INFERENCE OF POISSON COUNT PROCESSES USING LOW-RANK TENSOR DATA

Juan Andrés Bazerque, Gonzalo Mateos, and Georgios B. Giannakis

Dept. of ECE and DTC, University of Minnesota, Minneapolis, MN 55455

ABSTRACT

A novel regularizer capturing the tensor rank is introduced in this paper as the key enabler for completion of three-way data arrays with missing entries. The novel regularized imputation approach induces sparsity in the factors of the tensor's PARAFAC decomposition, thus reducing its rank. The focus is on count processes which emerge in diverse applications ranging from genomics to computer and social networking. Based on Poisson count data, a maximum a-posteriori (MAP) estimator is developed using the Kullback-Leibler divergence criterion. This probabilistic approach also facilitates incorporation of correlated priors regularizing the rank, while endowing the tensor imputation method with extra smoothing and prediction capabilities. Tests on simulated and real datasets corroborate the sparsifying regularization effect, and demonstrate recovery of 15% missing RNA-sequencing data with an inference error of $-12dB$.

Index Terms— Tensor, low-rank, missing data, Poisson processes.

1. INTRODUCTION

Recovering missing data is a fundamental task arising in various applications as diverse as medical imaging, bioinformatics, as well as social and computer networking. The key idea rendering recovery feasible is leveraging the regularity present among missing and available data. Low rank is an attribute capturing this regularity that can be readily exploited when data are organized into a matrix. Indeed, data matrices can be completed by fitting them to a partial set of their entries while regularizing their ranks [5]. In this context, the nuclear norm has been advocated as a practical rank regularizer with analytical performance guarantees [9].

Relation to prior works. The focus here is on 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 [11]. Leveraging the low-rank structure for tensor completion is challenging, since even computing the tensor rank is NP-hard. 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, or, employing the parallel factor (PARAFAC) decomposition [16, 11]. In the context of tensor completion, an approach falling under the first category can be found in [10], while the PARAFAC decomposition was dealt with in [2].

The imputation approach presented here builds on a novel regularizer accounting for the tensor rank. It relies on redefining the matrix nuclear norm in terms of low-rank factors, which admit a neat connection with the atomic-norm [6]. While least-squares (LS)

is typically adopted as the fitting criterion for matrix and tensor completion, implicitly assuming Gaussian data, the method here targets count data available in the form of network traffic data, genome sequencing, and social media interactions, modeled as a Poisson process that leads to a MAP estimator in terms of the Kullback-Leibler divergence [8]. This probabilistic approach also facilitates the incorporation of correlated priors that regularize the rank, while endowing the tensor imputation method with extra smoothing and prediction capabilities.

2. PRELIMINARIES

2.1. Nuclear-norm minimization for matrix completion

Leveraging dependencies implied by the property of low-rank is a popular method for estimating missing values of a data matrix $\mathbf{Z} \in \mathbb{R}^{N \times M}$ [9]. For the imputation to be feasible, a binding assumption that relates available entries with the missing ones is required. An alternative is to postulate that $\mathbf{Z}$ has low rank $R \ll \min(N, M)$, which implies that the vector $\mathbf{s}(\mathbf{Z})$ of its singular values is *sparse*.

Finding matrix $\hat{\mathbf{Z}}$ with rank not exceeding R, which approximates $\mathbf{Z}$ in the available entries specified by a binary matrix $\mathbf{\Delta} \in \{0, 1\}^{N \times M}$, can be formulated as the following convex optimization problem [9] ($*$ denotes Hadamard product, subscript F the Frobenious norm, and superscript T transposition)

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

where $\|\mathbf{X}\|_* := \|\mathbf{s}(\mathbf{X})\|_1$ denotes the nuclear norm, and $\mu \geq 0$ is a rank-controlling parameter tuned to ensure $\text{rank}(\hat{\mathbf{Z}}) \leq R$.

