Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing

Benjamin D. Haeffele BHAEFFELE@JHU.EDU
Eric D. Young EYOUNG@JHU.EDU
René Vidal RVIDAL@JHU.EDU
Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland USA

Abstract

Recently, convex solutions to low-rank matrix factorization problems have received increasing attention in machine learning. However, in many applications the data can display other structures beyond simply being low-rank. For example, images and videos present complex spatio-temporal structures, which are largely ignored by current low-rank methods. In this paper we explore a matrix factorization technique suitable for large datasets that captures additional structure in the factors by using a projective tensor norm, which includes classical image regularizers such as total variation and the nuclear norm as particular cases. Although the resulting optimization problem is not convex, we show that under certain conditions on the factors, any local minimizer for the factors yields a global minimizer for their product. Examples in biomedical video segmentation and hyperspectral compressed recovery show the advantages of our approach on high-dimensional datasets.

1. Introduction

In many large datasets the relevant information often lies in a low-dimensional subspace of the ambient space, leading to a large interest in representing data with low-rank approximations. A common formulation for this problem is as a regularized loss problem of the form

$$\min_X \ell(Y, X) + \lambda R(X), \tag{1}$$

where $Y \in \mathbb{R}^{t\times p}$ is the data matrix, $X \in \mathbb{R}^{t\times p}$ is the low-rank approximation, $\ell(\cdot)$ is a loss function that measures how well X approximates Y, and $R(\cdot)$ is a regularization function that promotes various desired properties in X (low-rank, sparsity, group-sparsity, etc.). When ℓ and R are convex functions of X, and the dimensions of Y are not too large, the above problem can be solved efficiently using existing algorithms, which have achieved impressive results. However, when t is the number of frames in a video and p is the number of pixels, for example, optimizing over $O(tp)$ variables can be prohibitive.

To address this, one can exploit the fact that if X is low-rank, then there exist matrices $A \in \mathbb{R}^{t\times r}$ and $Z \in \mathbb{R}^{p\times r}$ (which we will refer to as the column and row spaces of X, respectively) such that $Y \approx X = AZ^T$ and $r \ll \min(t, p)$. This leads to the following *matrix factorization* problem, in which we search for A and Z that minimize

$$\min_{A,Z} \ell(Y, AZ^T) + \lambda\tilde{R}(A, Z), \tag{2}$$

where $\tilde{R}(\cdot,\cdot)$ is now a regularizer on the factors A and Z. Notice that by working directly with a factorized formulation such as (2), we can reduce the size of the optimization problem from $O(tp)$ to $O(r(t + p))$. Additionally, in many applications of low-rank modeling the factors obtained from the factorization often contain information relevant to the problem and can be used as features for further analysis, such as in classical PCA. Placing regularization directly on the factors thus allows one to promote additional structure on the factorized matrices A and Z beyond simply being a low-rank approximation, e.g. in sparse dictionary learning the matrix Z should be sparse. However, the price to be paid for these advantages is that the resulting optimization problems are typically not convex due to the product of A and Z, which poses significant challenges.

Despite the growing availability of tools for low-rank recovery and approximation and the utility of deriving features from low-rank representations, many techniques fail to incorporate additional information about the underlying row and columns spaces which are often known *a priori*. In computer vision, for example, a collection of images of

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

an object taken under different illuminations has not only a low-rank representation (Basri & Jacobs, 2003), but also significant spatial structure relating to the statistics of the scene, such as sparseness on a particular wavelet basis or low total variation (Rudin et al., 1992).

To capture this additional structure in the problem, we explore a low-rank matrix factorization technique based on several very interesting formulations which have been proposed to provide convex relaxations of structured matrix factorization (Bach et al., 2008; Bach, 2013). While our proposed technique is not convex, we show that a rank-deficient local minimum gives a global minimum, suggest an optimization strategy which is highly parallelizable and can be performed using a potentially highly reduced set of variables, and illustrate the advantages of our approach for large scale problems with examples in biomedical video segmentation and hyperspectral compressed recovery.

2. Background and Preliminaries

2.1. Notation

For $q \in [1, \infty]$, we denote the l_q norm of a vector $x \in \mathbb{R}^t$ as $\|x\|_q = (\sum_{i=1}^t |x_i|^q)^{1/q}$, where x_i is the ith entry of x. Also, we denote the ith column of a matrix $X \in \mathbb{R}^{t \times p}$ by X_i, its trace as $\text{Tr}(X)$, and its Frobenius norm as $\|X\|_F$. For a function $W(X)$, we denote its Fenchel dual as

$$W^*(X) \equiv \sup_Z \text{Tr}(Z^T X) - W(Z). \tag{3}$$

For a norm $\|X\|$, with some abuse of notation we denote its dual norm as $\|X\|^* \equiv \sup_{\|Z\| \leq 1} \text{Tr}(Z^T X)$. The space of $n \times n$ positive semidefinite matrices is denoted as S_n^+. For a function f, if f is non-convex we use ∂f to denote the general subgradient of f; if f is convex the general subgradient is equivalent to the regular subgradient and will also be denoted as ∂f, with the specific subgradient definition being known from the context (see Rockafellar & Wets, 2009, Chap. 8).

2.2. Proximal Operators

In our optimization algorithm we will make use of proximal operators, which are defined as follows.

Definition 1 *The proximal operator of a closed convex function $\theta(x)$ is defined as*

$$\textbf{prox}_\theta(y) \equiv \arg\min_x \frac{1}{2}\|y - x\|_2^2 + \theta(x). \tag{4}$$

2.3. Projective Tensor Norm

To find structured matrix factorizations, we will use the following matrix norm.

