Robust Kronecker-Decomposable Component Analysis for Low-Rank Modeling

Mehdi Bahri[1] Yannis Panagakis[1,2] Stefanos Zafeiriou[1,3]
[1]Imperial College London, UK [2]Middlesex University, London, UK [3]University of Oulu, Finland
mehdi.b.tn@gmail.com, {i.panagakis, s.zafeiriou}@imperial.ac.uk

Abstract

Dictionary learning and component analysis are part of one of the most well-studied and active research fields, at the intersection of signal and image processing, computer vision, and statistical machine learning. In dictionary learning, the current methods of choice are arguably K-SVD and its variants, which learn a dictionary (i.e., a decomposition) for sparse coding via Singular Value Decomposition. In robust component analysis, leading methods derive from Principal Component Pursuit (PCP), which recovers a low-rank matrix from sparse corruptions of unknown magnitude and support. However, K-SVD is sensitive to the presence of noise and outliers in the training set. Additionally, PCP does not provide a dictionary that respects the structure of the data (e.g., images), and requires expensive SVD computations when solved by convex relaxation. In this paper, we introduce a new robust decomposition of images by combining ideas from sparse dictionary learning and PCP. We propose a novel Kronecker-decomposable component analysis which is robust to gross corruption, can be used for low-rank modeling, and leverages separability to solve significantly smaller problems. We design an efficient learning algorithm by drawing links with a restricted form of tensor factorization. The effectiveness of the proposed approach is demonstrated on real-world applications, namely background subtraction and image denoising, by performing a thorough comparison with the current state of the art.

1. Introduction

Sparse dictionary learning [40, 45, 34] and Robust Principal Component Analysis (RPCA) [6] are two popular methods of learning representations that assume different structure and answer different needs, but both are special cases of structured matrix factorization. Assuming a set of N data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ represented as the columns of a matrix $\mathbf{X}$, we seek a decomposition of $\mathbf{X}$ into meaningful components of a given structure, in the generic form:

$$\min_{\mathbf{Z}} l(\mathbf{X}, \mathbf{Z}) + g(\mathbf{Z}) \tag{1}$$

Where $l(\cdot)$ is a loss function, generally the ℓ_2 error, and $g(\cdot)$ a possibly non-smooth regularizer that encourages the desired structure, often in the form of specific sparsity requirements.

When $\mathbf{Z}$ is taken to be factorized in the form $\mathbf{DR}$, we obtain a range of different models depending on the choice of the regularization. Imposing sparsity on the dictionary $\mathbf{D}$ leads to sparse PCA, while sparsity only in the code $\mathbf{R}$ yields sparse dictionary learning. The current methods of choice for sparse dictionary learning are K-SVD [1] and its variants [32], which seek an overcomplete dictionary $\mathbf{D}$ and a sparse representation $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_n]$ by minimizing the following constrained objective:

$$\min_{\mathbf{R},\mathbf{D}} ||\mathbf{X} - \mathbf{DR}||_{\mathrm{F}}^2, \text{ s.t. } ||\mathbf{r}_i||_0 \leq T_0, \tag{2}$$

where $||.||_{\mathrm{F}}$ is the Frobenius norm and $||.||_0$ is the ℓ_0 pseudo-norm, counting the number of non-zero elements. Problem (2) is solved in an iterative manner that alternates between sparse coding of the data samples on the current dictionary, and a process of updating the dictionary atoms to better fit the data using the Singular Value Decomposition (SVD). When used to reconstruct images, K-SVD is trained on overlapping image patches to allow for overcompleteness [14]. Problems of this form suffer from a high computational burden that limits their applicability to small patches that only capture local information, and prevent them from scaling to larger images.

The recent developments in robust component analysis are attributed to the advancements in compressive sensing [7, 13, 8]. In this area, the emblematic model is the Robust Principal Component Analysis, proposed in [6]. RPCA assumes that the observation matrix $\mathbf{X}$ is the sum of a low-rank component $\mathbf{A}$, and of a sparse matrix $\mathbf{E}$ that collects the gross errors, or outliers. Finding the minimal rank solution for $\mathbf{A}$ and the sparsest solution for $\mathbf{E}$ is combinatorial and NP-hard. Therefore, RPCA is obtained by solving the Principal Component Pursuit (PCP) in which the discrete rank and ℓ_0-norm functions are approximated by their convex envelopes as follows:

$$\min_{\mathbf{A},\mathbf{E}} ||\mathbf{A}||_* + ||\mathbf{E}||_1, \text{ s.t. } \mathbf{X} = \mathbf{A} + \mathbf{E} \tag{3}$$

where $||.||_*$ is the nuclear norm and $||.||_1$ is the standard ℓ_1 norm. Like sparse dictionary learning, Robust PCA is a special case of (1). The nuclear-norm relaxation is convex but the cost of performing SVDs on the full-size matrix $\mathbf{A}$ is high. Factorization-based formulations exploit the fact that a rank-r matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ can be decomposed in $\mathbf{A} = \mathbf{U}\mathbf{V}^\mathsf{T}$, $\mathbf{U} \in \mathbb{R}^{m\times r}, \mathbf{V} \in \mathbb{R}^{n\times r}$, to impose low-rankness. These formulations are non-convex and algorithms may get stuck in local optima or saddle points. We refer to [28] for recent developments on the convergence of non-convex matrix factorizations, [44, 36, 50, 47] for low-rank problems, and [16] specifically for matrix completion.

