Sparse Robust Matrix Tri-factorization with Application to Cancer Genomics

Seung-Jun Kim[†] TaeHyun Hwang[‡] Georgios B. Giannakis[§]

***Abstract*—Nonnegative matrix tri-factorization (NMTF) $\mathbf{X} \approx \mathbf{FSG}^T$ with all matrices nonnegative can reveal simultaneous row and column clusters of $\mathbf{X}$, as well as the associations among the two. In this work, a sparsity-promoting variant is proposed and a simple multiplicative algorithm is developed. The resulting sparse NMTF is further robustified to cope with presence of outliers in the data. A synthetic example illustrates the efficacy of the method. A novel application to cancer patient clustering and pathway analysis is presented using real datasets.**

I. Introduction

Matrix factorization is an important tool for feature extraction and dimensionality reduction tasks with wide range of applications in the areas including engineering, psychometrics, marketing, and computational biology. Using singular value decomposition (SVD), principal component analysis (PCA) factorizes the underlying data matrix under orthogonality constraints to uncover salient uncorrelated variables influencing the data. The k-means clustering may be viewed as matrix factorization under the hard constraint that each data vector (matrix column) is approximated by one of the cluster centroids [16].

When the data are nonnegative, it often makes sense to require also the factors to be nonnegative. Nonnegative matrix factorization (NMF) can help identify easily interpretable parts that comprise the overall data, especially when additive structures can be presumed. For instance, in an article published in *Nature* in 1999, NMF applied to facial images was shown to yield image segments containing different parts of the face [12]. PCA-type approaches might not be good candidates in such cases, as the resulting factors are not guaranteed to be nonnegative. Moreover, strict orthogonality constraints may prevent discovery of possibly overlapping structures that can be naturally present in some datasets.

In its most primitive form, nonnegative matrix (bi-)factorization seeks to obtain factors $\mathbf{F} \in \mathbb{R}_+^{m\times k}$ and $\mathbf{G} \in \mathbb{R}_+^{n\times k}$ that approximate the data matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ such that $\mathbf{X} \approx \mathbf{FG}^T$ in some sense (e.g., minimizing square-errors), as in

$$\min_{\mathbf{F}\geq 0,\mathbf{G}\geq 0} \frac{1}{2}\|\mathbf{X} - \mathbf{FG}^T\|_F^2 \tag{1}$$

where $\mathbf{F} \geq 0$ constrains each element of $\mathbf{F}$ to be nonnegative. The (maximum) rank of the factors k is usually chosen to be much smaller than $\min\{m, n\}$ to effect dimensionality reduction.

It has been also argued that NMF can be viewed as co-clustering of rows and columns of $\mathbf{X}$, especially if $\mathbf{F}$ and $\mathbf{G}$ have orthogonal columns [4]. In this interpretation, the a-th row of $\mathbf{F}$ corresponds to the cluster membership indicator for the a-th row of $\mathbf{X}$; and at the same time, the b-th row of $\mathbf{G}$ corresponds to the cluster membership indicator for the b-th column of $\mathbf{X}$. Application of NMF for document co-clustering was reported in [17].

A number of algorithms have been developed for computing nonnegative factors. The most well-known class comprises the multiplicative update rules proposed in the seminal work by Lee and Seung [13]. Gradient descent-type algorithms have also been studied partly due to the slow convergence of multiplicative updates. Alternating nonnegative least-squares approaches typically enjoy theoretically well-grounded convergence properties [10]. Extensions have been made to incorporate various prior knowledge on the structures of the factors, such as smoothness and sparsity [8].

Nonnegative matrix tri-factorization (NMTF) aims to approximate the data matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ using three factor matrices $\mathbf{F} \in \mathbb{R}_+^{m\times k_1}$, $\mathbf{S} \in \mathbb{R}_+^{k_1\times k_2}$ and $\mathbf{G} \in \mathbb{R}_+^{n\times k_2}$ such that $\mathbf{X} \approx \mathbf{FSG}^T$ [5]. Essentially, the relevant optimization problem is now given by (cf. (1))

$$\min_{\mathbf{F}\geq 0,\mathbf{S}\geq 0,\mathbf{G}\geq 0} \frac{1}{2}\|\mathbf{X} - \mathbf{FSG}^T\|_F^2. \tag{2}$$

To facilitate the clustering interpretation, $\mathbf{F}$ and $\mathbf{G}$ were additionally constrained to be column-orthogonal in [5].

By having three factors instead of two, one can gain a number of benefits. First, under orthogonality constraints, bi-factorization may be too restrictive; an

[†] Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. E-mail: seungjun@umn.edu.

[‡] Masonic Cancer Center, University of Minnesota, Twin Cities. thwang@cs.umn.edu.

