P}}$ 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 described in Section IV-A, 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 [28], 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 $\mathbf{K}_{\mathcal{P}}(p',p) := \mathbb{E}[\mathrm{Tr}(\mathbf{X}_{p'}^T\mathbf{X}_p)]$, that is, the covariance between slices $\mathbf{X}_{p'}$ and $\mathbf{X}_p$ taking $\langle \mathbf{X},\mathbf{Y}\rangle := \mathrm{Tr}(\mathbf{X}^T\mathbf{Y})$ as the standard inner product in the matrix space. Under this alternative definition for $\mathbf{K}_{\mathcal{P}}$, and corresponding definitions for $\mathbf{K}_{\mathcal{N}}$, and $\mathbf{K}_{\mathcal{M}}$, it is shown in Appendix E that

$$\mathbf{K}_{\mathcal{M}} = \theta^2\mathbf{R}_A, \quad \mathbf{K}_{\mathcal{N}} = \theta^2\mathbf{R}_B, \quad \mathbf{K}_{\mathcal{P}} = \theta^2\mathbf{R}_C \tag{17}$$

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

$$\mathbb{E}[\|\underline{\mathbf{X}}\|_F^2] = R\theta^3. \tag{18}$$

Since sample estimates for $\mathbf{K}_{\mathcal{M}}$, $\mathbf{K}_{\mathcal{N}}$, $\mathbf{K}_{\mathcal{P}}$, 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; see also the numerical tests in Section VI-C.

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 prescribed function of a few attributes, such as age or income [1].

D. Block Successive Upper-Bound Minimization Algorithm

An iterative algorithm is developed here for solving (13), by cyclically minimizing the cost over $\mathbf{A} \rightarrow \mathbf{B} \rightarrow \mathbf{C}$. This alternating-minimization procedure is typically adopted to fit PARAFAC models, and is also known as block-coordinate descent (BCD) in the optimization parlance [29]. 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 fixed parameters taking on their previous iteration values. Accordingly, the partial cost to minimize reduces to the convex function

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

where μ was identified with and substituted for σ^2. Function (19) is quadratic in $\mathbf{A}$ and can be readily minimized after re-writing it in terms of $\mathbf{a} := \mathrm{vec}(\mathbf{A})$. 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 square linear system of MR equations. The alternative pursued here relies on the so-called block successive upper-bound minimization (BSUM) algorithm [29]. As it will become clear later on, this way the computational complexity in updating $\mathbf{A}$ is reduced from $\mathcal{O}((MR)^3)$ to $\mathcal{O}(MR^3)$ per iteration, and likewise for $\mathbf{B}$ and $\mathbf{C}$.

BSUM follows the same cyclic architecture as BCD, but one instead minimizes a judiciously chosen upper-bound $g(\mathbf{A},\bar{\mathbf{A}})$ of $f(\mathbf{A})$. As such, it blends the properties of BCD and majorization-minimization algorithms. The majorizing function $g(\mathbf{A},\bar{\mathbf{A}})$ depends on the current iterate $\bar{\mathbf{A}}$, and should be crafted such that: i) it is simpler to optimize than $f(\mathbf{A})$; and ii) satisfies certain local-tightness conditions; see also [29] and properties i)-iii) in Lemma 1.

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

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

where $\lambda := \lambda_{\max}(\mathbf{R}_A^{-1})$ is the maximum eigenvalue of $\mathbf{R}_A^{-1}$, and $\mathbf{\Theta} := (\lambda\mathbf{I} - \mathbf{R}_A^{-1})\bar{\mathbf{A}}$. 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 (see Appendix F for a proof).

Lemma 1: Function $g(\mathbf{A},\bar{\mathbf{A}})$ *in* (20) 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) $f(\mathbf{A}) \leq g(\mathbf{A},\bar{\mathbf{A}})$, $\forall\mathbf{A}$.

The computational advantage of minimizing $g(\mathbf{A},\bar{\mathbf{A}})$ instead of $f(\mathbf{A})$ comes from the separability of $g(\mathbf{A},\bar{\mathbf{A}})$ across rows of $\mathbf{A}$. To appreciate this, consider the Khatri-Rao product $\mathbf{\Pi} := \mathbf{C} \odot \mathbf{B} := [\mathbf{c}_1 \otimes \mathbf{b}_1, \ldots \mathbf{c}_R \otimes \mathbf{b}_R]$, defined by the column-wise Kronecker products $\mathbf{c}_r \otimes \mathbf{b}_r$. Let also matrix $\mathbf{Z} := [\mathbf{Z}_1,\ldots,\mathbf{Z}_P] \in \mathbb{N}^{M\times NP}$ denote the mode-1 unfolding of $\underline{\mathbf{Z}}$ (along its tube dimension; see e.g., [13, p.30],) and likewise for $\mathbf{\Delta} := [\mathbf{\Delta}_1,\ldots,\mathbf{\Delta}_P] \in \{0,1\}^{M\times NP}$ and $\mathbf{X} := [\mathbf{X}_1,\ldots,\mathbf{X}_P] \in \mathbb{R}_+^{M\times NP}$. Using the following identity that relates the unfolded tensor with its factors [11]

$$\mathbf{X} := [\mathbf{X}_1,\ldots,\mathbf{X}_P] = \mathbf{A}\mathbf{\Pi}^T \tag{21}$$

it is possible to rewrite (20) as

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

which can be decomposed as

$$g(\mathbf{A},\bar{\mathbf{A}}) = \sum_{m=1}^{M}\left[\frac{1}{2}\|\mathrm{diag}(\boldsymbol{\delta}_m)(\mathbf{z}_m - \mathbf{\Pi}\mathbf{a}_m)\|_2^2 + \mu\left((\lambda/2)\|\mathbf{a}_m\|^2 - \boldsymbol{\theta}_m^T\mathbf{a}_m + \boldsymbol{\theta}_m^T\bar{\mathbf{a}}_m\right)\right] \tag{22}$$

where $\mathbf{z}_m^T$, $\mathbf{a}_m^T$, $\boldsymbol{\delta}_m^T$, $\boldsymbol{\theta}_m^T$, and $\bar{\mathbf{a}}_\mathbf{m}^T$, 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 their gradients to zero, namely (define $\mathbf{D}_m := \mathrm{diag}(\boldsymbol{\delta}_m)$ for convenience)

$$(\mathbf{\Pi}^T\mathbf{D}_m\mathbf{\Pi} + \lambda\mu\mathbf{I})\mathbf{a}_m - \mathbf{\Pi}^T\mathbf{D}_m\mathbf{z}_m - \mu\boldsymbol{\theta} = \mathbf{0},\ m = 1,\ldots,M.$$