Several iterative algorithms have been proposed to solve (1), and are effective in tackling low- to medium-size matrix completion problems; see e.g., [15]. However, most algorithms require computation of singular values per iteration and become prohibitively expensive when dealing with high-dimensional data. To solve (1) efficiently, consider the alternative characterization of the nuclear norm $\|\mathbf{X}\|_* = \min_{\{\mathbf{B}, \mathbf{C}\}} \frac{1}{2}(\|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2)$, where the minimization is over all possible bilinear factorizations of $\mathbf{X} = \mathbf{B}\mathbf{C}^T$. Accordingly, one can arrive at the following equivalent reformulation of (1) [15]

$$\min_{\{\mathbf{B}, \mathbf{C}\}} \|(\mathbf{Z} - \mathbf{B}\mathbf{C}^T) * \mathbf{\Delta}\|_F^2 + \frac{\mu}{2}(\|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2). \quad (2)$$

The equivalence implies that by finding the global minimum of (2) [which could entail considerably less variables than (1)], one can recover the optimal solution of (1) as $\hat{\mathbf{Z}} := \hat{\mathbf{B}}\hat{\mathbf{C}}^T$ However, since (2) is *nonconvex*, it may have stationary points which need not be globally optimum. Interestingly, the next proposition offers a condition under which a stationary point of (2) is globally optimum for (1) [12].

Proposition 1: *Let* $\{\hat{\mathbf{B}}, \hat{\mathbf{C}}\}$ *be a stationary point of* (2) . *If* $\|\mathbf{Z} - \hat{\mathbf{B}}\hat{\mathbf{C}}^T\| < \mu/2$, *then* $\hat{\mathbf{Z}} := \hat{\mathbf{B}}\hat{\mathbf{C}}^T$ *is globally optimal for* (1).

The Frobenious-norm regularization for controlling the rank in (2), will be useful to obtain tensor counterparts in Section 3.
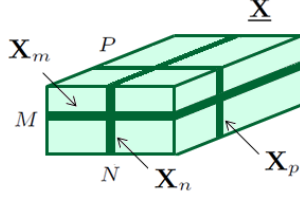
Fig. 1. Tensor slices along the row, column, and tube dimensions.

2.2. 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 the means to define its rank [16, 11]. For given $R \in \mathbb{N}$, consider matrices $\mathbf{A} \in \mathbb{R}^{N \times R}$, $\mathbf{B} \in \mathbb{R}^{M \times R}$, and $\mathbf{C} \in \mathbb{R}^{P \times R}$, such that

$$\underline{\mathbf{X}}(m, n, p) = \sum_{r=1}^{R} \mathbf{A}(m, r)\mathbf{B}(n, r)\mathbf{C}(p, r). \quad (3)$$

The rank of $\underline{\mathbf{X}}$ is defined as 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 (3) holds with $R = R^*$.

To appreciate why the aforementioned rank definition is natural, rewrite (3) 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

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \gamma_r(\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r) \quad (4)$$

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 matrix factors will be used in the sequel

$$\mathbf{X}_p = \mathbf{A}\text{diag}\left[\mathbf{e}_p^T \mathbf{C}\right] \mathbf{B}, \quad p = 1, \ldots, P \quad (5)$$

where the diagonal matrix $\mathbf{D} = \text{diag}[\mathbf{u}]$ 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. (3)], and one can also write $\mathbf{X}_m = \mathbf{B}\text{diag}\left[\mathbf{e}_m^T \mathbf{A}\right] \mathbf{C}$, or, $\mathbf{X}_n = \mathbf{C}\text{diag}\left[\mathbf{e}_n^T \mathbf{B}\right] \mathbf{A}$ in terms of slices along the first (row), or, second (column) dimensions.

3. RANK REGULARIZATION FOR TENSORS

Generalizing the nuclear-norm regularization technique (1) 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 [11]. The Frobenious-norm regularization outlined in Section 2.1 offers a viable option for low-rank tensor completion under the PARAFAC model however, by solving

$$\hat{\underline{\mathbf{Z}}} := \arg\min_{\{\underline{\mathbf{X}}, \mathbf{A}, \mathbf{B}, \mathbf{C}\}} \| (\underline{\mathbf{Z}} - \underline{\mathbf{X}}) * \underline{\mathbf{\Delta}}\|_F^2 + \frac{\mu}{2}\left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2\right)$$

$$\text{s. to} \quad \mathbf{X}_p = \mathbf{A}\text{diag}\left[\mathbf{e}_p^T \mathbf{C}\right] \mathbf{B}, \quad p = 1, \ldots, P \quad (6)$$

where $\|\underline{\mathbf{X}}\|_F^2 := \sum_m \sum_n \sum_p \underline{\mathbf{X}}^2(m, n, p)$ defines the tensor's Frobenious norm, and $(\underline{\mathbf{X}} * \underline{\mathbf{\Delta}})(m, n, p) := \underline{\mathbf{X}}(m, n, p)\underline{\mathbf{\Delta}}(m, n, p)$ the Hadamard product.