[§] Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. georgios@umn.edu.

978-1-4673-1878-5/12/$31.00 ©2012 IEEE

additional factor $\mathbf{S}$ can furnish necessary degrees of freedom to obtain "good" factorization [5]. Also, an important flexibility is that the number k_1 of the column clusters can be different from the number k_2 of the row clusters. This is useful when the rows and the columns correspond to different entities, say, documents and the words contained in them, respectively. Finally, in some applications, direct interpretation of $\mathbf{S}$ is of interest. Specifically, $\mathbf{S}$ can reveal how different row clusters are associated with column clusters, providing a summary of the interaction structure [9].

The goal of this work is to extend NMTF to incorporate sparsity and robustness. Sparsity constraints in NMF were shown to yield more "local" features in an instance of a facial images dataset in [8], which are easier to interpret. Without promoting sparsity, NMF sometimes converged to "global" image segments that do not visually correspond to different parts of the face. Also, along the arguments of variable selection applications, sparsity can help pick the most relevant variables, which is instrumental when such analyses serve as a preliminary step for more costly verification processes, e.g., as in medicine.

To address outliers that may be present in the data due to, e.g., contaminated samples in biological experiments, noisy measurements, and other types of errors, the universal sparsity-controlling outlier rejection (USPACOR) framework is adopted in the context of NMTF [7]. By capitalizing on typical sparsity of outliers, erroneous data entries are effectively compensated to align with the NMTF structure.

The rest of the paper is organized as follows. Sec. II provides the problem formulation for sparse NMTF, and develops a multiplicative update algorithm. Incorporation of robustness is discussed in Sec. III. Tests using simple synthetic data are described in Sec. IV. Results based on real datasets in a bioinformatics application are reported in Sec. V, followed by conclusions in Sec. VI.

II. Sparse NMTF

A. Problem Statement

To promote sparsity of the factors in NMTF, appropriate penalty terms can be added to the objective function of (2). Among widely-used sparsity-inducing penalties is the one based on the ℓ_1-norm, which offers the tightest convex relaxation of the cardinality, or the "ℓ_0-norm." To facilitate derivation of multiplicative update rules, as will be detailed in Sec. II-B, squared ℓ_1-norm penalties are employed.

Based on the preceding discussion, given a nonnegative data matrix $\mathbf{X} \in \mathbb{R}_+^{m \times n}$, the proposed NMTF seeks to find factor matrices $\mathbf{F} \in \mathbb{R}_+^{m \times k_1}$, $\mathbf{S} \in \mathbb{R}_+^{k_1 \times k_2}$, and $\mathbf{G} \in \mathbb{R}_+^{n \times k_2}$ that are sparse, by solving the following optimization problem:

$$\min_{\mathbf{F}\geq 0, \mathbf{S}\geq 0, \mathbf{G}\geq 0} \frac{1}{2}\left(\|\mathbf{X} - \mathbf{F}\mathbf{S}\mathbf{G}^T\|_F^2 + \lambda_F \|\mathbf{F}\|_1^2 \right. \\ \left. + \lambda_S \|\mathbf{S}\|_1^2 + \lambda_G \|\mathbf{G}\|_1^2\right) \qquad (3)$$

where $\|\mathbf{F}\|_1$ denotes the ℓ_1-norm of $\mathbf{F}$, which is equal to the sum of the absolute values of all the entries of $\mathbf{F}$. Since the entries of $\mathbf{F}$ are nonnegative, $\|\mathbf{F}\|_1$ is equal to the sum of all the entries of $\mathbf{F}$.

Prior works advocated enforcing orthogonality of $\mathbf{F}$ and $\mathbf{G}$ to strengthen the clustering interpretation [5], as well as to obtain a more distinctive set of centroids [18]. On the other hand, it has also been argued that orthogonality fails to capture natural semantic structures in certain applications, as it precludes overlapping (soft) clusters [17]. Here, we do not enforce orthogonality to allow for soft clusters. However, it is emphasized that sparsity does not forestall orthogonality; in fact, sparse nonnegative vectors are more likely to be orthogonal.

B. Algorithm

The most popular algorithms for NMF are based on multiplicative update rules [13]. Since (1) is nonconvex, alternating optimization is used to obtain a locally optimal solution. In particular, iteration of the following was shown to yield a non-increasing sequence of objectives:

$$G_{a\mu} \leftarrow \frac{(\mathbf{F}^T\mathbf{X})_{a\mu}}{(\mathbf{F}^T\mathbf{F}\mathbf{G})_{a\mu}} \qquad (4)$$

$$F_{ia} \leftarrow \frac{(\mathbf{X}^T\mathbf{G}^T)_{ia}}{(\mathbf{F}\mathbf{G}\mathbf{G}^T)_{ia}} \qquad (5)$$

where $G_{a\mu}$ (or, equivalently $(\mathbf{G})_{a\mu}$) denotes the (a, μ)-th entry of $\mathbf{G}$.