Accordingly, the majorization strategy reduces the computational load to M systems of R equations that can be solved in parallel, where R is typically small (cf. the low tensor rank assumption). 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

Algorithm 1: Low-rank tensor imputation (LRTI)

function UPDATE_FACTOR($\mathbf{A}, \mathbf{R}, \mathbf{\Pi}, \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu$)
  Set $\lambda = \lambda_{\max}(\mathbf{R}^{-1})$
  Unfold $\underline{\mathbf{\Delta}}$ and $\underline{\mathbf{Z}}$ over dimension of $\mathbf{A}$ into $\mathbf{\Delta}$ and $\mathbf{Z}$
  Set $\mathbf{\Theta} = (\lambda\mathbf{I} - \mathbf{R}^{-1})\mathbf{A}$
  for $m = 1, \ldots, M$ **do**
    Select rows $\mathbf{z}_m^T, \boldsymbol{\delta}_m^T$, and $\boldsymbol{\theta}_m^T$, and set $\mathbf{D}_m = \text{diag}(\boldsymbol{\delta}_m)$
    Compute $\mathbf{a}_m = (\mathbf{\Pi}^T\mathbf{D}_m\mathbf{\Pi} + \lambda\mu\mathbf{I})^{-1}(\mathbf{\Pi}^T\mathbf{D}_m\mathbf{z}_m + \mu\boldsymbol{\theta}_m)$
    Update $\mathbf{A}$ with row $\mathbf{a}_m^T$
  end for
  return $\mathbf{A}$
end function
Initialize $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ randomly.
while $|\text{cost} - \text{cost\_old}| < \epsilon$ **do**
  $\mathbf{A} = \text{UPDATE\_FACTOR}(\mathbf{A}, \mathbf{R}_A, (\mathbf{C} \odot \mathbf{B}), \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu)$
  $\mathbf{B} = \text{UPDATE\_FACTOR}(\mathbf{B}, \mathbf{R}_B, (\mathbf{A} \odot \mathbf{C}), \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu)$
  $\mathbf{C} = \text{UPDATE\_FACTOR}(\mathbf{C}, \mathbf{R}_C, (\mathbf{B} \odot \mathbf{A}), \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu)$
  Recalculate cost in **(13)**
end while
return $\hat{\underline{\mathbf{X}}}$ with slices $\hat{\mathbf{X}}_\mathbf{p} = \mathbf{A}\text{diag}[\mathbf{e}_p^T\mathbf{C}]\mathbf{B}^T$

presented in Algorithm 1, where analogous updates for $\mathbf{B}$ and $\mathbf{C}$ are carried out cyclically per iteration.

Remark 6: A different algorithm for solving (13) was put forth in the conference precursor of this paper [5], which cyclically minimizes the columns of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. Distinct from Algorithm 1 that entails *parallel* row-wise updates per factor, iterates in [5] involve *sequential* updates across columns and factors, thus incurring a per iteration complexity of $\mathcal{O}(R(M^3 + N^3 + P^3))$. Because the factor matrices are tall $[\min(M, N, P) \gg R]$, the aforementioned computational load is markedly higher than the one incurred by Algorithm 1, namely $\mathcal{O}((M + N + P)R^3)$.

By virtue of properties i)-iii) in Lemma 1, convergence of Algorithm 1 follows readily from that of the BSUM algorithm [29].

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 conditioned on $\underline{\mathbf{X}}$. 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 [11]. Such a model is also well-suited for sparse tensor data, since the Poisson distribution has mass at the origin.

Suppose that the entries z_{mnp} of $\underline{\mathbf{Z}}$ are Poisson distributed, with probability mass function

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

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

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

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 [cf. (8)] as a more suitable criterion [11]. 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 $\underline{\mathbf{\Delta}}$) becomes

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

over the feasible set

$$\mathcal{T} := \{\underline{\mathbf{X}}, \mathbf{A}, \mathbf{B}, \mathbf{C} : \mathbf{A} \geq \mathbf{0}, \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \mathbf{X}_p = \mathbf{A}\text{diag}\left[\mathbf{e}_p^T\mathbf{C}\right]\mathbf{B}^T,\ p = 1, \ldots, P\},$$

where the symbol $\geq$ should be understood to imply entry-wise nonnegativity.

Remark 7: The parameter μ in (25) was introduced to add flexibility in varying the sparsity level of $\hat{\underline{\mathbf{Z}}}$. However, derivation of the Poisson MAP estimator with Gaussian priors leads to $\mu = 1$, which is used as the default value in the applications of Section VI and is corroborated to be a reasonable choice in Fig. 4. The reason behind μ taking a specific value is that in the Poisson distribution (23) the mean and variance are related (in fact they are equal). This should be contrasted with the MAP estimator in Section IV, where μ equals σ^2 which is a free parameter under the Gaussian data model (11).

With the aid of the Representer 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. BSUM Algorithm

A K-L counterpart of the LRTI algorithm is developed in this section, that provably converges to a stationary point of (25). An alternating-minimization scheme is adopted once again, which

optimizes (a suitable upper-bound of) (25) cyclically 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 $\mathbf{Z} := [\mathbf{Z}_1, \ldots, \mathbf{Z}_P] \in \mathbb{N}^{M \times NP}$ denote the mode-1 unfolding of $\underline{\mathbf{Z}}$, and likewise for $\mathbf{\Delta} := [\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_P] \in \{0,1\}^{M \times NP}$ and $\mathbf{X} := [\mathbf{X}_1, \ldots, \mathbf{X}_P] \in \mathbb{R}_+^{M \times NP}$. Based on these definitions, (24) can be written as

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

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 Khatri-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

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

A closed-form minimizer $\mathbf{A}^\star$ for $f(\mathbf{A})$ is not available, but since $f(\mathbf{A})$ is convex one could in principle resort to an iterative procedure to obtain $\mathbf{A}^\star$. To avoid extra inner iterations, the approach here relies again on the BSUM algorithm [29].