Definition 2 *Given vector norms $\|\cdot\|_a$ and $\|\cdot\|_z$, the Projective Tensor Norm of a matrix $X \in \mathbb{R}^{t \times p}$ is defined as*

$$\|X\|_P \equiv \inf_{A,Z:AZ^T=X} \sum_i \|A_i\|_a \|Z_i\|_z \tag{5}$$

$$= \inf_{A,Z:AZ^T=X} \frac{1}{2} \sum_i (\|A_i\|_a^2 + \|Z_i\|_z^2). \tag{6}$$

It can be shown that $\|X\|_P$ is a valid norm on X; however, a critical point is that for general norms $\|\cdot\|_a$ and $\|\cdot\|_z$ the summation in (5) and (6) might need to be over an infinite number of columns of A and Z (Bach et al., 2008; Ryan, 2002, Sec. 2.1). A particular case where this sum is known to be bounded is when $\|\cdot\|_a = \|\cdot\|_2$ and $\|\cdot\|_z = \|\cdot\|_2$. In this case $\|X\|_P$ reverts to the nuclear norm $\|X\|_*$ (sum of singular values of X), which is widely used as a convex relaxation of matrix rank and can optimally recover low-rank matrices under certain conditions (Recht et al., 2010).

More generally, the projective tensor norm provides a natural framework for *structured matrix factorizations*, where appropriate norms can be chosen to reflect the desired properties of the row and column spaces of the matrix. For instance, the projective tensor norm was studied in the context of sparse dictionary learning, where it was referred to as the Decomposition Norm (Bach et al., 2008). In this case, one can use combinations of the l_1 and l_2 norms to produce a tradeoff between the number of factorized elements (number of columns in A and Z) and the sparsity of the factorized elements (Bach et al., 2008). Finally, recent work has shown that the projective tensor norm can be considered a special case of a much more general matrix factorization framework based on gauge functions. This allows additional structure to be placed on the factors A and Z (for example non-negativity), while still resulting in a convex regularizer, offering significant potential extensions for future work (Bach, 2013).

3. Structured Matrix Factorizations

Motivated by the introductory discussion, in this section we describe the link between traditional convex loss problems (1), which offer guarantees of global optimality, and factorized formulations (2), which offer additional flexibility in modeling the data structure and recovery of features that can be used in subsequent analysis. Following (Bach et al., 2008), we use the projective tensor norm as a regularizer, leading to the following structured low-rank matrix factorization problem.

$$\min_X \ell(Y, X) + \lambda \|X\|_P. \tag{7}$$

Given the definition of the projective tensor norm, this problem is equivalently minimized by solutions to the non-

convex problem (see supplement)

$$\min_{A,Z} \ell(Y, AZ^T) + \lambda \sum_i \|A_i\|_a \|Z_i\|_z. \tag{8}$$

Since we are interested in capturing certain structures in the column and row spaces of X, while at the same time capturing low-rank structures in X, in this paper we consider norms of the form

$$\begin{aligned} \|\cdot\|_a &= \nu_a \|\cdot\|_{\tilde{a}} + \|\cdot\|_2 \\ \|\cdot\|_z &= \nu_z \|\cdot\|_{\tilde{z}} + \|\cdot\|_2. \end{aligned} \tag{9}$$

Here $\|\cdot\|_{\tilde{a}}$ and $\|\cdot\|_{\tilde{z}}$ are norms that model the desired properties of the column and row spaces of X, respectively, and ν_a and ν_z balance the tradeoff between those desired properties and the rank of the solution (recall that when $\nu_a = \nu_z = 0$, $\|X\|_P$ reduces to the nuclear norm $\|X\|_*$).

3.1. Matrix Factorization as Semidefinite Optimization

While (7) is a convex function of the product $X = AZ^T$, it is still non-convex with respect to A and Z jointly. However, if we define a matrix Q to be the concatenation of A and Z

$$Q \equiv \begin{bmatrix} A \\ Z \end{bmatrix} \implies QQ^T = \begin{bmatrix} AA^T & AZ^T \\ ZA^T & ZZ^T \end{bmatrix}, \tag{10}$$

we see that AZ^T is a submatrix of the positive semidefinite matrix QQ^T. After defining the function $F : S_n^+ \to \mathbb{R}$

$$F(QQ^T) = \ell(Y, AZ^T) + \lambda \|AZ^T\|_P, \tag{11}$$

it is clear that the proposed formulation (7) can be recast as an optimization over a positive semidefinite matrix $X = QQ^T$. At first this seems to be a circular argument, since while $F(X)$ is a convex function of X, this says nothing about finding Q (or A and Z). However, recent results for semidefinite programs in standard form show that one can minimize $F(X)$ by solving for Q directly without introducing any additional local minima, provided that the rank of Q is larger than the rank of the true solution X^{true} (Burer & Monteiro, 2005)[1]. Additionally, if the rank of the true solution is not known *a priori*, the following key result shows that when $F(X)$ is twice differentiable, it is often possible to optimize $F(QQ^T)$ with respect to Q and still be assured of a global minimum.

Theorem 1 *(Bach et al., 2008, Prop. 4) Let $F : S_n^+ \to \mathbb{R}$ be a twice differentiable convex function with compact level sets. If Q is a rank deficient local minimum of $f(Q) = F(QQ^T)$, then $X = QQ^T$ is a global minimum of $F(X)$.*

[1]Note, however, that for general norms $\|\cdot\|_a$ and $\|\cdot\|_z$ Q may not contain a factorization which achieves the infimum in (5)

Unfortunately, while many common loss functions are convex and twice differentiable, for the problems we study here we cannot directly apply this result due to the fact that the projective tensor norm is clearly non-differentiable for general norms $\|\cdot\|_a$ and $\|\cdot\|_z$. In what follows we extend the above result to the non-differentiable case and describe an algorithm to minimize (8) suitable to large problems.