Here we employ a similar technique to optimize the objective in (12) with respect to (w.r.t.) $\mathbf{F}$, $\mathbf{S}$, and $\mathbf{G}$ in an alternating fashion. First, it is observed that the optimization w.r.t. $\mathbf{F}$ with other variables fixed can be equivalently written as follows; see also [10]

$$\min_{\mathbf{F}\geq 0} \|\mathbf{X}^T - \mathbf{G}\mathbf{S}^T\mathbf{F}^T\|_F^2 + \lambda_F \|\mathbf{F}\|_1^2 \\ = \min_{\mathbf{F}\geq 0} \left\| \begin{bmatrix} \mathrm{vec}(\mathbf{X}^T) \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{I} \otimes (\mathbf{G}\mathbf{S}^T) \\ \sqrt{\lambda_F}\mathbf{1}^T \end{bmatrix} \mathrm{vec}(\mathbf{F}^T) \right\|_2^2 \qquad (6)$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{1}$ a column vector with all entries equal to 1. Applying the multiplicative update rule to the cost in (6) yields

$$F_{a\mu} \leftarrow F_{a\mu} \frac{\left(\mathbf{X}\mathbf{G}\mathbf{S}^T\right)_{a\mu}}{(\mathbf{F}\mathbf{S}\mathbf{G}^T\mathbf{G}\mathbf{S}^T)_{a\mu} + \lambda_F \|\mathbf{F}\|_1}. \qquad (7)$$

Similarly, the optimization w.r.t. $\mathbf{S}$ is equivalent to

$$\min_{\mathbf{S}\geq 0}\left\|\begin{bmatrix}\mathrm{vec}(\mathbf{X})\\ 0\end{bmatrix}-\begin{bmatrix}\mathbf{G}\otimes\mathbf{F}\\ \sqrt{\lambda_S}\mathbf{1}^T\end{bmatrix}\mathrm{vec}(\mathbf{S})\right\|_2^2 \tag{8}$$

from which one can derive the update rule

$$S_{\mu b}\leftarrow S_{\mu b}\frac{\left(\mathbf{F}^T\mathbf{X}\mathbf{G}\right)_{\mu b}}{(\mathbf{F}^T\mathbf{F}\mathbf{S}\mathbf{G}^T\mathbf{G})_{\mu b}+\lambda_S\|\mathbf{S}\|_1}. \tag{9}$$

Likewise, the update rule for $\mathbf{G}$ is given by

$$G_{ib}\leftarrow G_{ib}\frac{\left(\mathbf{X}^T\mathbf{F}\mathbf{S}\right)_{ib}}{(\mathbf{G}\mathbf{S}^T\mathbf{F}^T\mathbf{F}\mathbf{S})_{ib}+\lambda_G\|\mathbf{G}\|_1}. \tag{10}$$

In practice, one adds a small $\epsilon > 0$ (say, 10^{-10}) to the denominators in (7), (9)–(10) to prevent division by zero.

C. Avoiding Inadmissible Zeros

One of the issues associated with multiplicative update rules is that when an entry in the factors becomes zero, it is forever stuck at zero. In theory, the entries should never become zero provided that all the entries in the factors are initialized with positive values. However, due to the finite precision of calculations, zero entries may well appear in practice. Therefore, the fixed point obtained from multiplicative updates might not satisfy the Karush-Kuhn-Tucker (KKT) optimality conditions.

To avoid such "inadmissible zeros," the KKT conditions can be examined during iterations [1]. The KKT conditions for (3) can be written as

$$\mathbf{F}\geq 0,\quad \mathbf{S}\geq 0,\quad \mathbf{G}\geq 0 \tag{11a}$$

$$(\mathbf{F}\mathbf{S}\mathbf{G}^T-\mathbf{X})\mathbf{G}\mathbf{S}^T+\lambda_F\|\mathbf{F}\|_1\mathbf{1}\mathbf{1}^T\geq 0 \tag{11b}$$

$$\mathbf{F}^T(\mathbf{F}\mathbf{S}\mathbf{G}^T-\mathbf{X})\mathbf{G}+\lambda_S\|\mathbf{S}\|_1\mathbf{1}\mathbf{1}^T\geq 0 \tag{11c}$$

$$(\mathbf{G}\mathbf{S}^T\mathbf{F}^T-\mathbf{X}^T)\mathbf{F}\mathbf{S}+\lambda_G\|\mathbf{G}\|_1\mathbf{1}\mathbf{1}^T\geq 0 \tag{11d}$$