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

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

where $\lambda := \lambda_{\max}(\mathbf{R}_A^{-1})$, and parameters s_{mr}, t_{mr}, and u_{mr} are defined in terms of $\bar{\mathbf{A}}$, $\mathbf{Z}$, $\mathbf{\Delta}$, $\mathbf{\Pi}$, and $\mathbf{\Theta} := \left(\lambda\mathbf{I} - \mathbf{R}_A^{-1}\right)\bar{\mathbf{A}}$ 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 $u_{mr} := \frac{1}{\lambda\mu}\left(\theta_{mr}\bar{a}_{mr} + \sum_{k=1}^{NP} \delta_{mk} z_{mk} \bar{a}_{mr} \pi_{kr} \upsilon_{mrk}\right)$, with $\upsilon_{mrk} := \log(\bar{a}_{mr}\pi_{kr} / \sum_{r'=1}^{R} \bar{a}_{mr'}\pi_{kr}) / \sum_{r'=1}^{R} \bar{a}_{mr'}\pi_{kr'}$. 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 G for a proof.)

Lemma 2: Function $g(\mathbf{A}, \bar{\mathbf{A}})$ in (27) *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 $\mathbf{A} = \mathbf{A}_g^\star$ with entries $a_{g,mr}^\star := t_{mr} + \sqrt{t_{mr}^2 + s_{mr}}$.

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

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

function UPDATE_FACTOR($\mathbf{A}, \mathbf{R}, \mathbf{\Pi}, \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu$)
  Set $\lambda = \lambda_{\max}(\mathbf{R}^{-1})$
  Unfold $\underline{\mathbf{\Delta}}$ and $\underline{\mathbf{Z}}$ over dimension of $\mathbf{A}$ into $\mathbf{\Delta}$ and $\mathbf{Z}$
  Compute $\mathbf{S} = \frac{\mathbf{A}}{\lambda\mu} \circledast \left(\frac{\mathbf{\Delta} \circledast \mathbf{Z}}{\mathbf{A}\mathbf{\Pi}^T}\mathbf{\Pi}\right)$ (element-wise division)
  Compute $\mathbf{T} = \frac{1}{2\lambda\mu}\left(\mu(\lambda\mathbf{I} - \mathbf{R}^{-1})\mathbf{A} - \mathbf{\Delta}\mathbf{\Pi}\right)$
  Update $\mathbf{A}$ with entries $a_{mr} = t_{mr} + \sqrt{t_{mr}^2 + s_{mr}}$
  return $\mathbf{A}$
end function
Initialize $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ randomly.
while $|\mathrm{cost} - \mathrm{cost\_old}| < \epsilon$ **do**
  $\mathbf{A} = $ UPDATE_FACTOR$(\mathbf{A}, \mathbf{R}_A, (\mathbf{C} \odot \mathbf{B}), \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu)$
  $\mathbf{B} = $ UPDATE_FACTOR$(\mathbf{B}, \mathbf{R}_B, (\mathbf{A} \odot \mathbf{C}), \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu)$
  $\mathbf{C} = $ UPDATE_FACTOR$(\mathbf{C}, \mathbf{R}_C, (\mathbf{B} \odot \mathbf{A}), \underline{\mathbf{\Delta}}, \underline{\mathbf{Z}}, \mu)$
  Recalculate cost in **(25)**
end while
return $\hat{\underline{\mathbf{X}}}$ with slices $\hat{\mathbf{X}}_{\mathbf{p}} = \mathbf{A}\mathrm{diag}(\mathbf{e}_p^T\mathbf{C})\mathbf{B}^T$

By virtue of properties i)–iii) in Lemma 2, convergence of Algorithm 2 follows readily from the general convergence theory available for the BSUM algorithm [29].

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

A related algorithm, abbreviated as CP-APR can be found in [11], 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 [11] – 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 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.

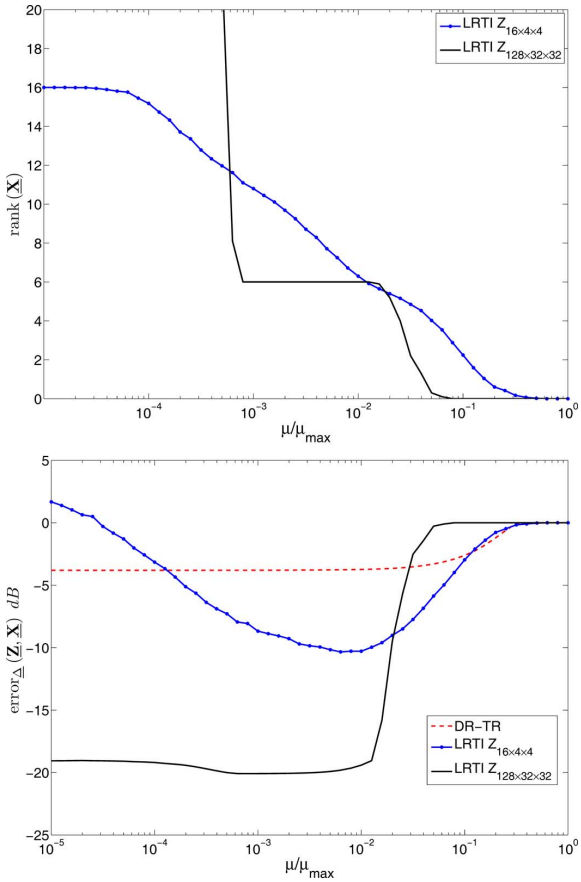

Fig. 3. Performance of the low-rank tensor imputation method as a function of μ; (top) rank of the tensor as recovered by (8) averaged over 100 test repetitions, compared to the DR-TR algorithm in [15]; (bottom) relative recovery error.