3.2. Local Minima Achieve Global Minimum

In this subsection, we extend the results from Theorem 1 to functions $F : S_n^+ \to \mathbb{R}$ of the form

$$F(X) = G(X) + H(X), \tag{12}$$

where $G : S_n^+ \to \mathbb{R}$ is a twice differentiable convex function with compact level sets and $H : S_n^+ \to \mathbb{R}$ is a (possibly non-differentiable) proper convex function such that F is lower semi-continuous. Before presenting our main result, define $g(Q) = G(QQ^T)$, $h(Q) = H(QQ^T)$, $f(Q) = g(Q) + h(Q) = F(QQ^T)$ and note the following.

Lemma 1 *If Q is a local minimum of $f(Q) = F(QQ^T)$, where $F : S_n^+ \to \mathbb{R}$ is a function form in (12), then $\exists \Lambda \in \partial H(QQ^T)$ such that $0 = 2\nabla G(QQ^T)Q + 2\Lambda Q$.*

Proof. If Q is a local minimum of $f(Q)$, then it is necessary that $0 \in \partial f(Q)$ (Rockafellar & Wets, 2009, Thm. 10.1). Let $V(Q) = QQ^T$. Then $\partial f(Q) \subseteq \nabla V(Q)^T \partial F(QQ^T) = \nabla V(Q)^T (\nabla G(QQ^T) + \partial H(QQ^T))$ (Rockafellar & Wets, 2009, Thm. 10.6). From the symmetry of $\nabla G(QQ^T)$ and $\partial H(QQ^T)$, we get $\nabla V(Q)^T \nabla G(QQ^T) = 2\nabla G(QQ^T)Q$ and $\nabla V(Q)^T \partial H(QQ^T) = 2\partial H(QQ^T)Q$, as claimed. ■

Theorem 2 *Let $F : S_n^+ \to \mathbb{R}$ be a function of the form in (12). If Q is a rank-deficient local minimum of $f(Q) = F(QQ^T)$, then $X = QQ^T$ is a global minimum of $F(X)$.*

Proof. We begin by introducing another variable subject to an equality constraint

$$\min_{X \succeq 0} G(X) + H(X) = \min_{X \succeq 0, Y} G(X) + H(Y) \text{ s.t. } X = Y. \tag{13}$$

This gives the Lagrangian

$$L(X, Y, \Lambda) = G(X) + H(Y) + \text{Tr}(\Lambda^T (X - Y)). \tag{14}$$

Minimizing the Lagrangian w.r.t. Y we obtain

$$\min_Y H(Y) - \text{Tr}(\Lambda^T Y) = -H^*(\Lambda) \tag{15}$$

Let $k(Q, \Lambda) = G(QQ^T) + \text{Tr}(\Lambda^T QQ^T)$ and let X_Λ denote a value of X which minimizes the Lagrangian w.r.t. X for a fixed value of Λ. Assuming strong duality, we have

$$\min_{X \succeq 0} F(X) = \max_\Lambda \min_{X \succeq 0} G(X) + \text{Tr}(\Lambda^T X) - H^*(\Lambda). \tag{16}$$

Algorithm 1 (Structured Low-Rank Approximation)

Input: Y, A^0, Z^0, λ, NumIter
Initialize $\hat{A}^1 = A^0$, $\hat{Z}^1 = Z^0$
for $k = 1$ **to** NumIter **do**
  \\Calculate gradient of loss function w.r.t. A
  \\evaluated at the extrapolated point $\hat{A}$
  $G_A^k = \nabla_A \ell(Y, \hat{A}^k (Z^k)^T)$
  $P = \hat{A}^k - G_A^k / L_A^k$
  \\Calculate proximal operator of $\|\cdot\|_a$
  \\for every column of A
  for $i = 1$ **to** number of columns in A **do**
    $A_i^k = \mathbf{prox}_{\lambda \|Z_i\|_z \|\cdot\|_a / L_A^k}(P_i)$
  end for
  \\Repeat similar process for Z
  $G_Z^k = \nabla_Z \ell(Y, A^k (\hat{Z}^k)^T)$
  $W = \hat{Z}^k - G_Z^k / L_Z^k$
  for $i = 1$ **to** number of columns in Z **do**
    $Z_i^k = \mathbf{prox}_{\lambda \|A_i\|_a \|\cdot\|_z / L_Z^k}(W_i)$
  end for
  \\Update extrapolation based on prior iterates
  $\hat{A}^{k+1} = Extrapolate_A(A^k, A^{k-1})$
  $\hat{Z}^{k+1} = Extrapolate_Z(Z^k, Z^{k-1})$
end for

From Theorem 1, if we fix the value of Λ, then a rank-deficient local minimum of $k(Q, \Lambda)$ minimizes the Lagrangian w.r.t. X for $X_\Lambda = QQ^T$. In particular, if we fix Λ such that it satisfies Lemma 1, we then have $\frac{\partial}{\partial Q} k(Q, \Lambda) = 2\nabla G(QQ^T)Q + 2\Lambda Q = 0$, so $X_\Lambda = QQ^T$ is a global minimum of the Lagrangian w.r.t. X for a fixed Λ that satisfies Lemma 1. Additionally, since we chose Λ to satisfy Lemma 1, then we have $\Lambda \in \partial H(QQ^T) \Rightarrow X_\Lambda = QQ^T \in \partial H^*(\Lambda)$ (due to the easily shown fact that $X_\Lambda \in \partial H^*(\Lambda) \Leftrightarrow \Lambda \in \partial H(X_\Lambda)$). Combining these results, we have that (QQ^T, Λ) is a primal-dual saddle point, so $X = QQ^T$ is a global minimum of $F(X)$. ■