Different from the matrix case, it is unclear whether the proposed regularization in (6) bears any relation with the tensor rank. Interestingly, the analysis provided next corroborates the capability of (6) to produce a low-rank tensor $\hat{\underline{\mathbf{Z}}}$, for sufficiently large μ. In this direction, consider the alternative completion problem stated in terms of the normalized tensor representation (4)

$$\hat{\underline{\mathbf{Z}}}' := \arg\min_{\{\underline{\mathbf{X}}, \boldsymbol{\gamma}, \{\mathbf{u}_r\}, \{\mathbf{v}_r\}, \{\mathbf{w}_r\}\}} \| (\underline{\mathbf{Z}} - \underline{\mathbf{X}}) * \underline{\mathbf{\Delta}}\|_F^2 + \frac{\mu}{2}\|\boldsymbol{\gamma}\|_{2/3}^{2/3}$$

$$\text{s. to} \quad \underline{\mathbf{X}} = \sum_{r=1}^{R} \gamma_r(\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r) \quad (7)$$

where $\boldsymbol{\gamma} := [\gamma_1, \ldots, \gamma_R]^T$, the nonconvex $\ell_{2/3}$ (pseudo)-norm is given by $\|\boldsymbol{\gamma}\|_{2/3} := \left(\sum_{r=1}^{R} |\gamma_r|^{2/3}\right)^{3/2}$, and the unit-norm constraint on the factors' columns is left implicit. Problems (6) and (7) are equivalent as established by the following proposition (a proof is omitted due to lack of space; see [4])

Proposition 2: *The solutions of (6) and (7) 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 (6) to produce a low-rank approximant tensor $\underline{\mathbf{X}}$, consider transforming (7) once more by rewriting it in the constrained-error form

$$\hat{\underline{\mathbf{Z}}}'' := \arg\min_{\{\underline{\mathbf{X}}, \boldsymbol{\gamma}, \{\mathbf{u}_r\}, \{\mathbf{v}_r\}, \{\mathbf{w}_r\}\}} \|\boldsymbol{\gamma}\|_{2/3} \quad (8)$$

$$\text{s. to} \quad \| (\underline{\mathbf{Z}} - \underline{\mathbf{X}}) * \underline{\mathbf{\Delta}}\|_F^2 \leq \sigma^2, \quad \underline{\mathbf{X}} = \sum_{r=1}^{R} \gamma_r(\mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r).$$

For any value of σ^2 there exists a corresponding Lagrange multiplier λ such that (7) and (8) 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 (8).] The $\ell_{2/3}$-norm $\|\boldsymbol{\gamma}\|_{2/3}$ in (8) produces a sparse vector $\boldsymbol{\gamma}$ when minimized [7], sharing this well-documented property of the ℓ_1-norm as their norm-one balls, depicted in Fig. 2, share the "pointy geometry" which is responsible for inducing sparsity.

With (6) equivalently rewritten as in (8), its low-rank inducing property is now revealed. As $\boldsymbol{\gamma}$ in (8) becomes sparse, some of its entries γ_r are zeroed, and the corresponding outer-products $\gamma_r(\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$ drop from the sum in (4), thus lowering the rank of $\underline{\mathbf{X}}$.