$$\left[(\mathbf{F}\mathbf{S}\mathbf{G}^T-\mathbf{X})\mathbf{G}\mathbf{S}^T+\lambda_F\|\mathbf{F}\|_1\mathbf{1}\mathbf{1}^T\right]\odot\mathbf{F}=0 \tag{11e}$$

$$\left[\mathbf{F}^T(\mathbf{F}\mathbf{S}\mathbf{G}^T-\mathbf{X})\mathbf{G}+\lambda_S\|\mathbf{S}\|_1\mathbf{1}\mathbf{1}^T\right]\odot\mathbf{S}=0 \tag{11f}$$

$$\left[(\mathbf{G}\mathbf{S}^T\mathbf{F}^T-\mathbf{X}^T)\mathbf{F}\mathbf{S}+\lambda_G\|\mathbf{G}\|_1\mathbf{1}\mathbf{1}^T\right]\odot\mathbf{G}=0 \tag{11g}$$

where $\odot$ represents element-wise multiplication.

It can be deduced that the KKT optimality conditions are equivalent to: *i)* if $F_{a\mu} > 0$ (likewise, $S_{\mu b} > 0$ or $G_{ib} > 0$), the multiplicative factor in (7) ((9) or (10), respectively) must be equal to 1; and *ii)* otherwise, the corresponding factor must be less than or equal to 1.

Thus, whenever an entry in $\mathbf{F}$, $\mathbf{S}$ or $\mathbf{G}$ is zero, one can check the corresponding factor. If the factor is greater than 1, the zero entry is replaced by a small positive number κ to prevent convergence to an inadmissible fixed point. The overall algorithm is tabulated in Table I.

```
Parameters: $\kappa > 0$, $\kappa_{\mathrm{tol}} > 0$, $\epsilon > 0$ all small;
            $\lambda_F \geq 0$, $\lambda_S \geq 0$, $\lambda_G \geq 0$ [and $\lambda_O \geq 0$]
Initialize $\mathbf{F}$, $\mathbf{S}$, and $\mathbf{G}$ with positive entries
    [For the robust version: set $\mathbf{O} = 0$]
While not converged, repeat:
    $\alpha_{a\mu} = \frac{(\mathbf{X}\mathbf{G}\mathbf{S}^T)_{a\mu}}{(\mathbf{F}\mathbf{S}\mathbf{G}^T\mathbf{G}\mathbf{S}^T)_{a\mu}+\lambda_F\|\mathbf{F}\|_1+\epsilon}$
        [For robustness, replace $\mathbf{X}$ by $\mathbf{X}-\mathbf{O}$]
    $\tilde{F}_{a\mu} = \begin{cases}\kappa, & \text{if } F_{a\mu} < \kappa_{\mathrm{tol}} \text{ and } \alpha_{a\mu} > 1\\ 0, & \text{otherwise}\end{cases}$
    $\mathbf{F} \leftarrow (\mathbf{F}+\tilde{\mathbf{F}})\odot\boldsymbol{\alpha}$
    $\beta_{\mu b} = \frac{(\mathbf{F}^T\mathbf{X}\mathbf{G})_{\mu b}}{(\mathbf{F}^T\mathbf{F}\mathbf{S}\mathbf{G}^T\mathbf{G})_{\mu b}+\lambda_S\|\mathbf{S}\|_1+\epsilon}$
        [For robustness, replace $\mathbf{X}$ by $\mathbf{X}-\mathbf{O}$]
    $\tilde{S}_{\mu b} = \begin{cases}\kappa, & \text{if } S_{\mu b} < \kappa_{\mathrm{tol}} \text{ and } \beta_{\mu b} > 1\\ 0, & \text{otherwise}\end{cases}$
    $\mathbf{S} \leftarrow (\mathbf{S}+\tilde{\mathbf{S}})\odot\boldsymbol{\beta}$
    $\gamma_{ib} = \frac{(\mathbf{X}^T\mathbf{F}\mathbf{S})_{ib}}{(\mathbf{G}\mathbf{S}^T\mathbf{F}^T\mathbf{F}\mathbf{S})_{ib}+\lambda_G\|\mathbf{G}\|_1+\epsilon}$
        [For robustness, replace $\mathbf{X}$ by $\mathbf{X}-\mathbf{O}$]
    $\tilde{G}_{ib} = \begin{cases}\kappa, & \text{if } G_{ib} < \kappa_{\mathrm{tol}} \text{ and } \gamma_{ib} > 1\\ 0, & \text{otherwise}\end{cases}$
    $\mathbf{G} \leftarrow (\mathbf{G}+\tilde{\mathbf{G}})\odot\boldsymbol{\gamma}$
    [For robustness: update $\mathbf{O}$ using (13)]
```

TABLE I
SPARSE (ROBUST) NMTF ALGORITHM.