1.1. Separable dictionaries

Analytical separable dictionaries have been proposed as generalizations of 1-dimensional transforms in the form, *e.g.*, of tensor-product wavelets, and have since fallen out of flavour for models that drop orthogonality in favour of overcompleteness to achieve geometric invariance [40]. However, recent work in the compressed sensing literature has shown a regain of interest for separable dictionaries for very high-dimensional signals where scalability is a hard-requirement. Notably, high-resolution spatial angular representations of diffusion-MRI signals, such as in HARDI (*High Angular Resolution Diffusion Imaging*), represent each voxel by a 3D signal, yielding prohibitive sampling times. Traditional compressed sensing techniques can only handle signal sparsity bounded below by the number of voxels (1 atom per voxel) and still cannot meet the needs of medical imaging, but Kronecker extensions [42, 41] of Orthogonal Matching Pursuit (OMP) [37, 11], Dual ADMM [5], and FISTA [4] allow sparser signals.

The recent Separable Dictionary Learning (SeDiL) [23] considers a dictionary that factorizes into the Kronecker product of two smaller dictionaries $\mathbf{A}$ and $\mathbf{B}$, and matrix observations $\mathcal{X} = (\mathbf{X}_i)_i$. The observations have sparse representations $\mathcal{R} = (\mathbf{R}_i)_i$ in the bases $\mathbf{A}, \mathbf{B}$, and the learning problem is recast as:

$$\min_{\mathbf{A},\mathbf{B},\boldsymbol{\mathcal{R}}} \frac{1}{2}\sum_i ||\mathbf{X}_i - \mathbf{A}\mathbf{R}_i\mathbf{B}^\mathsf{T}||_\mathrm{F}^2 + \lambda g(\boldsymbol{\mathcal{R}}) + \kappa r(\mathbf{A}) + \kappa r(\mathbf{B}) \tag{4}$$

Where the regularizers g and r promote, respectively, sparsity in the representations, and low mutual-coherence of the dictionary $\mathbf{D} = \mathbf{B} \otimes \mathbf{A}$. Here, $\mathbf{D}$ is constrained to have orthogonal columns, *i.e.*, the pair $\mathbf{A}, \mathbf{B}$ shall lie on the product manifold of two product of sphere manifolds. A different approach is taken in [25]. A separable 2D dictionary is learnt in a two-step strategy similar to that of K-SVD. Each matrix observation $\mathbf{X}_i$ is represented as $\mathbf{A}\mathbf{R}_i\mathbf{B}^\mathsf{T}$. In the first step, the sparse representations $\mathbf{R}_i$ are found by 2D OMP. In the second step, a CP decomposition is performed on a tensor of residuals via Regularized Alternating Least Squares to solve $\min_{\mathbf{A},\mathbf{B},\boldsymbol{\mathcal{R}}} ||\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{R}} \times_1 \mathbf{A} \times_2 \mathbf{B}||_\mathrm{F}$[1].

1.2. Contributions

In this paper, we propose a novel method for separable dictionary learning based on a robust tensor factorization that learns simultaneously the dictionary and the sparse representations. We extend the previous approaches by introducing an additional level of structure to make the model robust to outliers. We do not seek overcompleteness, but rather promote sparsity in the dictionary to learn a low-rank representation of the input tensor. In this regard, our method combines ideas from both Sparse Dictionary Learning and Robust PCA. We propose a non-convex parallelizable ADMM algorithm and provide experimental evidence of its effectiveness. Finally, we compare the performance of our method against several tensor and matrix factorization algorithms on computer vision benchmarks, and show our model systematically matches or outperforms the state of the art. We make the code available online[2] along with supplementary material.

Notations Throughout the paper, matrices (vectors) are denoted by uppercase (lowercase) boldface letters *e.g.*, $\mathbf{X}$, ($\mathbf{x}$). $\mathbf{I}$ denotes the identity matrix of compatible dimensions. The i^{th} column of $\mathbf{X}$ is denoted as $\mathbf{x}_i$. Tensors are considered as the multidimensional equivalent of matrices (second-order tensors), and vectors (first-order tensors), and denoted by bold calligraphic letters, *e.g.*, $\boldsymbol{\mathcal{X}}$. The *order* of a tensor is the number of indices needed to address its elements. Consequently, each element of an M^{th}-order tensor $\boldsymbol{\mathcal{X}}$ is addressed by M indices, *i.e.*, $(\boldsymbol{\mathcal{X}})_{i_1,i_2,\ldots,i_M} \doteq x_{i_1,i_2,\ldots,i_M}$.

The sets of real and integer numbers are denoted by $\mathbb{R}$ and $\mathbb{Z}$, respectively. An M^{th}-order real-valued tensor $\boldsymbol{\mathcal{X}}$ is defined over the tensor space $\mathbb{R}^{I_1\times I_2\times\cdots\times I_M}$, where $I_m \in \mathbb{Z}$ for $m = 1, 2, \ldots, M$.

The mode-n product of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_M}$ with a matrix $\mathbf{U} \in \mathbb{R}^{J\times I_n}$, denoted by $\boldsymbol{\mathcal{X}} \times_n \mathbf{U}$, is defined element-wise as

$$(\boldsymbol{\mathcal{X}} \times_m \mathbf{x})_{i_1,\ldots,i_{n-1},i_{n+1},\ldots,i_M} = \sum_{i_n=1}^{I_n} x_{i_1,i_2,\ldots,i_M} x_{i_n}. \tag{5}$$

The *Kronecker* product of matrices $\mathbf{A} \in \mathbb{R}^{I\times K}$, and $\mathbf{B} \in \mathbb{R}^{L\times M}$, is denoted by $\mathbf{A} \otimes \mathbf{B}$, and yields a matrix of dimensions $I \cdot L \times K \cdot M$.

Finally, we define the tensor *Tucker rank* as the vector of the ranks of its mode-n unfoldings (*i.e.*, its mode-n ranks), and the tensor *multi-rank* [48, 31] as the vector of the ranks of its frontal slices. More details about tensors, such as the definitions of tensor slices and mode-n unfoldings, can be found in [26] for example.

[1] *c.f.* notations for the product $\times_n$

[2] https://github.com/mbahri/KDRSDL

2. Model and updates

We first formulate our structured factorization, and describe a tailored optimization procedure.