3.3. Minimization Algorithm

Before we begin the discussion of our algorithm, we note that the particular method we present here assumes that the gradients of the loss function $\ell(Y, AZ^T)$ w.r.t. A and w.r.t. Z (denoted as $\nabla_A \ell(Y, AZ^T)$ and $\nabla_Z \ell(Y, AZ^T)$, respectively) are Lipschitz continuous with Lipschitz constants L_A^k and L_Z^k (in general the Lipschitz constant of the gradient will depend on the current value of the variables at that iteration, hence the superscript). Under these assumptions on ℓ, the bilinear structure of our objective function (8) gives convex subproblems if we update A or Z independently while holding the other fixed, making an alternating minimization strategy efficient and easy to implement. Specifically, the updates to our variables are made using accelerated proximal-linear steps similar to the FISTA algorithm, which entails solving a proximal operator of an extrapolated gradient step to update each variable (Beck & Teboulle, 2009; Xu & Yin, 2013). The general structure of the alternating minimization we use is given in Algorithm 1 (full details can be found in (Xu & Yin, 2013)), but the key point is that to update either A or Z the primary computational burden lies in calculating the gradient of the loss function and then calculating a proximal operator. The structure of the non-differentiable term in (8) allows the proximal operator to be separated into columns, greatly reducing the complexity of calculating the proximal operator and offering the potential for parallelization. Moreover, the following result provides a simple method to calculate the proximal operator of the l_2 norm combined with any norm.

Theorem 3 *Let $\|\cdot\|$ be any vector norm. The proximal operator of $\theta(x) = \lambda\|x\| + \lambda_2\|x\|_2$ is the composition of the proximal operator of the l_2 norm and the proximal operator of $\|\cdot\|$, i.e.,* $\mathbf{prox}_\theta(y) = \mathbf{prox}_{\lambda_2\|\cdot\|_2}(\mathbf{prox}_{\lambda\|\cdot\|}(y))$.

Proof. See supplement ■

Combining these results with Theorem 2, we have a potential strategy to search for structured low-rank matrix factorizations as we only need to find a rank-deficient local minimum to conclude that we have found a global minimum. However, there are a few critical caveats to note about the optimization problem. First, alternating minimization does not guarantee convergence to a local minimum. It has been shown that, subject to a few conditions[2], block convex functions will globally converge to a Nash equilibrium point via the alternating minimization algorithm we use here, and any local minima must also be a Nash equilibrium point (although unfortunately the converse is not true) (Xu & Yin, 2013). Of practical importance, this implies that multiple stationary points which are not local minima can be encountered and the variables cannot be initialized arbitrarily. For example, $(A, Z) = (0, 0)$ is a Nash equilibrium point of (8).[3] Nevertheless, we observe that empirically we obtain good results in our studied applications with very trivial initializations.

Second, although it can be shown that the projective tensor norm defined by (5) is a valid norm if the sum is taken over a potentially infinite number of columns of A and Z, for general vector norms $\|\cdot\|_a$ and $\|\cdot\|_z$ it is not necessarily known *a priori* if a finite number of columns of A and Z can achieve the infimum. Here we conjecture that for norms of the form given in (9) the infimum of (5) can

[2]The objective function as we have presented it in (8) does not meet these conditions as the non-differentiable elements are not separated into summable blocks, but by using the equivalence between (5) and (6) it can easily be converted to a form that does.

[3]More details about the difficulties associated with the problem and some techniques to bound or approximate the projective tensor norm can be found in (Bach, 2013)

be achieved or closely approximated by summing over a number of columns equal to the rank of AZ^T (again recall the equivalence with the nuclear norm when $\nu_a = \nu_z = 0$). We also note good empirical results by setting the number of columns of A and Z to be larger than the expected rank of the solution but smaller than full rank, a strategy that has been shown to be optimally convergent for semidefinite programs in standard form (Burer & Monteiro, 2005).

4. Applications

In this section we demonstrate our matrix factorization method on two image processing problems: spatiotemporal segmentation of neural calcium imaging data and hyperspectral compressed recovery. Such problems are well modeled by low-rank linear models with square loss functions under the assumption that the spatial component of the data has low total variation (and is optionally sparse in the row and/or column space). Specifically, in this section we consider the following objective

$$\min_{A,Z} \frac{1}{2}\|Y - \Phi(AZ^T)\|_F^2 + \lambda \sum_i \|A_i\|_a \|Z_i\|_z \tag{17}$$

$$\|\cdot\|_a = \nu_a \|\cdot\|_1 + \|\cdot\|_2 \tag{18}$$

$$\|\cdot\|_z = \nu_{z_1} \|\cdot\|_1 + \nu_{z_{TV}} \|\cdot\|_{TV} + \|\cdot\|_2, \tag{19}$$

where $\Phi(\cdot)$ is a linear operator, and ν_a, ν_{z_1} and $\nu_{z_{TV}}$ are non-negative scalars[4]. Recall that the anisotropic total variation of x is defined as (Birkholz, 2011)

$$\|x\|_{TV} \equiv \sum_i \sum_{j \in N_i} |x_i - x_j| \tag{20}$$

where N_i denotes the set of pixels in the neighborhood of pixel i.