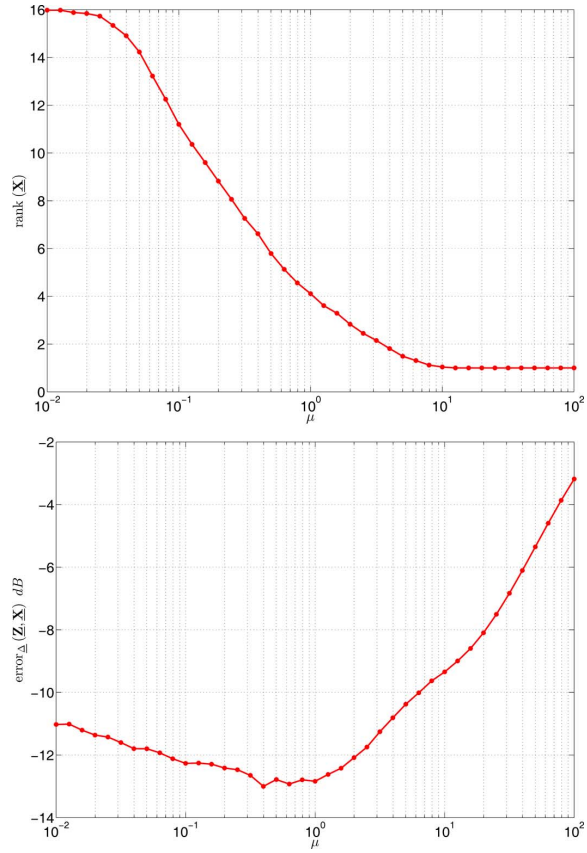

Fig. 4. Performance of the low-rank Poisson imputation method as function of the regularizing parameter μ; (top) rank of the recovered tensor averaged over 100 test repetitions, (bottom) relative recovery error.

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 μ varying on the interval $10^{-5}\mu_{\max}$ to $\mu_{\max}$, which is computed as in Corollary 1. The blue dotted line in Fig. 3 (bottom) shows that the LRTI algorithm is successful in recovering the missing entries of $\underline{\mathbf{Z}}$ up to -10 dB for a wide range of values of μ, presenting a minimum at $\mu = 10^{-2}\mu_{\max}$. This result is consistent with Fig. 3 (top, blue dotted line), 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 μ is increased, and confirms that the recovered tensor is null at $\mu_{\max}$ as asserted by Corollary 1.

Fig. 3 (bottom) also depicts the imputation error that results from applying the Douglas-Rachford (DR-TR) method for tensor recovery in [15]. Since the DR-TR method is not designed to capture the PARAFAC rank, the LRTI offers better performance in terms of recovery error when $\underline{\mathbf{Z}}$ indeed abides to a low-rank model. In addition, Fig. 3 depicts the LRTI results obtained for a larger tensor $\underline{\mathbf{Z}}$ of dimensions $M = 128$, $N = 32$, $P = 32$, and rank $R = 6$. Similar to the prior simulation setting where $M = 16$, $N = 4$, and $P = 4$, the minimum error is again attained at a similar value of $\mu/\mu_{\max}$, where the true rank is recovered.

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 $\underline{\mathbf{Z}}$ consist of realizations of Poisson random variables generated according to (23), with means specified by entries of $\underline{\mathbf{X}}$. Tensor $\underline{\mathbf{X}}$ 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 $\underline{\mathbf{Z}}$ 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 μ 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

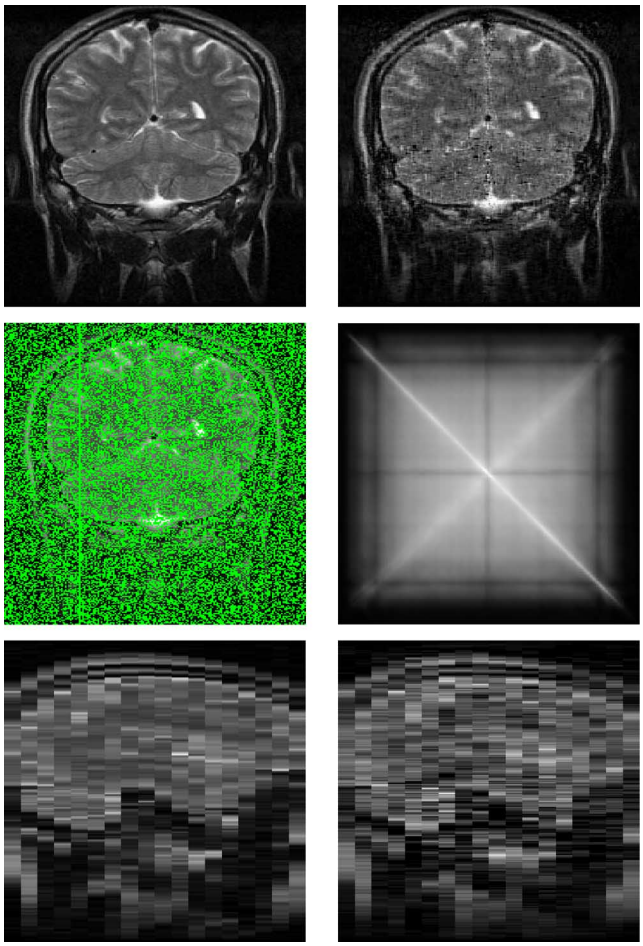

Fig. 5. Results of applying (14) to the MRI brain data set 657 [19]. (top) Original and recovered fibers $\mathbf{Z}_p$ and $\hat{\mathbf{Z}}_p$ for $p = 5$. (center) Input fiber $\mathbf{Z}_p,\ p = 5$ with missing data, and covariance matrix $\mathbf{K}_{\mathcal{N}}$. (bottom) Original and recovered columns $\mathbf{Z}_n$ and $\hat{\mathbf{Z}}_n$ for the position $n = 50$ in which the whole input slice is missing).

(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 $\underline{\hat{\mathbf{Z}}}$.

C. MRI Data

Estimator (14) was tested against a corrupted version of the MRI brain data set 657 from the Internet brain segmentation repository [19]. The tensor $\underline{\mathbf{Z}}$ 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 -11.49 dB. The original slice $\mathbf{Z}_p,\ p = 5$, its corrupted counterpart, and the resulting estimate are shown on top and center left.

Parameter μ is set equal to σ^2 as per Remark 5. The noise variance is estimated from 150 entries at each corner of $\mathbf{Z}_p,\ p = 1, \ldots, P$, which are assumed to contain background noise only. Covariance matrices $\mathbf{K}_{\mathcal{M}}$, $\mathbf{K}_{\mathcal{N}}$ and $\mathbf{K}_{\mathcal{P}}$ are estimated from six additional tensor samples containing complementary scans of the brain also available at [19]. Fig. 5 (center right) represents the covariance matrix $\mathbf{K}_{\mathcal{N}}$ 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.