2.1. Robustness and structure

Consider the following Sparse Dictionary Learning problem with Frobenius-norm regularization on the dictionary $\mathbf{D}$, where we decompose N observations $\mathbf{x}_i \in \mathbb{R}^{mn}$ on $\mathbf{D} \in \mathbb{R}^{mn \times r_1 r_2}$ with representations $\mathbf{r}_i \in \mathbb{R}^{r_1 r_2}$:

$$\min_{\mathbf{D},\mathbf{R}} \sum_i ||\mathbf{x}_i - \mathbf{D}\mathbf{r}_i||_2^2 + \lambda \sum_i ||\mathbf{r}_i||_1 + ||\mathbf{D}||_{\mathrm{F}} \quad (6)$$

We assume a Kronecker-decomposable dictionary $\mathbf{D} = \mathbf{B} \otimes \mathbf{A}$ with $\mathbf{A} \in \mathbb{R}^{m \times r_1}, \mathbf{B} \in \mathbb{R}^{n \times r_2}$. To model the presence of outliers, we introduce a set of vectors $\mathbf{e}_i \in \mathbb{R}^{mn}$ and, with $d = r_1 r_2 + mn$, define the block vectors and matrices:

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{r}_i \\ \mathbf{e}_i \end{bmatrix} \in \mathbb{R}^d \quad \mathbf{C} = \begin{bmatrix} \mathbf{B} \otimes \mathbf{A} & \mathbf{I} \end{bmatrix} \in \mathbb{R}^{mn \times d} \quad (7)$$

We obtain a two-level structured dictionary $\mathbf{C}$ and the associated sparse encodings $\mathbf{y}_i$. Breaking-down the variables to reduce dimensionality and discarding the constant $||\mathbf{I}||_{\mathrm{F}}$:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{R},\mathbf{E}} \sum_i ||\mathbf{x}_i - (\mathbf{B} \otimes \mathbf{A})\mathbf{r}_i - \mathbf{e}_i||_2^2 \\ &+\lambda \sum_i ||\mathbf{r}_i||_1 + \lambda \sum_i ||\mathbf{e}_i||_1 + ||\mathbf{B} \otimes \mathbf{A}||_{\mathrm{F}} \end{aligned} \quad (8)$$

Suppose now that the observations $\mathbf{x}_i$ were obtained by vectorizing two-dimensional data such as images, *i.e.*, $\mathbf{x}_i = \mathrm{vec}(\mathbf{X}_i), \mathbf{X}_i \in \mathbb{R}^{m \times n}$. We find preferable to keep the observations in matrix form as this preserves the spatial structure of the images, and - as we will see - allows us solve matrix equations efficiently instead of high-dimensional linear systems. Without loss of generality, we choose $r_1 = r_2 = r$ and $\mathbf{r}_i = \mathrm{vec}(\mathbf{R}_i), \mathbf{R}_i \in \mathbb{R}^{r \times r}$, and recast the problem as:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{E}}} \sum_i ||\mathbf{X}_i - \mathbf{A}\mathbf{R}_i\mathbf{B}^{\mathsf{T}} - \mathbf{E}_i||_{\mathrm{F}}^2 \\ &+\lambda \sum_i ||\mathbf{R}_i||_1 + \lambda \sum_i ||\mathbf{E}_i||_1 + ||\mathbf{B} \otimes \mathbf{A}||_{\mathrm{F}} \end{aligned} \quad (9)$$

Equivalently, enforcing the equality constraints and writing the problem explicitly as a structured tensor factorization:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{E}}} \quad \lambda||\boldsymbol{\mathcal{R}}||_1 + \lambda||\boldsymbol{\mathcal{E}}||_1 + ||\mathbf{B} \otimes \mathbf{A}||_{\mathrm{F}} \\ &\text{s.t} \qquad\quad \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{R}} \times_1 \mathbf{A} \times_2 \mathbf{B} + \boldsymbol{\mathcal{E}} \end{aligned} \quad (10)$$

Where the matrices $\mathbf{X}_i, \mathbf{R}_i$, and $\mathbf{E}_i$ are concatenated as the frontal slices of 3-way tensors. Figure 1 illustrates the decomposition.

We impose $r \leq \min(m, n)$ as a natural upper bound on the rank of the frontal slices of $\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{R}} \times_1 \mathbf{A} \times_2 \mathbf{B}$, and on its mode-1 and mode-2 ranks.

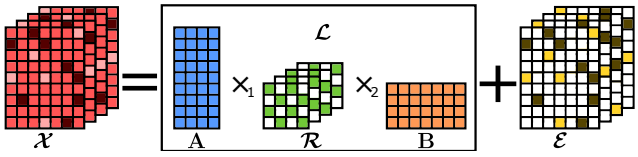


Figure 1: Illustration of the decomposition.

2.2. An efficient algorithm

Problem (10) is not jointly convex, but is convex in each component individually. We resort to an alternating-direction method and propose a non-convex ADMM procedure that operates on the frontal slices.