Through (8), it is also possible to relate (6) with the atomic norm in [6]. It is established in [6] that if $R = R^*$ is the rank of $\underline{\mathbf{X}}$, then the ℓ_1-norm of $\boldsymbol{\gamma}$ is a proper norm for $\underline{\mathbf{X}}$, named atomic norm and denoted by $\|\underline{\mathbf{X}}\|_\star := \|\boldsymbol{\gamma}\|_1$. Thus, by replacing $\|\boldsymbol{\gamma}\|_{2/3}$ with $\|\underline{\mathbf{X}}\|_\star$, (8) becomes a convex problem in the variable $\underline{\mathbf{X}}$. Still, the complexity of solving such a variant of (8) resides in that finding $\|\underline{\mathbf{X}}\|_\star$ through the PARAFAC decomposition is NP-hard [11]. In this regard, it is remarkable that arriving to (8) had the sole purpose of demonstrating the low-rank inducing property, and that (6) is to be solved by the algorithm developed in the ensuing section. Such an algorithm will neither require computing the PARAFAC decomposition of the variable $\underline{\mathbf{X}}$, nor knowing its rank. Indeed, the number of columns in $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ can be set to an overestimate of the rank of $\underline{\mathbf{Z}}$, as for instance, the general overbound $\bar{R} := \min\{MN, NP, PM\} \geq \text{rank}(\underline{\mathbf{Z}})$, and the low-rank of $\underline{\mathbf{X}}$ will be induced by regularization as argued earlier. To carry out a
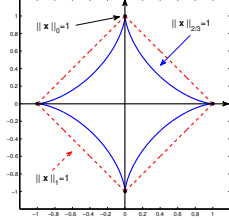
Fig. 2. $\ell_{2/3}$-norm ball compared to the ℓ_0- and ℓ_1-norm balls

fair comparison, only convergence to a stationary point of (6) will be guaranteed in this paper.

Remark 1 From a probabilistic vantage point, adoption of an LS criterion in (6) implicitly assumes that the random variation in $\underline{\mathbf{Z}}$ adheres to a Gaussian distribution. The next section deals with Poisson-distributed tensor data, a natural alternative to the Gaussian model when integer data is obtained by counting independent events [8]. Interpreting the Frobenius-norm regularization in (6) as $\mathcal{N}(\mathbf{0}, (1/\mu)\mathbf{I})$ priors on the independent columns of the factor matrices, the model in Section 4 also incorporates general per-factor covariance matrices; see e.g., [3].

4. INFERENCE FOR LOW-RANK POISSON TENSORS

Suppose that the entries z_{mnp} of $\underline{\mathbf{Z}}$ are Poisson distributed, i.e.,

$$P(z_{mnp} = k) = \frac{x_{mnp}^k e^{-x_{mnp}}}{k!} \qquad (9)$$

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

$$l_{\underline{\Delta}}(\underline{\mathbf{Z}}; \underline{\mathbf{X}}) = \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{p=1}^{P} \delta_{mnp}[z_{mnp} \log(x_{mnp}) - x_{mnp}] \quad (10)$$

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

The choice of a Poisson distribution (9) over a Gaussian one for counting data, prompts for minimization of the Kullback-Leibler divergence (10) instead of LS as a more suitable criterion [8]. Still, the entries of $\underline{\mathbf{X}}$ are unconnected in model (10), and a binding PARAFAC modeling assumption is natural for feasibility of the tensor approximation task under missing data. On top of that, augmenting (10) with a Frobenius-norm regularizing term on the model factors offers an attractive way of controlling the rank of the solution $\hat{\underline{\mathbf{Z}}}$ [4], and avoid well-known indeterminacies of the PARAFAC model [1]. Accordingly, the maximum-a-posteriori (MAP) estimator of $\underline{\mathbf{X}}$ given Poisson-distributed data (entries of $\underline{\mathbf{Z}}$ indexed by $\underline{\Delta}$), and with Gaussian priors for the PARAFAC factors of $\underline{\mathbf{X}}$, becomes

$$\hat{\underline{\mathbf{Z}}} := \arg\min_{\{\underline{\mathbf{X}}, \mathbf{A}, \mathbf{B}, \mathbf{C}\}} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{p=1}^{P} \delta_{mnp}(x_{mnp} - z_{mnp} \log(x_{mnp}))$$
$$+ \frac{\mu}{2} \left[\mathrm{Tr}\left(\mathbf{A}^T \mathbf{K}_A^{-1} \mathbf{A}\right) + \mathrm{Tr}\left(\mathbf{B}^T \mathbf{K}_B^{-1} \mathbf{B}\right) + \mathrm{Tr}\left(\mathbf{C}^T \mathbf{K}_C^{-1} \mathbf{C}\right) \right]$$
$$\text{s. to} \quad \mathbf{X}_p = \mathbf{A}\mathrm{diag}\left[\mathbf{e}_p^T \mathbf{C}\right] \mathbf{B}^T, \quad p = 1, \ldots, P. \quad (11)$$