III. Robust NMTF

In this section, NMTF that is robust against additive outliers in the data is developed. Outliers can originate from erroneous experiments or noisy data acquisition. Moreover, data points that do not conform to the NMTF structure can also be determined as outliers.

The robust (and sparse) NMTF problem can be formulated using a sparse outlier matrix $\mathbf{O}\in\mathbb{R}^{m\times n}$, as

$$\min_{\mathbf{F}\geq 0,\mathbf{S}\geq 0,\mathbf{G}\geq 0,\mathbf{X}-\mathbf{O}\geq 0}\frac{1}{2}\left(\|\mathbf{X}-\mathbf{F}\mathbf{S}\mathbf{G}^T-\mathbf{O}\|_F^2 +\lambda_F\|\mathbf{F}\|_1^2+\lambda_S\|\mathbf{S}\|_1^2+\lambda_G\|\mathbf{G}\|_1^2\right)+\lambda_O\|\mathbf{O}\|_1. \tag{12}$$

Nonzero entries in $\mathbf{O}$ indicate the location of the data points that are not conforming to the NMTF [7]. In (12), the entries of $\mathbf{X}-\mathbf{O}$, which correspond to the data after compensating for the outliers, are constrained to be nonnegative to preserve the nonnegative modality of the original data.

To develop a simple algorithm for robust sparse NMTF, alternating minimization is again adopted. Thus, the multiplicative updates w.r.t. $\mathbf{F}$, $\mathbf{S}$ and $\mathbf{G}$ are identical to (7), (9) and (10), respectively, except that $\mathbf{X}$ is replaced by $\mathbf{X}-\mathbf{O}$. To perform the update w.r.t. $\mathbf{O}$, it is observed from (12) that the optimization w.r.t. $\mathbf{O}$ decouples to individual entries of $\mathbf{O}$. Therefore, a closed-form solution is obtained as

$$O_{ab}\leftarrow\min\left\{X_{ab},\mathcal{S}(X_{ab}-(\mathbf{F}\mathbf{S}\mathbf{G}^T)_{ab},\lambda_O)\right\} \tag{13}$$

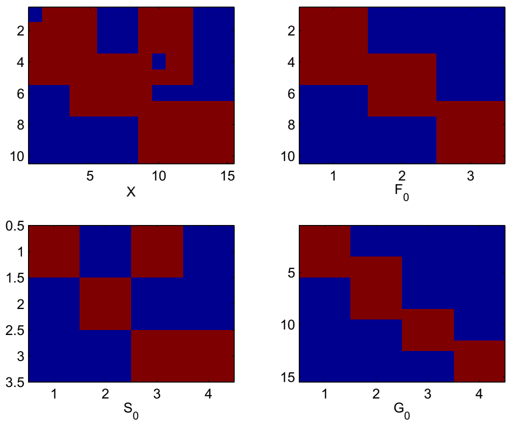


Fig. 1. Data matrix and true matrix factors.

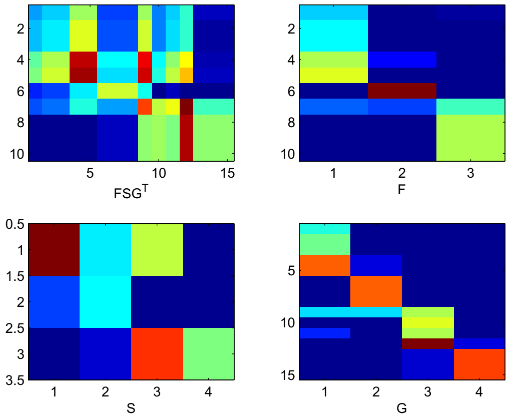


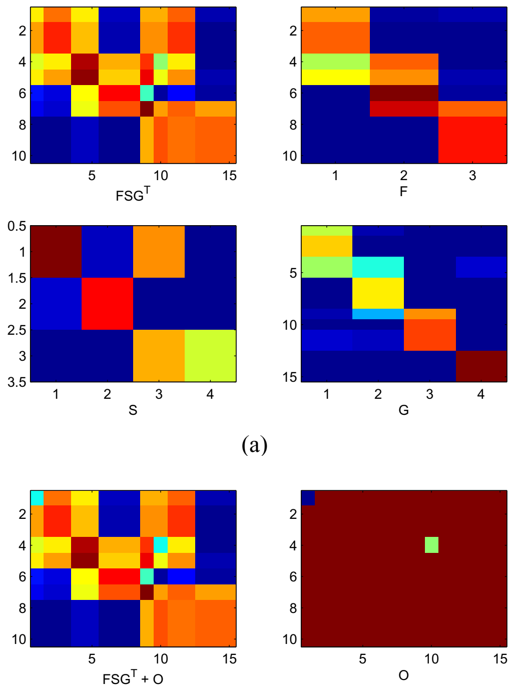
Fig. 2. Orthogonal NMTF [18].

where $\mathcal{S}(x, \lambda) \triangleq \operatorname{sign}(x) \max\{|x| - \lambda, 0\}$ is the soft thresholding function. The overall algorithm is given in Table I, with the instructions in the brackets followed.