Minimizing $||\mathbf{B} \otimes \mathbf{A}||_{\mathrm{F}}$ presents a challenge: the product is high-dimensional, the two bases are coupled, and the loss is non-smooth. Using the identity $||\mathbf{A} \otimes \mathbf{B}||_p = ||\mathbf{A}||_p ||\mathbf{B}||_p$ where $||.||_p$ denotes the Schatten-p norm (this follows from the compatibility of the Kronecker product with the singular value decomposition), and remarking that $||\mathbf{B}||_{\mathrm{F}}||\mathbf{A}||_{\mathrm{F}} \leq \frac{||\mathbf{A}||_{\mathrm{F}}^2 + ||\mathbf{B}||_{\mathrm{F}}^2}{2}$, we minimize a simpler upper bound ($||\mathbf{A}\mathbf{B}^{\mathsf{T}}||_*$ [39]). The resulting sub-problems are smaller, and therefore more scalable. In order to obtain exact proximal steps for the encodings $\mathbf{R}_i$, we introduce a split variable $\mathbf{K}_i$ such that $\forall i,\ \mathbf{K}_i = \mathbf{R}_i$. Thus, we solve:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\boldsymbol{\mathcal{R}},\boldsymbol{\mathcal{K}}\boldsymbol{\mathcal{E}}} \quad \lambda||\boldsymbol{\mathcal{R}}||_1 + \lambda||\boldsymbol{\mathcal{E}}||_1 + \tfrac{1}{2}(||\mathbf{A}||_{\mathrm{F}}^2 + ||\mathbf{B}||_{\mathrm{F}}^2) \\ &\text{s.t} \qquad\quad \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{K}} \times_1 \mathbf{A} \times_2 \mathbf{B} + \boldsymbol{\mathcal{E}} \\ &\text{s.t} \qquad\quad \boldsymbol{\mathcal{R}} = \boldsymbol{\mathcal{K}} \end{aligned} \quad (11)$$

Introducing the tensors of Lagrange multipliers $\boldsymbol{\Lambda}$ and $\boldsymbol{\mathcal{Y}}$, such that the i^{th} frontal slice corresponds to the i^{th} constraint, and the dual step sizes μ and $\mu_{\boldsymbol{\mathcal{K}}}$, we formulate the Augmented Lagrangian of problem (11):

$$\begin{aligned} &\min \lambda \sum_i ||\mathbf{R}_i||_1 + \lambda \sum_i ||\mathbf{E}_i||_1 + \frac{1}{2}(||\mathbf{A}||_{\mathrm{F}}^2 + ||\mathbf{B}||_{\mathrm{F}}^2) + \\ &\sum_i \langle \boldsymbol{\Lambda}_i,\ \mathbf{X}_i - \mathbf{A}\mathbf{K}_i\mathbf{B}^{\mathsf{T}} - \mathbf{E}_i \rangle + \sum_i \langle \mathbf{Y}_i,\ \mathbf{R}_i - \mathbf{K}_i \rangle + \\ &\frac{\mu}{2} \sum_i ||\mathbf{X}_i - \mathbf{A}\mathbf{K}_i\mathbf{B}^{\mathsf{T}} - \mathbf{E}_i||_{\mathrm{F}}^2 + \frac{\mu_{\boldsymbol{\mathcal{K}}}}{2} \sum_i ||\mathbf{R}_i - \mathbf{K}_i||_{\mathrm{F}}^2 \end{aligned} \quad (12)$$

We can now derive the ADMM updates. Each $\mathbf{E}_i$ is given by shrinkage after rescaling:

$$\mathbf{E}_i = \mathcal{S}_{\lambda/\mu}(\mathbf{X}_i - \mathbf{A}\mathbf{K}_i\mathbf{B}^{\mathsf{T}} + \frac{1}{\mu}\boldsymbol{\Lambda}_i) \quad (13)$$

A similar rule is immediate to derive for $\mathbf{R}_i$, and solving for $\mathbf{A}$ and $\mathbf{B}$ is straightforward with some matrix algebra. We therefore focus on the computation of the split variable $\mathbf{K}_i$.

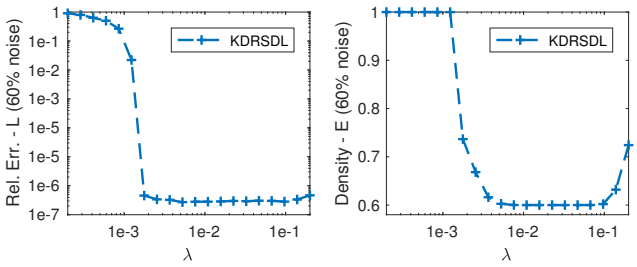


Figure 2: Recovery performance with 60% corruption. Relative ℓ_2 error and density.

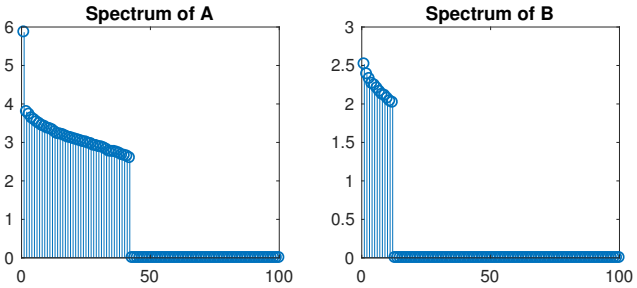


Figure 3: Sample spectrums of $\mathbf{A}$ and $\mathbf{B}$, the ground truth is 42 for $\mathbf{A}$ and 12 for $\mathbf{B}$, and is attained. $r = 100$.

Differentiating, we find $\mathbf{K}_i$ satisfies:

$$\mu_{\boldsymbol{\mathcal{K}}}\mathbf{K}_i + \mu\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{K}_i\mathbf{B}^\mathsf{T}\mathbf{B} \tag{14}$$
$$= \mathbf{A}^\mathsf{T}(\mathbf{\Lambda}_i + \mu(\mathbf{X}_i - \mathbf{E}_i))\mathbf{B} + \mu_{\boldsymbol{\mathcal{K}}}\mathbf{R}_i + \mathbf{Y}_i$$

The key for an efficient algorithm is here to recognize equation (14) is a *Stein* equation, and can be solved in cubical time and quadratic space in r by solvers for discrete-time Sylvester equations - such as the Hessenberg-Schur method [19] - instead of the naive $O(r^6)$ time, $O(r^4)$ space solution of vectorizing the equation in an r^2 linear system.