It is apparent from (11) that the mutually-independent columns of $\mathbf{A}$ are modeled as $\mathbf{a}_r \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_A)$, and likewise for $\mathbf{B}$ and $\mathbf{C}$.

Algorithm 1 : Low-rank Poisson-tensor imputation (LRPTI)

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

Relative to (6), through these covariances one can incorporate side information in the form of correlations among tensor slices (along the row, column, and tube dimensions), and enhance the estimation performance when large amounts of data are missing. Interestingly, this also endows (11) with *prediction* (or extrapolation) capabilities, useful when full slices of data in $\underline{\mathbf{Z}}$ are not observed [4]. With the aid of Representer's Theorem, it is also possible to interpret (11) as a variational estimator in reproducing-kernel Hilbert spaces [3].

4.1. Block successive upper-bound minimization algorithm

A block coordinate-descent algorithm is developed in this section, that provably converges to a stationary point of (11). This iterative alternating-minimization procedure sequentially optimizes (11) with respect to one factor matrix, while holding the others fixed.

In the sequel, the goal is to arrive at a suitable expression for the cost function in (11), 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 unfolding of $\underline{\mathbf{Z}}$ along its third (tube) dimension, and likewise for $\Delta := [\Delta_1, \ldots, \Delta_P] \in \{0, 1\}^{M \times NP}$ and $\mathbf{X} := [\mathbf{X}_1, \ldots, \mathbf{X}_P] \in \mathbb{R}_+^{M \times NP}$. Based on these definitions, (10) can be written as

$$l_{\Delta}(\mathbf{Z}; \mathbf{X}) = \mathbf{1}_M^T(\Delta * [\mathbf{X} - \mathbf{Z} * \log(\mathbf{X})])\mathbf{1}_{NP} \qquad (12)$$

where $\mathbf{1}_M$, $\mathbf{1}_{NP}$ are all-one vectors of dimensions M and NP respectively, and $\log(\cdot)$ should be understood as an element-wise operator. The log-likelihood function (12) can be expressed in terms of the factors $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ by resorting to 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 $\mathbf{c}_r \otimes \mathbf{b}_r$. The following identity [8] is useful

$$\mathbf{X} := [\mathbf{X}_1, \ldots, \mathbf{X}_P] = \mathbf{A}\mathbf{\Pi}^T. \qquad (13)$$

Substituting (13) into (12) one arrives at the desired expression for the cost in (11) as a function of $\mathbf{A}$, namely

$$f(\mathbf{A}) := \mathbf{1}_M^T(\Delta * [\mathbf{A}\mathbf{\Pi} - \mathbf{Z} * \log(\mathbf{A}\mathbf{\Pi}^T)])\mathbf{1}_{NP} + \frac{\mu}{2}\mathrm{Tr}(\mathbf{A}^T\mathbf{K}_A^{-1}\mathbf{A}). \qquad (14)$$

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 on the so-termed block successive upper-bound minimization

(BSUM) algorithm [14]. In BSUM one minimizes a judiciously chosen upper-bound $g(\mathbf{A}, \bar{\mathbf{A}})$ of $f(\mathbf{A})$, which: i) depends on the current iterate $\bar{\mathbf{A}}$; ii) should be simpler to optimize; and iii) satisfies certain local-tightness technical conditions; see also Lemma 1 and [14].