For $\mu = \sigma^2$, $\mathrm{rank}(\underline{\hat{\mathbf{X}}}) = R = 100$, i.e., the rank is not reduced but remains equal to the number of columns R set for $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. The results are weakly dependent on the selection of R, with a reconstruction error in the interval $[-10.50, -12.66]$ dB for R between 50 and 200. If μ is increased the rank of the estimated tensor is reduced, but the recovery error is increased. For instance, selecting $\mu = 0.01\mu_{\max}$ as suggested by the simulation studies of Section VI-A, results in $\mathrm{rank}(\underline{\hat{\mathbf{X}}}) = 69 < R$, but the recovery error slightly increases to -10.51 dB. Further increasing μ up to 0.1 $\mu_{\max}$, results in a lower $\mathrm{rank}(\underline{\hat{\mathbf{X}}}) = 14 < R$, with a larger error at -7.8 dB. It is thus noticed that (14) is able to regularize the tensor taking into account correlations, but without necessarily forcing a reduced rank.

These properties are further appreciated when comparing the performance of LRTI with state-of-the-art methods for tensor completion. The missing entries of $\underline{\mathbf{Z}}$ were imputed via the CP-WOPT algorithm in the Tensor Toolbox 2.5 [2]. CP-WOPT was run 100 times with candidate values for the rank between 1 and 100, yielding higher reconstruction errors in the interval $[0, -5.98]$ dB.

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 [13], 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 [26] 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 [26] 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 parameter μ is set equal to 1 as per Remark 7, resulting in $\mathrm{rank}(\underline{\hat{\mathbf{X}}}) = NP = 6$ and a recovery error of -15 dB.

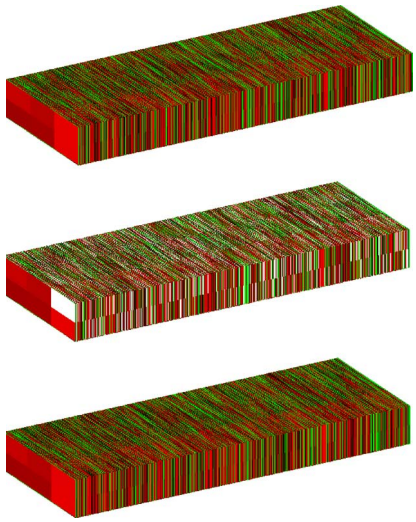

Fig. 6. Imputation of sequencing count data via LRPTI; (top) original data; (center) data with missing entries; (bottom) recovered tensor.

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 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

A. Proof of Proposition 1

The equivalence between (2) and (4) stated in a) follows immediately from (3). Indeed, if (4) is minimized in two steps

$$\min_{\mathbf{X}} \left\{ \min_{\substack{\mathbf{B},\mathbf{C} \\ \text{s. to } \mathbf{B}\mathbf{C}^T=\mathbf{X}}} \frac{1}{2}\|(\mathbf{Z}-\mathbf{X})\circledast\boldsymbol{\Delta}\|_F^2+\frac{\mu}{2}(\|\mathbf{B}\|_F^2+\|\mathbf{C}\|_F^2) \right\} \tag{28}$$

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

$$\min_{\mathbf{X}} \left\{ \frac{1}{2}\|(\mathbf{Z}-\mathbf{X})\circledast\boldsymbol{\Delta}\|_F^2+\mu \left[\min_{\substack{\mathbf{B},\mathbf{C} \\ \text{s. to } \mathbf{B}\mathbf{C}^T=\mathbf{X}}} \frac{1}{2}(\|\mathbf{B}\|_F^2+\|\mathbf{C}\|_F^2) \right] \right\} \tag{29}$$

and by recognizing (3) as the problem within the square brackets in (29), the equivalence follows.

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

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

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 $U(\mathbf{B},\mathbf{C}) \neq U(\bar{\mathbf{B}},\bar{\mathbf{C}})$. 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

$$\begin{aligned} dU = \mathrm{Tr}\left[\left(\boldsymbol{\Delta}\circledast(\mathbf{Z}-\bar{\mathbf{X}})\right)\left(\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\right)\right] + \frac{1}{2}\|\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\|_F^2 \\ +\frac{\mu}{2}\left(\|\mathbf{B}\|_F^2-\|\bar{\mathbf{B}}\|_F^2+\|\mathbf{C}\|_F^2-\|\bar{\mathbf{C}}\|_F^2\right) < 0. \end{aligned} \tag{30}$$

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

$$\mathbf{Q} := \begin{bmatrix} \mathbf{B} \\ \mathbf{C} \end{bmatrix} [\,\mathbf{B}^T \quad \mathbf{C}^T\,] = \begin{pmatrix} \mathbf{B}\mathbf{B}^T & \mathbf{X} \\ \mathbf{X}^T & \mathbf{C}\mathbf{C}^T \end{pmatrix} \tag{31}$$

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

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

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

$$\mathbf{Q}_\theta := \begin{bmatrix} \mathbf{B}_\theta \\ \mathbf{C}_\theta \end{bmatrix} [\,\mathbf{B}'_\theta \quad \mathbf{C}'_\theta\,] = \begin{pmatrix} \mathbf{B}_\theta\mathbf{B}'_\theta & \mathbf{X}_\theta \\ \mathbf{X}'_\theta & \mathbf{C}_\theta\mathbf{C}'_\theta \end{pmatrix} \tag{33}$$

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

Expanding the cost difference dU_θ as in (30) results in

$$\begin{aligned} dU_\theta &:= U(\mathbf{B}_\theta,\mathbf{C}_\theta) - U(\bar{\mathbf{B}},\bar{\mathbf{C}}) \\ &= \mathrm{Tr}\left[\left(\boldsymbol{\Delta}\circledast(\mathbf{Z}-\bar{\mathbf{X}})\right)\left(\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X}_\theta)\right)\right] \\ &\quad + \frac{\mu}{2}\left(\|\mathbf{B}_\theta\|_F^2-\|\bar{\mathbf{B}}\|_F^2+\|\mathbf{C}_\theta\|_F^2-\|\bar{\mathbf{C}}\|_F^2\right) \\ &\quad + \frac{1}{2}\|\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X}_\theta)\|_F^2. \end{aligned}$$