In practice, we solve a slightly different problem with $\alpha\sum_i ||\mathbf{R}_i||_1 + \lambda\sum_i ||\mathbf{E}_i||_1$. This introduces an additional degree of freedom and corresponds to rescaling the coefficients of $\mathbf{R}_i$. We found the modified problem to be numerically more stable and to allow for tuning of the relative importance of $||\boldsymbol{\mathcal{R}}||_1$ and $||\boldsymbol{\mathcal{E}}||_1$. We obtain Algorithm 1.

3. Discussion

We show experimentally that our algorithm successfully recovers the components of the decomposition (10), we then prove the formulation encourages two different notions of low-rankness on tensors, and finally discuss the issues of non-convexity and convergence.

3.1. Validation on synthetic data

We generated synthetic data following the model's assumptions by first sampling two random bases $\mathbf{A}$ and $\mathbf{B}$ of known ranks $r_\mathbf{A}$ and $r_\mathbf{B}$, N Gaussian slices for the core $\boldsymbol{\mathcal{R}}$, and forming the ground truth $\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{R}} \times_1 \mathbf{A} \times_2 \mathbf{B}$. We modeled additive random sparse Laplacian noise with a tensor $\boldsymbol{\mathcal{E}}$ whose entries are 0 with probability p, and 1 or -1 with equal probability otherwise. We generated data for $p = 70\%$ and $p = 40\%$, leading to a noise density of, respectively, 30% and 60%. We measured the reconstruction error on $\boldsymbol{\mathcal{L}}$, $\boldsymbol{\mathcal{E}}$, and the density of $\boldsymbol{\mathcal{E}}$ for varying values of λ, and $\alpha = 1\mathrm{e}{-2}$. Our model achieved near-exact recovery of both $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{E}}$, and exact recovery of the density of $\boldsymbol{\mathcal{E}}$, for suitable values of λ. Experimental evidence is presented in Figure 2 for the 60% noise case.

The algorithm appears robust to small changes in λ, which suggests not only one value can lead to optimal results, and that a simple criterion that provides consistently good reconstruction may be derived, as in Robust PCA [6]. In the 30% noise case, we did not observe an increase in the density of $\boldsymbol{\mathcal{E}}$ as λ increases, and the ℓ_2 error on both $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{L}}$ was of the order of $1\mathrm{e}{-7}$ (*c.f.* supplementary material).

3.2. Low-rank solutions

Seeing the model from the perspective of Robust PCA (3), which seeks a low-rank representation $\mathbf{A}$ of the dataset $\mathbf{X}$, we minimize the rank of the low-rank tensor $\boldsymbol{\mathcal{L}}$. More precisely, we show in theorem 3.1 that we simultaneously penalize the Tucker rank and the multi-rank of $\boldsymbol{\mathcal{L}}$.

Theorem 3.1. *Algorithm 1 encourages low mode-1 and mode-2 rank, and thus, low-rankness in each frontal slice of $\boldsymbol{\mathcal{L}}$, for suitable choices of the parameters λ and α.*

Proof. From the equivalence of norms in finite-dimensions, $\exists k \in \mathbb{R}_+^*, ||\mathbf{A}\otimes\mathbf{B}||_* \leq k||\mathbf{A}\otimes\mathbf{B}||_\mathrm{F}$. We minimise $\lambda||\boldsymbol{\mathcal{E}}||_1 + \alpha||\boldsymbol{\mathcal{R}}||_1 + ||\mathbf{A}\otimes\mathbf{B}||_\mathrm{F}$. By choosing $\alpha = \frac{\alpha'}{k}, \lambda = \frac{\lambda'}{k}$, we penalize $\mathrm{rank}(\mathbf{A}\otimes\mathbf{B}) = \mathrm{rank}(\mathbf{A})\mathrm{rank}(\mathbf{B})$. Given that the rank is a non-negative integer, $\mathrm{rank}(\mathbf{A})$ or $\mathrm{rank}(\mathbf{B})$ decreases necessarily. Therefore, we minimize the mode-1 and mode-2 ranks of $\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{R}} \times_1 \mathbf{A} \times_2 \mathbf{B}$. Additionally, $\forall i, \mathrm{rank}(\mathbf{A}\mathbf{R}_i\mathbf{B}^\mathsf{T}) \leq \min(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B}), \mathrm{rank}(\mathbf{R}_i))$.

Exhibiting a valid k may help in the choice of parameters. We show $\forall\mathbf{A}\in\mathbb{R}^{m\times n}$, $||\mathbf{A}||_* \leq \sqrt{\min(m,n)}||\mathbf{A}||_\mathrm{F}$: We know $\forall\mathbf{x}\in\mathbb{R}^n, ||\mathbf{x}||_1 \leq \sqrt{n}||\mathbf{x}||_2$. By definition, the Schatten-p norm of $\mathbf{A}$ is the ℓ_p norm of its singular values. Recalling the nuclear norm and the Frobenius norm are the Schatten-1 and Schatten-2 norms, $\mathbf{A}$ has $\min(m,n)$ singular values, hence the result. □

In practice, we find that on synthetic data designed to test the model, we effectively recover the ranks of $\mathbf{A}$ and $\mathbf{B}$ regardless of the choice of r, as seen in Figure 3.