4.1. Neural Calcium Imaging Segmentation

Calcium imaging is a rapidly growing microscopy technique in neuroscience that records fluorescent images from neurons that have been loaded with either synthetic or genetically encoded fluorescent calcium indicator molecules. When a neuron fires an electrical action potential (or spike), calcium enters the cell and binds to the fluorescent calcium indicator molecules, changing the fluorescence properties of the molecule. By recording movies of the calcium-induced fluorescent dynamics it is possible to infer the spiking activity from large populations of neurons with single neuron resolution (Stosiek et al., 2003). If we are given the fluorescence time series from a single neuron, inferring the spiking activity from the fluorescence time series is well modeled via a Lasso style estimation,

$$\hat{s} = \arg\min_{s \geq 0} \frac{1}{2}\|y - Ds\|_2^2 + \lambda \|s\|_1, \tag{21}$$

where $y \in \mathbb{R}^t$ is the fluorescence time series (normalized by the baseline fluorescence), $\hat{s} \in \mathbb{R}^t$ denotes the estimated spiking activity (each entry of $\hat{s}$ is monotonically related to the number of action potentials the neuron has during that imaging frame), and $D \in \mathbb{R}^{t \times t}$ is a matrix that applies a convolution with a known decaying exponential to model the change in fluorescence resulting from a neural action potential (Vogelstein et al., 2010). One of the challenges in neural calcium imaging is that the data can have a significant noise level, making manual segmentation challenging. Additionally, it is also possible to have two neurons overlap in the spatial domain if the focal plane of the microscope is thicker than the size of the distinct neural structures in the data, making simultaneous spatiotemporal segmentation necessary. A possible strategy to address these issues would be to extend (21) to estimate spiking activity for the whole data volume via the objective

$$\hat{S} = \arg\min_{S \geq 0} \frac{1}{2}\|Y - DS\|_F^2 + \lambda \|S\|_1, \tag{22}$$

where now each column of $Y \in \mathbb{R}^{t \times p}$ contains the fluorescent time series for a single pixel and the corresponding column of $\hat{S} \in \mathbb{R}^{t \times p}$ contains the estimated spiking activity for that pixel. However, due to the significant noise often present in the actual data, solving (22) directly typically gives poor results. To address this issue, Pnevmatikakis et al. (2013) have suggested adding an additional low-rank regularization to (22) based on the knowledge that if two pixels are from the same neural structure they should have identical spiking activities, giving S a low-rank structure with the rank of S corresponding to the number of neural structures in the data. Specifically, they propose an objective to promote low-rank and sparse spike estimates,

$$\hat{S} = \arg\min_{S \geq 0} \frac{1}{2}\|Y - DS\|_F^2 + \lambda \|S\|_1 + \lambda_2 \|S\|_* \tag{23}$$

and then estimate the temporal and spatial features by performing a non-negative matrix factorization of $\hat{S}$.

It can be shown that problem (17) is equivalent to a standard Lasso estimation when both the row space and column space are regularized by the l_1 norm (Bach et al., 2008), while combined l_1, l_2 norms of the form (18) and (19) with $\nu_{z_{TV}} = 0$ promote solutions that are simultaneously sparse and low rank. Thus, the projective tensor norm can generalize the two prior methods for calcium image processing by providing regularizations that are sparse or simultaneously sparse and low-rank. Here we further extend these formulations by noting that if two pixels are neighboring

[4] It is straightforward to extend the method to include non-negative constrains on A and Z, but we found this had little effect on the experimental results. The results presented here are all without constraints on the sign for simplicity of presentation.

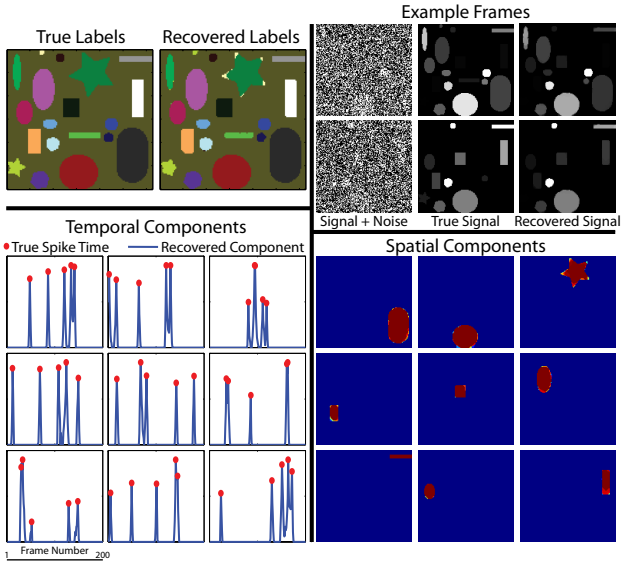

Figure 1. Results of experiment with phantom calcium imaging dataset. *Top Left*: True regions and regions recovered via k-means clustering on the spatial components. *Top Right*: Two example frames from the dataset, showing the signal with added noise (left), true signal (middle), and recovered signal (right). *Bottom Left*: First 9 most significant recovered temporal components (columns of A). The estimated temporal feature is shown as a blue line, while the true spike times are shown as red dots. *Bottom Right*: First 9 most significant spatial features (columns of Z).

each other it is likely that they are from the same neural structure and thus have identical spiking activity, implying low total variation in the spatial domain. We demonstrate the flexible nature of our formulation (17) by using it to process calcium image data with regularizations that are either sparse, simultaneously sparse and low-rank, or simultaneously sparse, low-rank, and with low total variation. Additionally, by optimizing (17) to simultaneously estimate temporal spiking activity A and neuron shape Z, with $\Phi(AZ^T) = DAZ^T$, we inherently find spatial and temporal features in the data (which are largely non-negative even though we do not explicitly constrain them to be) directly from our optimization without the need for an additional matrix factorization step.