IV. Tests with Synthetic Data

The proposed robust sparse NMTF algorithm is illustrated using a simple synthetic example. Fig. 1 depicts the true factor matrices $\mathbf{F}_0$, $\mathbf{S}_0$ and $\mathbf{G}_0$ with $m = 10$, $n = 15$, $k_1 = 3$ and $k_2 = 4$ (best viewed in colors). The entries of the factors assume binary values: the dark red areas represent entries with value 1, and dark blue areas correspond to value 0. To test the case of overlapping clusters, the columns of $\mathbf{F}_0$ and $\mathbf{G}_0$ were chosen to be non-orthogonal. To generate the data matrix $\mathbf{X}$, first $\mathbf{F}_0\mathbf{S}_0\mathbf{G}_0^T$ was formed, followed by binary quantization. Then, entries $X_{1,1}$ and $X_{4,10}$, which were originally ones, were flipped to zeros to test robustness. The resulting $\mathbf{X}$ is shown in the upper-left panel in Fig. 1.

As a benchmark, the orthogonal NMTF in [18] was employed on $\mathbf{X}$ with results shown in Fig. 2. Although relevant clusters are roughly identified, it can be seen that orthogonality constraints interfere with discovering correct clusters and associations. Also, many of the discovered structures are “bluish” (close to 0), and thus are not clearly contrasted from the dark blue background.



(a)

(b)

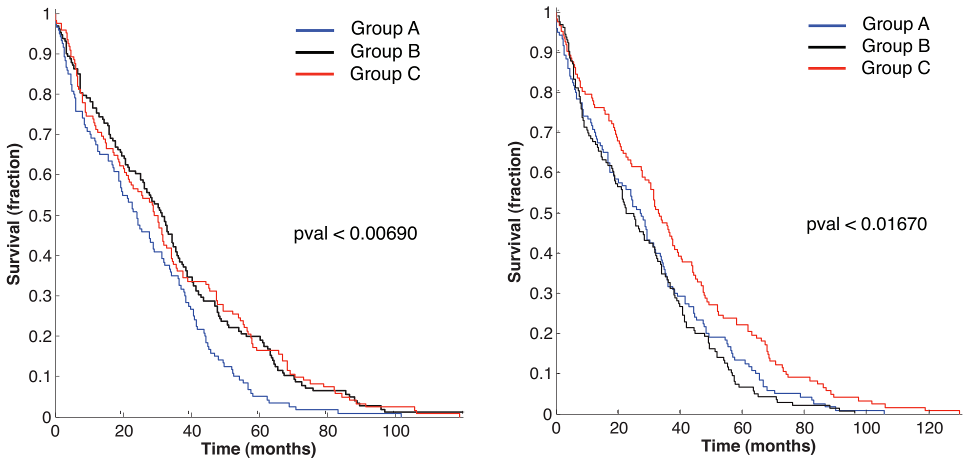
Fig. 3. Robust sparse NMTF.

It turns out that the novel NMTF algorithm is quite sensitive to initialization. It is customary to use basic clustering techniques such as k-means for initialization of NMF. In our test, orthogonal NMTF in Fig. 2 was used as initial factors. The resulting NMTF is shown in Fig. 3(a), where it can be seen that the correct structures are much more clearly identified (with “reddish” colors). Moreover, the overlaps in the clusters are better revealed.

The upper-left panel in Fig. 3(a) corresponds to the reconstruction *after* compensating for the outliers. The reconstruction *before* compensation is shown in the left panel of Fig. 3(b), which is closer to the original $\mathbf{X}$. In the right panel of Fig. 3(b) is shown the $\mathbf{O}$ matrix, where the dark red background now represents zeros, and the blue dots signify negative values. It can be seen that the locations of the outliers have been correctly indicated.

V. Application to Cancer Patient Clustering and Pathway Discovery

Identification of patient subpopulations that share common pathway activity is essential to understanding the complexities of genomic alterations, and to develop efficient therapeutic strategies (e.g., pathway-specific therapeutics) in cancer genomics. The proposed NMTF



(a) microarray gene expression

(b) copy number alteration

Fig. 4. Kaplan-Meier survival plots.

algorithm is used to: *i)* identify patient subgroups that have significantly different survival outcomes; and *ii)* assess the association of pathway activities with overall survival by integrating microarray gene expression or copy number alterations with pathway database.