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

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

(15)

where $\lambda := \lambda_{\max}(\mathbf{K}_A^{-1})$ is the largest eigenvalue of $\mathbf{K}_A$, and the parameters s_{rm}, t_{rm}, and u_{rm} are defined in terms of $\bar{\mathbf{A}}$, $\mathbf{Z}$, $\mathbf{\Delta}$, $\mathbf{\Pi}$, and $\mathbf{\Theta} := \left(\frac{1}{\lambda}\mathbf{I} - \mathbf{K}_A^{-1} \right)\bar{\mathbf{A}}$ 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 $u_{mr} := \frac{1}{\lambda\mu} \left(\theta_{mr}\bar{a}_{mr} + \sum_{k=1}^{NP} \delta_{mk} z_{mk} \bar{a}_{mr} \pi_{kr} v_{mrk} \right)$, with $v_{mrk} := (\log(\bar{a}_{mr}\pi_{kr}) - \log(\sum_{r'=1}^{R} \bar{a}_{mr'} \pi_{kr'})) / \sum_{r'=1}^{R} \bar{a}_{mr'} \pi_{kr'}$. As asserted in the following lemma, function $g(\mathbf{A}, \bar{\mathbf{A}})$ majorizes $f(\mathbf{A})$ at $\bar{\mathbf{A}}$ and satisfies the technical conditions required for the convergence of BSUM (see [4] for a proof).

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

    i) $f(\bar{\mathbf{A}}) = g(\bar{\mathbf{A}}, \bar{\mathbf{A}})$;

    ii) $\frac{d}{d\mathbf{A}} f(\mathbf{A})|_{\mathbf{A}=\bar{\mathbf{A}}} = \frac{d}{d\mathbf{A}} g(\mathbf{A}, \bar{\mathbf{A}})|_{\mathbf{A}=\bar{\mathbf{A}}}$; and,

    iii) $f(\mathbf{A}) \leq g(\mathbf{A}, \bar{\mathbf{A}})$, $\forall \mathbf{A}$.

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

Lemma 1 highlights the convenience of adopting $g(\mathbf{A}, \bar{\mathbf{A}})$ in the proposed block-coordinate descent algorithm, since it is separable across the entries of its matrix argument [cf. (15)], and hence it admits a closed-form minimizer given by the MR scalars $a_{g,mr}^{\star}$. The updates $\mathbf{A} \leftarrow \mathbf{A}_g^{*}$ are tabulated under Algorithm 1 for solving (11), where analogous updates for $\mathbf{B}$ and $\mathbf{C}$ are carried out cyclically.

By virtue of properties i)-iii) in Lemma 1, convergence of Algorithm 1 follows readily from the general converge theory available for the BSUM algorithm [14].

Proposition 3: *The iterates $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ generated by Algorithm 1 converge to a stationary point of (11).*

5. NUMERICAL TESTS

5.1. Simulated Poisson data

Synthetic tensor-data of dimensions $M \times N \times P = 16 \times 4 \times 4$ were generated according to the low-rank Poisson-tensor model described in Section 4. Specifically, entries of $\underline{\mathbf{Z}}$ consist of realizations of Poisson random variables generated according to (9), with means specified by entries of $\underline{\mathbf{X}}$. Tensor $\underline{\mathbf{X}}$ is constructed from factors $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ as in (5). Matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ have $R = 2$ columns containing realizations of independent uniform random variables scaled to yield $E[x_{mnp}] = 1,000$.

Half of the entries of $\underline{\mathbf{Z}}$ were removed at random and reserved to evaluate performance. The remaining half of the data was used to recover $\underline{\mathbf{Z}}$ considering the removed data as missing entries. Method
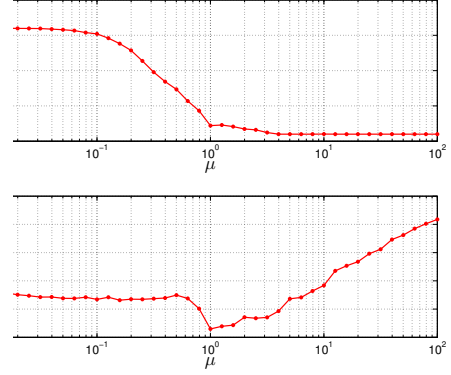

Fig. 3. Performance of the low-rank Poisson imputation method as function of the regularizing parameter μ; (top) rank of the recovered tensor, (bottom) relative recovery error.
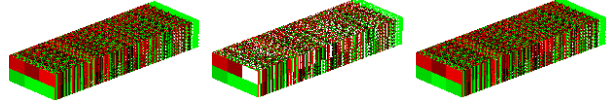

Fig. 4. Imputation of sequencing count data via LRPTI; (left) original data; (center) data with missing entries; (right) recovered tensor.