Simulation Data. We first tested our algorithm on a simulated phantom using the combined sparse, low-rank, and total variation regularization. The phantom was constructed with 19 non-overlapping spatial regions and 5 randomly timed action potentials and corresponding calcium dynamics per region. Gaussian white noise was added to the modeled calcium signal to produce an SNR of approximately -16dB (see top panels of Fig. 1). We initialized A to be an identity matrix and $Z = 0$.[5] Despite the high levels of noise and simple initialization, the recovered spatial factors (columns of Z) corresponded to the actual region shapes and the recovered temporal factors (columns of A) showed strong peaks near the true spike times (Fig. 1, bottom panels). Additionally, simple k-means clustering on the columns of Z recovered the true region labels with high accuracy (Fig. 1, top left panel), and, although we do not specifically enforce non-negative entries in A and Z, the recovered matrices had no negative entries.

***In vivo* Calcium Image Data.** We next tested our algorithm on actual calcium image data taken *in vivo* from the primary auditory cortex of a mouse that was transfected with the genetic calcium indicator GCaMP5 (Akerboom et al., 2012). The left panel of Figure 2 shows 5 manually labeled regions from the dataset (top row) and the corresponding spatial features recovered by our algorithm (bottom 3 rows) under the various regularization conditions. The right panel of Figure 2 displays a frame from the dataset taken at a time point when the corresponding region had a significant calcium signal, with the actual data shown in the top row and the corresponding reconstructed calcium signal for that time point under the various regularization conditions shown in the bottom 3 rows. We note that regions 1 and 2 correspond to the cell body and a dendritic branch of the same neuron. The manual labeling was purposefully split into two regions due to the fact that dendrites can have significantly different calcium dynamics from the cell body and thus it is often appropriate to treat calcium signals from dendrites as separate features from the cell body (Spruston, 2008).

The data shown in Figure 2 are particularly challenging to segment as the two large cell bodies (regions 1 and 3) are largely overlapping in space, necessitating a spatiotemporal segmentation. In addition to the overlapping cell bodies there are various small dendritic processes radiating perpendicular to (regions 4 and 5) and across (region 2) the focal plane that lie in close proximity to each other and have significant calcium transients. Additionally, at one point during the dataset the animal moves, generating a large artifact in the data. Nevertheless, optimizing (17) under the various regularization conditions, we observe that, as expected, the spatial features recovered by sparse regularization alone are highly noisy (Fig. 2, row 2). Adding low-rank regularization improves the recovered spatial features, but the features are still highly pixelated and contain numerous pixels outside of the desired regions (Fig. 2, row 3). Finally, by incorporating the total variation regularization our method produces coherent spatial features which are highly similar to the desired manual labelings (Fig. 2,

[5]For this application the first update of our alternating minimization was applied to Z, instead of A as shown in Algorithm 1.

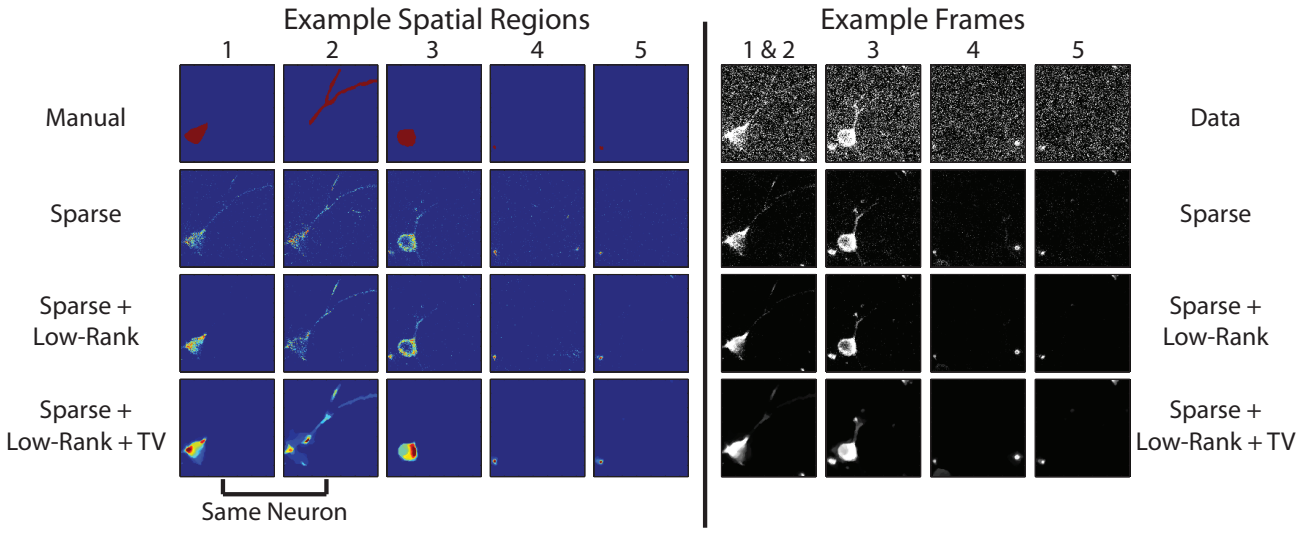

Figure 2. Results from the *in vivo* calcium imaging dataset. *Left*: Demonstration of spatial features for 5 example regions. (Top Row) Manually segmented regions. (Bottom 3 Rows) Corresponding spatial feature recovered by our method with various regularizations. Note that regions 1 and 2 are different parts of the same neurons - see discussion in the text . *Right*: Example frames from the dataset corresponding to time points where the example regions display a significant calcium signal. (Top Row) Actual Data. (Bottom 3 Rows) Estimated signal for the example frame with various regularizations.

rows 1 and 4), noting again that these features are found directly from the alternating minimization of (17) without the need to solve a secondary matrix factorization. For the two cases with low-rank regularization, A was initialized to be 100 uniformly sampled columns from an identity matrix (out of a possible 559), demonstrating the potential to reduce the problem size and achieve good results despite a very trivial initialization.