Algorithm 1 Kronecker-Decomposable Robust Sparse Dictionary Learning.

procedure KDRSDL($\boldsymbol{\mathcal{X}}, r, \lambda, \alpha$)
  $\mathbf{A}^0, \mathbf{B}^0, \boldsymbol{\mathcal{E}}^0, \boldsymbol{\mathcal{R}}^0, \boldsymbol{\mathcal{K}}^0 \leftarrow$ INITIALIZE($\boldsymbol{\mathcal{X}}$)
  while not converged **do**
    $\boldsymbol{\mathcal{E}}^{t+1} \leftarrow \mathcal{S}_{\lambda/\mu^t}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{K}}^t \times_1 \mathbf{A}^t \times_2 \mathbf{B}^t + \frac{1}{\mu^t}\boldsymbol{\Lambda}^t)$
    $\tilde{\boldsymbol{\mathcal{X}}}^{t+1} \leftarrow \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{E}}^{t+1}$
    $\mathbf{A}^{t+1} \leftarrow (\sum_i (\mu^t \tilde{\mathbf{X}}_i^{t+1} + \boldsymbol{\Lambda}_i^t)\mathbf{B}^t(\mathbf{K}_i^t)^\mathsf{T}) \,/\, (\mathbf{I} + \mu^t \sum_i \mathbf{K}_i^t(\mathbf{B}^t)^\mathsf{T}\mathbf{B}^t(\mathbf{K}_i^t)^\mathsf{T})$
    $\mathbf{B}^{t+1} \leftarrow (\sum_i (\mu^t \tilde{\mathbf{X}}_i^{t+1} + \boldsymbol{\Lambda}_i^t)^\mathsf{T}\mathbf{A}^{t+1}\mathbf{K}_i^t) \,/\, (\mathbf{I} + \mu^t \sum_i (\mathbf{K}_i^t)^\mathsf{T}(\mathbf{A}^{t+1})^\mathsf{T}\mathbf{A}^{t+1}\mathbf{K}_i^t)$
    for all i **do**
      $\mathbf{K}_i^{t+1} \leftarrow$ STEIN($-\frac{\mu^t}{\mu_{\boldsymbol{\mathcal{K}}}^t}(\mathbf{A}^{t+1})^\mathsf{T}\mathbf{A}^{t+1},\ (\mathbf{B}^{t+1})^\mathsf{T}\mathbf{B}^{t+1},\ \frac{1}{\mu_{\boldsymbol{\mathcal{K}}}^t}\left[(\mathbf{A}^{t+1})^\mathsf{T}(\boldsymbol{\Lambda}_i^t + \mu^t\tilde{\mathbf{X}}_i^{t+1})\mathbf{B}^{t+1} + \mathbf{Y}_i^t\right] + \mathbf{R}_i^t$)
      $\mathbf{R}_i^{t+1} \leftarrow \mathcal{S}_{\alpha/\mu_{\boldsymbol{\mathcal{K}}}^t}(\mathbf{K}_i^{t+1} - \frac{1}{\mu_{\boldsymbol{\mathcal{K}}}^t}\mathbf{Y}_i^t)$
    end for
    $\boldsymbol{\Lambda}^{t+1} \leftarrow \boldsymbol{\Lambda}^t + \mu^t(\tilde{\boldsymbol{\mathcal{X}}}^{t+1} - \boldsymbol{\mathcal{K}}^{t+1} \times_1 \mathbf{A}^{t+1} \times_2 \mathbf{B}^{t+1})$
    $\boldsymbol{\mathcal{Y}}^{t+1} \leftarrow \boldsymbol{\mathcal{Y}}^t + \mu_{\boldsymbol{\mathcal{K}}}^t(\boldsymbol{\mathcal{R}}^{t+1} - \boldsymbol{\mathcal{K}}^{t+1})$
    $\mu^{t+1} \leftarrow \min(\mu^*, \rho\mu^t)$
    $\mu_{\boldsymbol{\mathcal{K}}}^{t+1} \leftarrow \min(\mu_{\boldsymbol{\mathcal{K}}}^*, \rho\mu_{\boldsymbol{\mathcal{K}}}^t)$
  end while
  return $\mathbf{A}, \mathbf{B}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{E}}$
end procedure

3.3. Convergence and complexity

Convergence and initialization: The problem is non-convex, therefore global convergence is a priori not guaranteed. Recent work [24, 43] studies the convergence of ADMM for non-convex and possibly non-smooth objective functions with linear constraints. Here, the constraints are not linear. In [21, 22] the authors derive conditions for global optimality in specific non-convex matrix and tensor factorizations that may be extended to other formulations, including our model. However, the theoretical study of the global convergence of our algorithm is out of the scope of this paper and is left for future work. Instead, we provide experimental results and discuss the strategy implemented.

We propose a simple initialization scheme in the wake of [29]. We initialize the bases $\mathbf{A}$ and $\mathbf{B}$ and the core $\boldsymbol{\mathcal{R}}$ by performing SVD on each observation $\mathbf{X}_i = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i^\mathsf{T}$. We set $\mathbf{R}_i = \mathbf{S}_i$, $\mathbf{A} = \frac{1}{N}\sum_i \mathbf{U}_i$ and $\mathbf{B} = \frac{1}{N}\sum_i \mathbf{V}_i$. To initialize the dual-variables for the constraint $\mathbf{X}_i - \mathbf{A}\mathbf{R}_i\mathbf{B}^\mathsf{T} - \mathbf{E}_i = \mathbf{0}$, we take $\mu^0 = \frac{\eta N}{\sum_i ||\mathbf{X}_i||_\mathrm{F}}$ where η is a scaling coefficient, chosen in practice to be $\eta = 1.25$ as in [29]. Similarly, we chose $\mu_{\boldsymbol{\mathcal{K}}}^0 = \frac{\eta N}{\sum_i ||\mathbf{R}_i||_\mathrm{F}}$. These correspond to averaging the initial values for each individual slice and its corresponding constraint. Our convergence criterion corresponds to primal-feasibility of problem (11), and is given by $\max(\mathrm{err}_{rec}, \mathrm{err}_{split}) \leq \epsilon$ where $\mathrm{err}_{rec} = \max_i \frac{||\mathbf{X}_i - \mathbf{A}\mathbf{R}_i\mathbf{B}^\mathsf{T} - \mathbf{E}_i||_\mathrm{F}^2}{||\mathbf{X}_i||_\mathrm{F}^2}$ and $\mathrm{err}_{split} = \max_i \frac{||\mathbf{R}_i - \mathbf{K}_i||_\mathrm{F}^2}{||\mathbf{R}_i||_\mathrm{F}^2}$. Empirically, we obtained systematic convergence to a good solution, and a linear convergence rate, as shown in Figure 4. The parameter $\rho > 1$ can be tuned and affects the speed of convergence. High values may lead to the algorithm diverging, while values closer to 1 will lead to an increased number of iterations. We used $\rho = 1.2$ and set upper bounds $\mu^* = \mu^0 \times 10^7$ and $\mu_{\boldsymbol{\mathcal{K}}}^* = \mu_{\boldsymbol{\mathcal{K}}}^0 \times 10^7$ as usual with ADMM.