(11) was employed for recovery, as implemented in the LRPTI Algorithm, with regularization $\mu/2(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2)$, obtained by setting $\mathbf{K}_A = \mathbf{I}_M$, $\mathbf{K}_B = \mathbf{I}_N$, and $\mathbf{K}_C = \mathbf{I}_P$.

The relative recovery error between the recovered tensor and data $\underline{\mathbf{Z}}$ was computed, along with the rank of the recovered tensor, as a measure of performance. Fig. 3 shows these figures of merit averaged over 100 repetitions of the experiment, across values of the parameter μ varying on the interval 0.01 to 100. Fig 3 (bottom) shows that the LRPTI algorithm is successful in recovering the Poisson tensor up to $-15dB$ for a wide range of values of μ, presenting a minimum at $\mu = 1$. This result is consistent with Fig. 3 (top), which shows that rank $R^* = 2$ is recovered at the minimum error. Fig. 3 (top) also corroborates the low-rank inducing effect (6), with the recovered rank varying from the general upperbound $R = \bar{R} = NP = 16$ to $R = 1$, as μ is increased.

5.2. RNA sequencing data

The RNA-Seq method described in [13] exhaustively counts the number of Ribonucleic acid (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 [13] for a biological and a technological replicate of the original sample totalling $N = 3$ instances per primer selection. The data is thus organized in a tensor of dimensions $6,604 \times 3 \times 2$ as shown in Fig. 4 (left), with integer data that are modeled as Poisson counts. Fifteen percent of the data is removed and reserved for assessing performance. The missing data is represented in white in Fig. 4 (center).

The LRPTI algorithm is run with the data available in Fig. 4 (center) producing the recovered tensor depicted in Fig. 4 (right). The recovery error for this experiment was $-12dB$ ($\sim 6\%$).

6. REFERENCES

[1] E. Acar, D. M. Dunlavy, and T. G. Kolda, "A scalable optimization approach fitting canonical tensor decompositions," *Journal of Chemometrics,* vol. 25, no. 2, pp. 67-86, 2011.

[2] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mrup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems,* vol. 106, no. 1, pp. 41-56, 2011.

[3] J. A. Bazerque and G. B. Giannakis, "Nonparametric basis pursuit via sparse kernel-based learning," *IEEE Signal Processing Magazine*, July 2013 (too appear; see also arXiv:1302.5449v1 [cs.LG]).

[4] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Rank regularization and Bayesian inference for tensor completion and extrapolation," *IEEE Trans. Signal Process.*, Dec. 2012 (submitted; see also arXiv:1301.7619v1 [CS.IT]).

[5] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, pp. 925–936, June 2010.

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

[7] R. Chartrand, "Exact reconstruction of sparse signals via non-convex minimization," *IEEE Signal Processing Letters,* vol. 14, no. 10, pp. 707-710, Oct. 2007.

[8] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," *SIAM Journal on Matrix Analysis and Applications,* Dec. 2012 (to appear; see also arXiv:1112.2414v3 [math.NA]).

[9] M. Fazel, "Matrix rank minimization with applications" *PhD Thesis,* Electrical Engineering Dept., Stanford University, vol. 54, pp. 1-130, 2002.

[10] 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.

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

[12] M. Mardani, G. Mateos, and G. B. Giannakis, "In-network sparsity-regularized rank minimization: Algorithms and applications," *IEEE Trans. Signal Process.*, Apr. 2013 (to appear; see also arXiv:1203.1570v1 [cs.MA]).

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

[14] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Opt.* 2012 (submitted; see also arXiv:1209.2385v1 [math.OC]).

[15] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,"*SIAM Review*, vol. 52, no. 3, pp. 471-501, 2010.

[16] J. M. F. ten Berge and N. D. Sidiropoulos, "On uniqueness in CANDECOMP/PARAFAC," *Psychometrika,* vol. 67, no. 3, pp. 399-409, 2002.