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

$$\begin{aligned} dU_\theta &:= \theta\mathrm{Tr}\left[\left(\boldsymbol{\Delta}\circledast(\mathbf{Z}-\bar{\mathbf{X}})\right)\left(\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\right)\right] \\ &\quad + \frac{\mu\theta}{2}\left(\|\mathbf{B}\|_F^2-\|\bar{\mathbf{B}}\|_F^2+\|\mathbf{C}\|_F^2-\|\bar{\mathbf{C}}\|_F^2\right) \\ &\quad + \frac{\theta^2}{2}\|\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\|_F^2. \end{aligned}$$

Using (30), dU_θ can be expressed in terms of dU as

$$dU_\theta := \theta\left(dU - \frac{1}{2}\|\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\|_F^2\right) + \frac{\theta^2}{2}\|\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\|_F^2.$$

Since dU is strictly negative, so is $dU - \frac{1}{2}\|\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\|_F^2$, and hence

$$\lim_{\theta\to 0}\frac{1}{\theta}dU_\theta = \left(dU - \frac{1}{2}\|\boldsymbol{\Delta}\circledast(\bar{\mathbf{X}}-\mathbf{X})\|_F^2\right) < 0.$$

But then in every neighborhood of $(\bar{\mathbf{B}},\bar{\mathbf{C}})$ there is a point $(\mathbf{B}_\theta,\mathbf{C}_\theta)$ such that $U(\mathbf{B}_\theta,\mathbf{C}_\theta) < U(\bar{\mathbf{B}},\bar{\mathbf{C}})$, meaning $(\bar{\mathbf{B}},\bar{\mathbf{C}})$ 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 completes the proof. ■

B. Proof of Proposition 2

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

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

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

Without loss of generality, $\underline{\mathbf{X}}$ can be expressed 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

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

Focus first 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 $\{\mathbf{w}_r\}$, as well as for fixed products $\gamma_r := a_r b_r c_r$. The constraints and hence the LS part of the cost depend on γ_r only, and not on their particular factorizations $a_r b_r c_r$. Thus, the penalty is the only term that varies when γ_r is constant, rendering the inner-most minimization in (35) equivalent to

$$\min_{a_r,b_r,c_r} a_r^2 + b_r^2 + c_r^2 \quad \text{s. to } \gamma_r = a_r b_r c_r,\ r = 1,\ldots,R. \tag{36}$$

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

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

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

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

C. Proof of Corollary 1

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

Lemma 3: [17, p.58] If $\mathbf{E}\in\mathbb{R}^{m\times m}$ *satisfies* $\|\mathbf{E}\|_2 \le 1$, *then* $\mathbf{I}+\mathbf{E}$ *is invertible, and* $\left\|(\mathbf{I}+\mathbf{E})^{-1}\right\|_2 \le (1-\|\mathbf{E}\|_2)^{-1}$.

For any value of μ, and with $(\mathbf{A},\mathbf{B},\mathbf{C})$ being the minimizers of (8), the useful inequality

$$\mu\left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2\right) \le \|\underline{\boldsymbol{\Delta}}\circledast\underline{\mathbf{Z}}\|_F^2 \tag{37}$$

follows by comparing the cost at the minimum and at the origin

A second characterization of the minimum of (8) can be obtained from the first-order optimality condition. Upon vectorizing $\mathbf{A}$, the cost in (8) can be rewritten as

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

where $\mathbf{z}_p$, $\boldsymbol{\delta}_p$, and $\mathbf{a}$ denote the vectorizations of matrices $\mathbf{Z}_p$, $\mathbf{D}_p$, and $\mathbf{A}$, respectively. Additional regularization terms that vanish when taking derivatives w.r.t. $\mathbf{A}$ were removed from (38). Nulling the gradient of (38) w.r.t. $\mathbf{a}$ yields

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

with

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

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

$$\begin{aligned}\|\mathbf{E}\|_2 &\le \frac{1}{\mu}\|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2 \\ \|\boldsymbol{\zeta}\|_2 &\le \frac{1}{\mu}\|\underline{\boldsymbol{\Delta}}\circledast\underline{\mathbf{Z}}\|_F\|\mathbf{B}\|_F\|\mathbf{C}\|_F.\end{aligned}$$

According to Lemma 3, if μ is chosen large enough so that $\|\mathbf{E}\|_2 \le 1$, then the norm of $\mathbf{A}$ is bounded by

$$\|\mathbf{A}\|_F = \|\mathbf{a}\|_2 \le (\mu - \|\mathbf{B}\|_F^2\|\mathbf{C}\|_F^2)^{-1}\|\mathbf{B}\|_F\|\mathbf{C}\|_F\|\underline{\boldsymbol{\Delta}}\circledast\underline{\mathbf{Z}}\|_F \tag{39}$$

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

Yet a third characterization was obtained in Appendix B, where the norm of the factor columns were shown equal to each other, so that

$$\|\mathbf{A}\|_F = \|\mathbf{B}\|_F = \|\mathbf{C}\|_F. \tag{40}$$

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

$$\|\mathbf{A}\|_F^2 \leq \|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F^2/3\mu \tag{41}$$
$$\|\mathbf{A}\|_F \leq (\mu - \|\mathbf{A}\|_F^4)^{-1}\|\mathbf{A}\|_F^2\|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F. \tag{42}$$

Form (42), two complementary cases arise:

$$\mathbf{c1)}\ \|\mathbf{A}\|_F = 0;\ \text{and}$$
$$\mathbf{c2)}\ 1 \leq (1 - \|\mathbf{A}\|_F^4/\mu)^{-1}\|\mathbf{A}\|_F\|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F/\mu. \tag{43}$$

To argue that c2) is impossible, substitute (41) into (43) and square the result to obtain

$$1 \leq (1 - \|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F^4/9\mu^3)^{-2}\|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F^4/3\mu^3. \tag{44}$$

But by hypothesis $\mu \geq \|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F^{4/3}$ so that $\|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F^4/\mu^3 \leq 1$, and the right-hand side of (44) is bounded by 0.43, so that (44) does not hold. This implies that c1); i.e., $\|\mathbf{A}\|_F = \|\mathbf{B}\|_F = \|\mathbf{C}\|_F = 0$, must hold, which completes the proof.