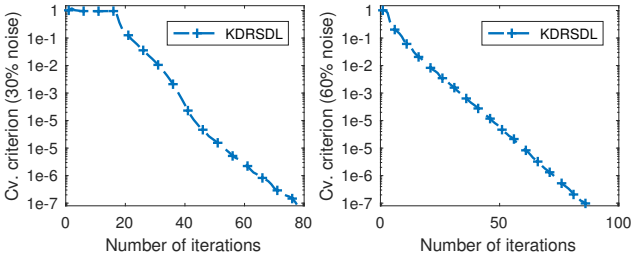


Figure 4: Convergence on synthetic data with 30% and 60% corruption.

Computational complexity: The time and space complexity per iteration of Algorithm 1 are $O(N(mnr + (m+n)r + mn + \min(m,n)r^2 + r^3 + r^2))$ and $O(N(mn + r^2) + (m+n)r + r^2)$. Since $r \leq \min(m,n)$, the terms in r are asymptotically negligible, but in practice it is useful to know how the computational requirements scale with the size of the dictionary. Similarly, the initialization procedure has cost $O(N(mn\min(m,n) + (\min(m,n))^3 + mn) + mn)$ in time and needs quadratic space per slice, assuming a standard algorithm is used for the SVD[9]. Several key steps in the algorithm, such as summations of independent terms, are trivially distributed in a *MapReduce* [12] way. Proximal operators are separable in nature and are therefore parallelizable. Consequently, highly parallel and distributed implementations are possible, and computational complexity can be further reduced by adaptively adopting sparse linear algebra structures and algorithms.

4. Experimental evaluation

We compared the performance of our model against a range of state-of-the-art tensor decomposition algorithms on four low-rank modeling computer vision benchmarks:

two for image denoising, and two for background subtraction. As a baseline, we report the performance of matrix Robust PCA implemented via inexact ALM (*RPCA*) [6, 29], and of Non-Negative Robust Dictionary Learning (*RNNDL*) [35]. We chose the following methods to include recent representatives of various existing approaches to low-rank modeling on tensors: The singleton version of Higher-Order Robust PCA (*HORPCA-S*) [18] optimizes the Tucker rank of the tensor through the sum of the nuclear norms of its unfoldings. In [46], the authors consider a similar model but with robust M-estimators as loss functions, either a Cauchy loss or a Welsh loss, and support both hard and soft thresholding; we tested the soft-thresholding models (*Cauchy ST* and *Welsh ST*). Non-convex Tensor Robust PCA (*NC TRPCA*) [2] adapts to tensors the matrix non-convex RPCA [33]. Finally, the two Tensor RPCA algorithms [31, 48] (*TRPCA '14* and *TRPCA '16*) work with slightly different definitions of the tensor nuclear norm as a convex surrogate of the tensor multi-rank.

For each model, we identified a maximum of two parameters to tune via grid-search in order to keep parameter tuning tractable. When criteria or heuristics for choosing the parameters were provided by the authors, we chose the search space around the value obtained from them. In all cases, the tuning process explored a wide range of parameters to maximize performance. The range of values investigated are provided as supplementary material.

When the performance of one method was significantly worse than that of the other, the result is not reported so as not to clutter the text, and is made available in the supplementary material. This is the case of Separable Dictionary Learning [23] whose drastically different nature renders unsuitable for robust low-rank modeling, but was compared for completeness. For the same reason, we did not compare our method against K-SVD [1], or [25].

4.1. Background subtraction

Background subtraction is a common task in computer vision and can be tackled by robust low-rank modeling: the static or mostly static background of a video sequence can effectively be represented as a low-rank tensor while the foreground forms a sparse component of outliers.

Experimental procedure We compared the algorithms on two benchmarks. The first is an excerpt of the *Highway* dataset [20], and consists in a video sequence of cars travelling on a highway; the background is completely static. We kept 400 gray-scale images re-sized to 48×64 pixels. The second is the *Airport Hall* dataset ([27]) and has been chosen as a more challenging benchmark since the background is not fully static and the scene is richer. We used the same excerpt of 300 frames (frames 3301 to 3600) as in [49], and kept the frames in their original size of 144×176 pixels.

We treat background subtraction as a binary classification problem. Since ground truth frames are available for our excerpts, we report the AUC [15] on both videos. The value of α was set to $1\mathrm{e}{-2}$ for both experiments.

(a) Cauchy ST (b) Welsh ST (c) TRPCA '16 (d) HORPCA-S (e) NCTRPCA

(f) RNNDL (g) RPCA (h) KDRSDL (i) Ground truth (j) Original

Figure 5: Results on *Airport Hall*. TRPCA '14 removed.

Results We provide the original, ground truth, and recovered frames in Figure 5 for the *Hall* experiment (*Highway* in supplementary material).

Table 1 presents the AUC scores of the algorithms, ranked in order of their mean performance on the two benchmarks. The two matrix methods rank high on both benchmarks and only half of the tensor algorithms match or outperform this baseline. Our proposed model matches the best performance on the *Highway* dataset and provides significantly higher performance than the other on the more challenging *Hall* benchmark. Visual inspection of the results show KDRSDL is the only method that doesn't fully capture the immobile people in the background, and therefore achieves the best trade-off between foreground detection and background-foreground contamination.