We conclude by noting that while adding total variation regularization improves performance for a segmentation task, it also can cause a dilative effect when reconstructing the estimated calcium signal (for example, distorting the size of the thin dendritic processes in the left two columns of the example frames in Figure 2). As a result, in a denoising task it might instead be desirable to only impose sparse and low-rank regularization. The fact that we can easily and efficiently adapt our model to account for many different features of the data depending on the desired task highlights the flexible nature and unifying framework of our proposed formulation (17).

4.2. Hyperspectral Compressed Recovery

In hyperspectral imaging (HSI), the data volume often displays a low-rank structure due to significant correlations in the spectra of neighboring pixels (Zhang et al., 2013). This fact, combined with the large data sizes typically encountered in HSI applications, has led to a large interest in developing compressed sampling and recovery techniques to compactly collect and reconstruct HSI datasets. In addition, the spatial domain of an HSI dataset typically can be modeled under the assumption that it displays properties common to natural scenes, which led Golbabaee & Vandergheynst (2012) to propose a combined nuclear norm and total variation regularization (NucTV) of the form

$$\min_X \|X\|_* + \lambda \sum_{i=1}^{t} \|(X^i)^T\|_{TV} \text{ s.t. } \|Y - \Phi(X)\|_F^2 \leq \epsilon. \quad (24)$$

Here $X \in \mathbb{R}^{t \times p}$ is the estimated HSI reconstruction with t spectral bands and p pixels, X^i denotes the ith row of X (or the ith spectral band), $Y \in \mathbb{R}^{t \times m}$ contains the observed samples (compressed at a subsampling ratio of m/p), and $\Phi(\cdot)$ denotes the compressed sampling operator. To solve (24), Golbabaee & Vandergheynst (2012) implemented a proximal gradient method, which required solving a total variation proximal operator for every spectral slice of the data volume in addition to solving the proximal operator of the nuclear norm (singular value thresholding) at every iteration of the algorithm (Combettes & Pesquet, 2011). For the large data volumes typically encountered in HSI, this can require significant computation per iteration. Here we demonstrate the use of our matrix factorization method to perform hyperspectral compressed recovery by optimizing (17), where $\Phi(\cdot)$ is a compressive sampling function that applies a random-phase spatial convolution at each wavelength (Romberg, 2009; Golbabaee & Vandergheynst, 2012), A contains estimated spectral features, and Z contains estimated spatial abundance features.[6] Compressed

[6]For HSI experiments, we set $\nu_a = \nu_{z_1} = 0$ in (18) and (19).

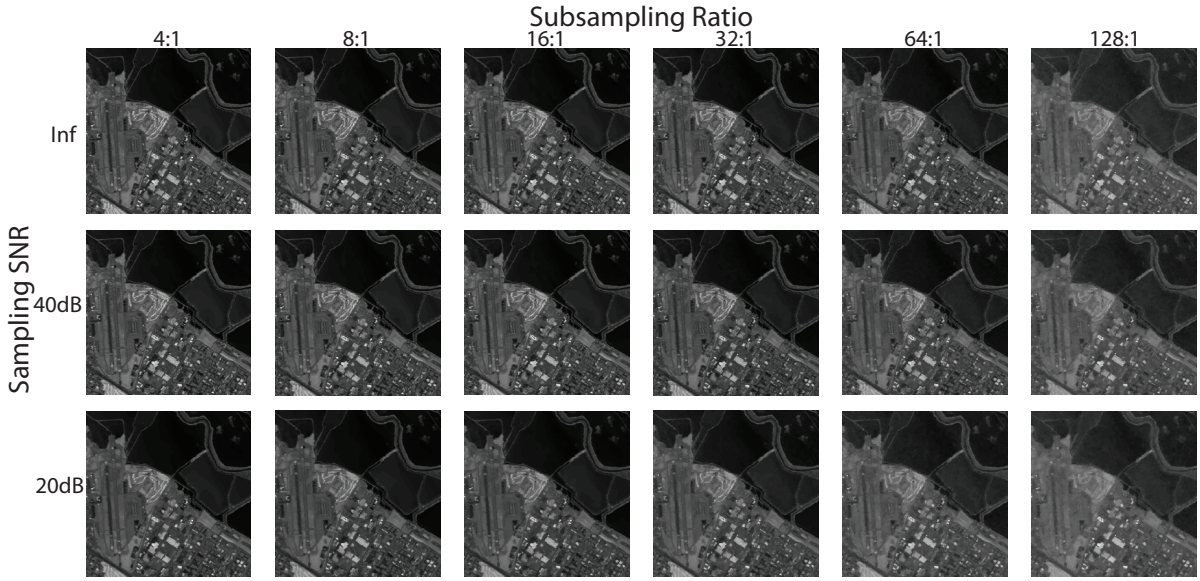

Figure 3. Hyperspectral compressed results. Example reconstructions from a single spectral band ($i = 50$) under different subsampling ratios and sampling noise levels. Compare with Golbabaee & Vandergheynst (2012, Fig. 2).

recovery experiments were performed on the dataset from Golbabaee & Vandergheynst (2012)[7] at various subsampling ratios and with different levels of sampling noise. We limited the number of columns of A and Z to 15 (the dataset is 256×256 pixels and 180 spectral bands), initialized one randomly selected pixel per column of Z to one and all others to zero, and initialized A as $A = 0$.