Still, the bound at 0.43 can be pushed to one by further reducing μ, and the proof remains valid under the slightly relaxed condition $\mu > (18/(5+\sqrt{21}))^{-1/3}\|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F^{4/3} \simeq 0.81\|\underline{\boldsymbol{\Delta}} \circledast \underline{\mathbf{Z}}\|_F^{4/3}$. ■

D. RKHS Imputation

Recursive application of the Representer Theorem yields finitely-parameterized minimizers $\hat{a}_r$, $\hat{b}_r$, and $\hat{c}_r$ 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 $\boldsymbol{k}_{\mathcal{M}}^T(m) := [k_{\mathcal{M}}(1,m), \ldots, k_{\mathcal{M}}(M,m)]$, and correspondingly $\boldsymbol{k}_{\mathcal{N}}^T(n) := [k_{\mathcal{N}}(1,n), \ldots, k_{\mathcal{N}}(N,n)]$, and $\boldsymbol{k}_{\mathcal{P}}^T(p) := [k_{\mathcal{P}}(1,p), \ldots, k_{\mathcal{P}}(P,p)]$, 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

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

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

$$\min_{\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}} \sum_{p=1}^{P} \left\| \left(\mathbf{Z}_p - \mathbf{K}_{\mathcal{M}}\hat{\mathbf{A}}\text{diag}\left[\mathbf{e}_p^T\mathbf{K}_{\mathcal{P}}\hat{\mathbf{C}}\right]\hat{\mathbf{B}}^T\mathbf{K}_{\mathcal{N}}\right) \circledast \boldsymbol{\Delta}_p \right\|_F^2 + \frac{\mu}{2}\left(\text{Tr}(\hat{\mathbf{A}}^T\mathbf{K}_{\mathcal{M}}\hat{\mathbf{A}}) + \text{Tr}(\hat{\mathbf{B}}^T\mathbf{K}_{\mathcal{N}}\hat{\mathbf{B}}) + \text{Tr}(\hat{\mathbf{C}}^T\mathbf{K}_{\mathcal{P}}\hat{\mathbf{C}})\right)$$

which is equivalent to (16) after 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).

E. Covariance Estimation

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

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

After summing over $p' = p$, one obtains

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

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

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

as stated in (18).

F. Proof of Lemma 1

Towards establishing properties i)–iii) in Lemma 1, consider expanding the difference between $g(\mathbf{A}, \bar{\mathbf{A}})$ and $f(\mathbf{A})$. One readily obtains

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

which is nonnegative from the definition of λ and, together with its gradient, vanish at $\bar{\mathbf{A}}$. ■

G. Proof of Lemma 2

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

$$f_1(\mathbf{A}) := \mathrm{Tr}\left(\mathbf{A}^T \mathbf{R}_A^{-1} \mathbf{A}\right) \tag{47}$$

$$g_1(\mathbf{A}, \bar{\mathbf{A}}) := \lambda \mathrm{Tr}\left(\mathbf{A}^T \mathbf{A}\right) - 2\mathrm{Tr}(\mathbf{\Theta}^T \mathbf{A}) + \mathrm{Tr}(\mathbf{\Theta}^T \bar{\mathbf{A}}) \tag{48}$$

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

$$f_2(\mathbf{A}) := -\mathbf{1}_M \mathbf{\Delta} \circledast \mathbf{Z} \log(\mathbf{A}\mathbf{\Pi}^T)\mathbf{1}_{NP} \tag{49}$$

$$g_2(\mathbf{A}, \bar{\mathbf{A}}) := -\sum_{r=1}^{R}\sum_{m=1}^{M}\sum_{k=1}^{NP} \delta_{mk} z_{mk} \alpha_{mkr} \log\left(\frac{a_{mr}\pi_{kr}}{\alpha_{mkr}}\right) \tag{50}$$

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 functions $g(\mathbf{A}, \bar{\mathbf{A}})$ and $f(\mathbf{A})$ in Lemma 2, as long as they are satisfied both by the pairs (47)-(48) and (49)-(50).

Focusing on the first pair, the arguments in Appendix F imply that properties i)-iii) are satisfied by $g_1(\mathbf{A}, \bar{\mathbf{A}})$ and $f_1(\mathbf{A})$. Considering the second pair, and expanding $f_2(\mathbf{A})$ yields

$$f_2(\mathbf{A}) = -\sum_{m=1}^{M}\sum_{k=1}^{NP} \delta_{mk} z_{mk} \log\left(\sum_{r=1}^{R} a_{mr}\pi_{kr}\right) \tag{51}$$

where the logarithm can be rewritten as (see also [11])

$$\log\left(\sum_{r=1}^{R} a_{mr}\pi_{kr}\right) = \log\left(\sum_{r=1}^{R} \alpha_{mkr} \frac{a_{mr}\pi_{kr}}{\alpha_{mkr}}\right) \tag{52}$$

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

and the inequality holds because of the concavity of the logarithm and the coefficients $\{\alpha_{mkr}\}_{r=1}^{R}$ summing up to one. Since substituting (53) for (52) in (51) results in (50), it follows that $g_2(\mathbf{A}, \bar{\mathbf{A}})$ and $f_2(\mathbf{A})$ satisfy property iii). The proof is complete after evaluating the pair of functions and their derivatives at $\bar{\mathbf{A}}$ to confirm that properties i) and ii) hold too.

The minimum $a^{\star}_{g,mr} := t_{mr} + \sqrt{t_{mr}^2 + s_{mr}}$ is obtained readily after equating to zero the derivative of the corresponding summand in (27), 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. Mach. Learn. Res.*, vol. 10, pp. 803–826, 2009.

[2] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometr. Intell. Lab. Syst.*, 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. Nat. Acad. Sci.*, vol. 97, no. 18, pp. 10101–10106, 2000.

[4] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. Allerton Conf. Commun., Contr., Comput.*, Monticello, Jun. 2010.

[5] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Nonparametric low-rank tensor imputation," in *Proc. IEEE Statistical Signal Process. Workshop*, Ann Arbor, MI, USA, Aug. 2012, pp. 888–891.

[6] J. F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimiz.*, vol. 20, pp. 1956–1982, Jan. 2010.

[7] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[8] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, Dec. 2012.

[9] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.

[10] J. Chen and Y. Saad, "On the tensor SVD and the optimal low-rank othogonal approximation of tensors," *SIAM J. Matrix Anal. Appl. (SIMAX)*, vol. 30, no. 4, pp. 1709–1734, 2009.

[11] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 4, pp. 1272–1299, 2012.