A. Set-up

Microarray gene expression and copy number alteration data for ovarian cancer patients were collected from The Cancer Genome Project Atlas (TCGA) portal (http://cancergenome.nih.gov). Both datasets contain 377 patients and 11,094 gene expression or copy number alterations. Also collected were 186 KEGG pathways, which contain 5,267 genes in total, from a molecular signature database. The microarray gene expression and copy number alteration data were transformed to nonnegative input matrices $\mathbf{X}$ as follows [11]. The rows and the columns of the original data matrix $\mathbf{Y}$ represent patients and genes, respectively. For each nonnegative element $Y_{ij} \geq 0$ in the original matrix $\mathbf{Y}$, set $X_{i,2j-1} = Y_{ij}$ and $X_{i,2j} = 0$. For each negative element $Y_{ij} < 0$, set $X_{i,2j-1} = 0$ and $X_{i,2j} = -Y_{ij}$. Thus, matrix factor $\mathbf{F}$ indicates the patient clusters, where the number of patient clusters k_1 is fixed to 20. To simplify interpretation, factor $\mathbf{G}$ was fixed to the known gene pathways.

B. Results

First, we investigated whether subgroups of patients that correlate with different clinical outcomes, such as survival, could be identified. After learning patient cluster $\mathbf{F}$, the patients were divided into three groups by examining each column of $\mathbf{F}$. Specifically, the patients were ranked based on the magnitude of the entries in each column of $\mathbf{F}$, and the top 120 patients out of 377 were collected in Group A, the bottom 120 in Group C, and the rest in Group B. To find the subgroup of patients that strongly correlate with survival outcomes, Kaplan-Meier curves were generated by plotting the proportion of surviving patients versus the number of months after initial diagnosis.

Interestingly, patient clusters with statistically significant difference in survival outcomes could be identified from both microarray gene expression and copy number alteration datasets. For example, it was found that Group A patients had significantly less chance of survival compared to Groups B and C in the 17th cluster (column of $\mathbf{F}$) from the microarray gene expression dataset, as shown in Fig. 4(a). The logrank test indicates that Groups A, B and C patients indeed have significantly different survival outcomes with p-values less than 0.0069 (at hazard ratio 1.2448). The median survival time for Group A was 23.95 months, compared to 32 and 29.55 months for Groups B and C. Similarly, the three patient groups in the 12th cluster from the copy number dataset had p-value less than 0.0167 with hazard ratio 1.2963, as depicted in Fig. 4(b). The median survival times were 27.25, 22.80 and 33.05 months for Groups A, B, and C.

To identify the pathway activities associated with the patient subpopulations with different survival outcomes, different rows in $\mathbf{S}$ (i.e., pathway activities corresponding to patient clusters) were examined. Specifically, the pathways were ranked based on the magnitudes of the entries in each row of $\mathbf{S}$. In the microarray gene expression dataset, many cancer-related pathways were found to be associated with the patient subgroups in the 17th

TABLE II
TOP RANKED PATHWAY ACTIVITIES.

Ranking	Pathway (Microarray gene expression)	Pathway (Copy number alteration)
1	KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION	KEGG PATHWAYS IN CANCER
2	KEGG COMPLEMENT AND COAGULATION CASCADES	KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
3	KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	KEGG RIBOSOME
4	KEGG CELL ADHESION MOLECULES CAMS	KEGG CELL ADHESION MOLECULES CAMS
5	KEGG PATHWAYS IN CANCER	KEGG UBIQUITIN MEDIATED PROTEOLYSIS
6	KEGG PURINE METABOLISM	KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION
7	KEGG CHEMOKINE SIGNALING PATHWAY	KEGG MAPK SIGNALING PATHWAY
8	KEGG HEMATOPOIETIC CELL LINEAGE	KEGG WNT SIGNALING PATHWAY
9	KEGG MAPK SIGNALING PATHWAY	KEGG HUNTINGTONS DISEASE
10	KEGG TGF BETA SIGNALING PATHWAY	KEGG CHEMOKINE SIGNALING PATHWAY

patient cluster, including ‘pathways in cancer,’ ‘mitogen-activated protein kinase (MAPK) signaling pathway,’ and ‘transforming growth factor beta (TGF-beta) signaling pathway.’ Deregulation of activities in MAPK and TGF-beta signaling are known to be involved in many types of cancers including ovarian, breast, lung, prostate, and renal cancers [3], [15].