Figure 3 shows examples of the recovered images at one wavelength (spectral band $i = 50$) for various subsampling ratios and sampling noise levels and Table 1 shows the reconstruction recovery rates $\left\|X_{true} - AZ^T\right\|_F / \left\|X_{true}\right\|_F$. We note that even though we optimized over a highly reduced set of variables ($[256 \times 256 \times 15 + 180 \times 15]/[256 \times 256 \times 180] \approx 8.4\%$) with very trivial initializations, we were able to achieve reconstruction error rates equivalent to or better than those in Golbabaee & Vandergheynst (2012)[8]. Additionally, by solving the reconstruction in a factorized form, our method offers the potential to perform blind hyperspectral unmixing directly from the compressed samples without ever needing to reconstruct the full dataset, an application extension we leave for future work.

[7]The data used are a subset of the publicly available AVARIS Moffet Field dataset. We made an effort to match the specific spatial area and spectral bands of the data for our experiments to that used in (Golbabaee & Vandergheynst, 2012) but note that slightly different data may have been used in our study.

[8]The entries for NucTV in Table 1 were adapted from (Golbabaee & Vandergheynst, 2012, Fig. 1)

Table 1. Hyperspectral imaging compressed recovery error rates.

	Our Method			NucTV		
Sample Ratio	Sampling SNR (dB)			Sampling SNR (dB)		
	∞	40	20	∞	40	20
4:1	0.0209	0.0206	0.0565	0.01	0.02	0.06
8:1	0.0223	0.0226	0.0589	0.03	0.04	0.08
16:1	0.0268	0.0271	0.0663	0.09	0.09	0.13
32:1	0.0393	0.0453	0.0743	0.21	0.21	0.24
64:1	0.0657	0.0669	0.1010			
128:1	0.1140	0.1186	0.1400			

5. Conclusions

We have proposed a highly flexible approach to projective tensor norm matrix factorization, which allows specific structure to be promoted directly on the factors. While our proposed formulation is not jointly convex in all of the variables, we have shown that under certain criteria a local minimum of the factorization is sufficient to find a global minimum of the product, offering the potential to solve the factorization using a highly reduced set of variables.

Acknowledgments

We thank John Issa and David Yue for their help in acquiring the calcium image data and the anonymous reviewers for their helpful comments. We also thank the support of NIH grants DC00115 and DC00032, a grant from the Kleberg Foundation, and NSF grant 11-1218709.

References

Akerboom, J., Chen, T.-W., Wardill, T. J., Tian, L., Marvin, J. S., Mutlu, S., $\cdots$, and Looger, L. L. Optimization of a GCaMP calcium indicator for neural activity imaging. *The Journal of Neuroscience*, 32:13819–13840, 2012.

Bach, F. Convex relaxations of structured matrix factorizations. *arXiv:1309.3117v1*, 2013.

Bach, F., Mairal, J., and Ponce, J. Convex sparse matrix factorizations. *arXiv:0812.1869v1*, 2008.

Basri, R. and Jacobs, D. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.

Birkholz, H. A unifying approach to isotropic and anisotropic total variation denoising models. *J. of Computational and Applied Mathematics*, 235:2502–2514, 2011.

Burer, S. and Monteiro, R. D. C. Local minima and convergence in low-rank semidenite programming. *Mathematical Programming, Series A*, (103):427–444, 2005.

Combettes, P. L. and Pesquet, J. C. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49, pp. 185–212. Springer-Verlag, 2011.

Golbabaee, M. and Vandergheynst, P. Joint trace/tv minimization: A new efficient approach for spectral compressive imaging. In *19th IEEE Internation Conference on Image Processing*, pp. 933–936, 2012.

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in Optimization*, 1:123–231, 2013.

Pnevmatikakis, E. A., Machado, T. A., Grosenick, L., Poole, B., Vogelstein, J. T., and Paninski, L. Rank-penalized nonnegative spatiotemporal deconvolution and demixing of calcium imaging data. Abstract: Computational and Systems Neuroscience (Cosyne), 2013.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501, 2010.

Rockafellar, R. T. and Wets, R. J-B. *Variational Analysis*. Springer, 3rd edition, 2009.

Romberg, J. Compressive sensing by random convolution. *SIAM J. Img. Sci.*, 2(4):1098–1128, November 2009. ISSN 1936-4954.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.

Ryan, R. A. *Introduction to Tensor Products of Banach Spaces*. Springer, 2002.

Spruston, N. Pyramidal neurons: Dendritic structure and synaptic integration. *Nature Reviews Neuroscience*, 9: 206–221, 2008.

Stosiek, C., Garaschuk, O., Holthoff, K., and Konnerth, A. In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12):7319–7324, 2003.

Vogelstein, J. T., Packer, A. M., Machado, T. A., Sippy, T., Babadi, B., Yuste, R., and Paninski, L. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology*, 104(6): 3691–3704, 2010.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal of Imaging Sciences*, 6(3):1758–1789, 2013.

Zhang, H., He, W., Zhang, L., Shen, H., and Yuan, Q. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Trans. on Geoscience and Remote Sensing*, PP:1–15, 2013.