[12] Y. Chi, Y. C. Eldar, and R. Calderbank, "PETRELS: Subspace estimation and tracking from partial observations," in *Proc. IEEE Int. Conf. on Acoust., Speech Signal Process.*, Kyoto, Japan, Mar. 2012.

[13] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis*. New York, NY, USA: Wiley, 2009.

[14] M. Fazel, "Matrix Rank Minimization with Applications," PhD, Stanford Univ., Elect. Eng. Dep., Stanford, CA, USA, 2002.

[15] 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.

[16] J. S. Goldstein, I. S. Reed, and L. L. Scharf, "A multistage representation of the Wiener filter based on orthogonal projections," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2943–2959, Jul. 1998.

[17] G. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: Johns Hopkins Univ.Press, 1996.

[18] J. Håstad, "Tensor rank is NP-complete," *J. Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

[19] Internet Brain Segmentation Repository, "MR Brain Data Set 657" Center for Morphometric Analysis at Massachusetts General Hospital [Online]. Available: http://www.cma.mgh.harvard.edu/ibsr/

[20] S. M. Kay, *Fundamental of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[21] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[22] J. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Alg. Appl.*, vol. 18, no. 2, pp. 95–138, 1977.

[23] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[24] M. Mardani, G. Mateos, and G. B. Giannakis, "Decentralized sparsity-regularized rank minimization: Algorithms and application," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 5374–5388, Nov. 1, 2013.

[25] M. Mardani, G. Mateos, and G. B. Giannakis, "Rank minimization for subspace tracking from incomplete data," in *Proc. IEEE Int. Conf. on Acoust., Speech Signal Process.*, Vancouver, Canada, May 2013, pp. 5681–5685.

[26] U. Nagalakshmi *et al.*, "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, Jun. 2008.

[27] M. Z. Nashed and Q. Sun, "Function spaces for sampling expansions," in *Multiscale Signal Analysis and Modelling*, X. Shen and A. Zayed, Eds. New York, NY, USA: Springer-Verlag, 2012, Lecture Notes in EE, pp. 81–104.

[28] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: The MIT Press, 2006.

[29] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Opt.*, 2012.

[30] B. Recht and C. Re, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Math. Programm. Comput.*, vol. 5, no. 2, pp. 201–226, 2013.

[31] 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," *Amer. J. Human Genet.*, vol. 78, pp. 629–644, 2006.

[32] N. D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *J. Chemometr.*, vol. 14, no. 3, pp. 229–239, 2000.

[33] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2377–2388, Aug. 2000.
[34] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 810–823, Mar. 2000.
[35] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," *Adv. Neural Inf. Process. Syst.*, vol. 17, pp. 1329–1336, 2005.
[36] A. Stegeman and N. D. Sidiropoulos, "On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition," *Linear Algebra Appl.*, vol. 420, no. 2, pp. 540–552, 2007.
[37] J. M. F. ten Berge and N. D. Sidiropoulos, "On uniqueness in CANDECOMP/PARAFAC," *Psychometrika*, vol. 67, no. 3, pp. 399–409, 2002.
[38] R. Tomioka, K. Hayashi, and H. Kashima, Estimation of Low-Rank Tensors via Convex Optimization 2011 [Online]. Available: arXiv:1010.0789v2 [stat.ML], submitted for publication
[39] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 49–95, 1996.
[40] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA, USA: SIAM, 1990.
[41] M. Welling and M. Weber, "Positive tensor factorization," *Pattern Recogn. Lett.*, vol. 22, pp. 1255–1261, 2001.

Juan Andrés Bazerque (M'13) received the M.Sc. degree in electrical engineering in 2009 and the Ph.D. degree in 2013, both from the University of Minnesota (UofM), Minneapolis, and the B.Sc. degree in electrical engineering from the Universidad de la República (UdelaR), Montevideo, Uruguay, in 2003. Since September 2013, he has been a Postdoctoral Research Associate in with the Spincom group at (UofM). From 2000 to 2006, he was a Teaching Assistant with the Department of Mathematics and Statistics, and with the Department of Electrical Engineering (UdelaR). From 2003 to 2006, he worked as a telecommunications engineer at the Uruguayan company Uniotel S.A. developing applications for Voice over IP. His broad research interests lie in the general areas of network theory, signal processing, and computational biology. His current research focuses on network identifiability, gene expression networks, chemical-genetic interactions, and sparsity-aware statistical modeling. Dr. Bazerque received the UofM's Distinguished M.Sc. Thesis Award in 2009 and the Best Student Paper award at the 2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communication (CROWNCOM) 2007.

Gonzalo Mateos (M'12) received the B.Sc. degree in electrical engineering from the Universidad de la República (UdelaR), Montevideo, Uruguay, in 2005 and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Minnesota, Minneapolis, in 2009 and 2011, respectively.

Since 2012, he has been a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering and the Digital Technology Center, University of Minnesota. From 2003 to 2006, he was an assistant with the Department of Electrical Engineering, UdelaR. From 2004 to 2006, he worked as a Systems Engineer at Asea Brown Boveri (ABB), Uruguay. His research interests lie in the areas of Big Data analytics, signal processing, and networking. His current research focuses on distributed signal processing, sparse linear regression, and statistical learning for social data analysis and network health monitoring.

Georgios B. Giannakis (F'97) received the Diploma in electrical engineering from the National Technical University of Athens, Greece, in 1981. From 1982 to 1986, he was with the University of Southern California (USC), where he received the M.Sc. degree in electrical engineering, in 1983, the M.Sc. degree in mathematics, in 1986, and the Ph.D. degree in electrical engineering in 1986.

Since 1999 he has been a professor with the University of Minnesota, where he now holds an ADC Chair in Wireless Telecommunications in the ECE Department, and serves as Director of the Digital Technology Center. His general interests span the areas of communications, networking, and statistical signal processing – subjects on which he has published more than 350 journal papers, 580 conference papers, 20 book chapters, two edited books and two research monographs (h-index 105). Current research focuses on sparsity and big data analytics, wireless cognitive radios, mobile *ad hoc* networks, renewable energy, power grid, gene-regulatory, and social networks. He is the (co-) inventor of 21 patents issued.

Dr. Giannakis is the (co-) recipient of eight Best Paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, and the G. W. Taylor Award for Distinguished Research from the University of Minnesota. He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.