Likewise, many cancer-related pathways popped up in the top ranked pathways associated with patient subgroups in the 12th patient cluster in the copy number alteration dataset. These included ‘pathways in cancer,’ ‘MAPK signaling pathway,’ and ‘Wnt signaling pathway.’ Alteration of Wnt signaling pathway have been suggested to play a central role in ovarian tumorigenesis [6]. Moreover, recent studies showed that ‘cytokine cytokine receptor interaction’ and ‘neuroactive ligand receptor interaction’ pathways, which are highly ranked in both microarray gene expression and copy number alteration datasets, could play a major role in ovarian tumorigenesis and survival [14], [2].

The list of top ranked pathway associated with patient clusters are listed in Table II. These results suggest that the proposed method can allow stratification of cancers at the pathway level, potentially leading to development of more targeted therapeutics.

VI. CONCLUSIONS

Sparsity-promoting NMTF was formulated and multiplicative rule-based algorithms were developed with provisions to avoid non-stationary fixed points. Compared to the NMTF under orthogonality constraints, the proposed method was shown to be effective in revealing overlapping clusters. A robust version was also derived based on the inherent sparsity of outliers. The outlier identification capability could determine the data points that do not conform to the NMTF structure. Novel application of the method to microarray gene expression and copy number alteration data of cancer patients showed promise in discovering relevant patient subgroups and associated critical pathways, potentially useful for targeted therapeutics.

REFERENCES

[1] E. C. Chi and T. G. Kolda, “On tensors, sparsity, and nonnegative factorizations,” Dec. 2011. [Online]. Available: http://arxiv.org/abs/1112.2414

[2] A. P. G. Crijns, R. S. N. Fehrmann *et al.*, “Survival-related profile, pathways, and transcription factors in ovarian cancer,” *PLoS Medicine*, vol. 6, no. 2, 2009.

[3] A. S. Dhillon, S. Hagan, W. Kolch, and O. Rath, “MAP kinase signalling pathways in cancer,” *Oncogene*, vol. 26, no. 22, pp. 3279–3290, 2007.

[4] C. Ding, X. He, and H. D. Simon, “On the equivalence of nonnegative matrix factorization and spectral clustering,” in *Proc. of the SIAM Data Mining Conf.*, Newport Beach, CA, Apr. 2005.

[5] C. Ding, T. Li, W. Peng, and H. Park, “Orthogonal nonnegative matrix tri-factorization for clustering,” in *Proc. of the ACM KDD Conf.*, Philadelphia, PA, Aug. 2006, pp. 126–135.

[6] T. A. Gatcliffe, B. J. Monk, K. Planutis, and R. F. Holcombe, “Wnt signaling in ovarian tumorigenesis,” *Int’l. J. Gynecological Cancer*, vol. 18, no. 5, pp. 954–962, 2008.

[7] G. B. Giannakis, G. Mateos, S. Farahmand, and H. Zhu, “USPACOR: Universal sparsity controlling outlier rejections,” in *Proc. of the ICASSP Conf.*, May 2011, pp. 1952–1955.

[8] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *J. Machine Learning Res.*, vol. 5, pp. 1457–1469, Nov. 2004.

[9] Y. Jin, E. Sharafuddin, and Z.-L. Zhang, “Unveiling core network-wide communication patterns through application traffic activity graph decomposition,” in *Proc. of the SIGMETRICS Conf.*, Seattle, WA, Jun. 2009, pp. 49–60.

[10] H. Kim and H. Park, “Non-negative matrix factorization based on alternating non-negative constrained least squares and active set method,” *SIAM J. Matrix Anal. and Appl.*, vol. 30, no. 2, pp. 713–730, 2008.

[11] P. M. Kim and B. Tidor, “Subsystem identification through dimensionality reduction of large-scale gene expression data,” *Genome Res.*, vol. 13, no. 7, pp. 1706–1718, Jul. 2003.

[12] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[13] ——, “Algorithms for non-negative matrix factorization,” in *Adv. in Neural Info. Proc. Syst.*, vol. 13. MIT Press, 2001.

[14] J. A. Malek, E. Mery *et al.*, “Copy number variation analysis of matched ovarian primary tumors and peritoneal metastasis,” *PLoS ONE*, vol. 6, no. 12, Dec. 2011.

[15] J. Massagué, “TGFβ in cancer,” *Cell*, vol. 134, no. 2, pp. 215–230, 2008.

[16] E. E. Papalexakis and N. D. Sidiropoulos, “Co-clustering as multilinear decomposition with sparse latent factors,” in *Proc. of the ICASSP Conf.*, May 2011, pp. 2064–2067.

[17] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proc. of the ACM SIGIR Conf.*, Toronto, Canada, Jul.-Aug. 2003, pp. 267–273.

[18] J. Yoo and S. Choi, “Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds,” *Information Processing and Management*, vol. 46, pp. 559–570, 2010.