Algorithm	Highway	Hall
KDRSDL (proposed)	0.94	0.88
TRPCA '16	0.94	0.86
NC TRPCA	0.93	0.86
RPCA (baseline)	0.94	0.85
RNNDL (baseline)	0.94	0.85
HORPCA-S	0.93	0.86
Cauchy ST	0.83	0.76
Welsh ST	0.82	0.71
TRPCA '14	0.76	0.61

Table 1: AUC on *Highway* and *Hall* ordered by mean AUC.

4.2. Image denoising

Many natural and artificial images exhibit an inherent low-rank structure and are suitably denoised by low-rank modeling algorithms. In this section, we assess the performance of the cohort on two datasets chosen for their popularity, and for the typical use cases they represent.

We consider collections of grayscale images, and color images represented as 3-way tensors. Laplacian (salt & pepper) noise was introduced separately in all frontal slices of

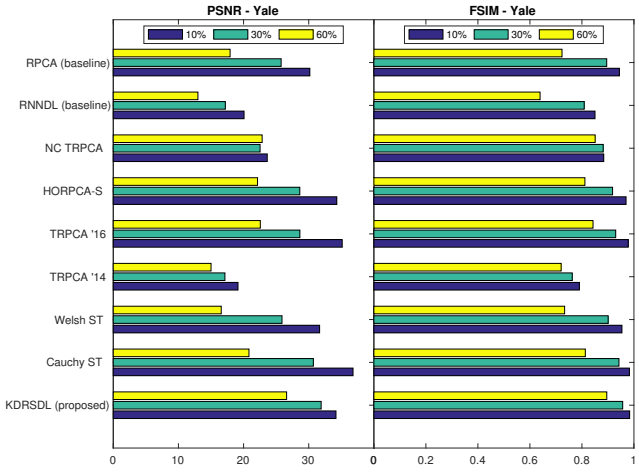
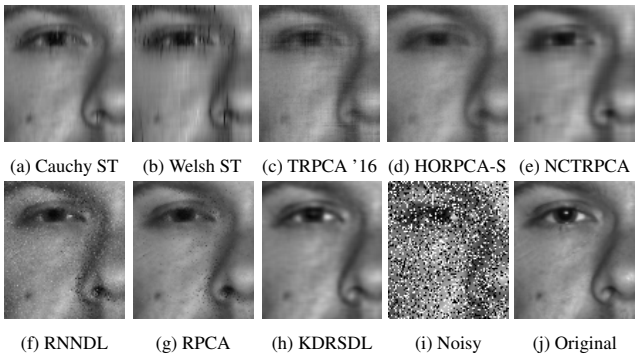


Figure 6: Mean PSNR and FSIM on the 64 images of the first subject of *Yale* at noise levels 10%, 30%, and 60%.

the observation tensor at three different levels: 10%, 30%, and 60% to simulate medium, high, and gross corruption. In these experiments we set the value of α to $1\mathrm{e}{-3}$ for noise levels up to 30%, and to $1\mathrm{e}{-2}$ at the 60% level.

We report two quantitative metrics designed to measure two key aspects of image recovery. The *Peak Signal To Noise Ratio (PSNR)* will be used as an indicator of the element-wise reconstruction quality of the signals, while the *Feature Similarity Index (FSIM, FSIMc for color images)* [30] evaluates the recovery of structural information. Quantitative metrics are not perfect replacements for subjective visual assessment of image quality; therefore, we present sample reconstructed images for verification. Our measure of choice for determining which images to compare visually is the FSIM(c) for its higher correlation with human evaluation than the PSNR.

Monochromatic face images Our face denoising experiment uses the Extented Yale-B dataset [17] of 10 different subject, each under 64 different lighting conditions. According to [38, 3], face images of one subject under various illuminations lie approximately on a 9-dimensional subspace, and are therefore suitable for low-rank modeling. We used the pre-cropped 64 images of the first subject and kept them at full resolution. The resulting collection of images constitutes a 3-way tensor of 64 images of size 192×168. Each mode corresponds respectively to the columns and rows of the image, and to the illumination component. All three are expected to be low-rank due to the spatial correlation within frontal slices and to the correlation between images of the same subject under different illuminations. We present the comparative quantitative performance of the methods tested in Figure 6, and provide visualizations of the reconstructed first image at the 30% noise level in Figure 7. We report the metrics averaged on the 64 images.



(a) Cauchy ST (b) Welsh ST (c) TRPCA '16 (d) HORPCA-S (e) NCTRPCA

(f) RNNDL (g) RPCA (h) KDRSDL (i) Noisy (j) Original

Figure 7: Results on the Yale benchmark with 30% noise. TRPCA '14 removed.

At the 10% noise level, nearly every method provided good to excellent recovery of the original images. We therefore omit this noise level (*c.f.* supplementary material). On the other hand, most methods, with the notable exception of KDRSDL, NC TRPCA, and TRPCA '16, failed to provide acceptable reconstruction in the gross corruption case. Thus, we present the denoised images at the 30% level, and compare the performance of the three best performing methods in Table 2 for the 60% noise level.

Clear differences appeared at the 30% noise level, as demonstrated both by the quantitative metrics, and by visual inspection of Figure 7. Overall, performance was markedly lower than at the 10% level, and most methods started to lose much of the details. Visual inspection of the results confirms a higher reconstruction quality for KDRSDL. We invite the reader to look at the texture of the skin, the white of the eye, and at the reflection of the light on the subject's skin and pupil. The latter, in particular, is very close in nature to the white pixel corruption of the salt & pepper noise. Out of all methods, KDRSDL provided the best reconstruction quality: it is the only algorithm that removed all the noise and for which all the aforementioned details are distinguishable in the reconstruction.

Noisy	KDRSDL	NC TRPCA	TRPCA '16
PSNR	26.6057	22.8502	22.566
FSIM	0.8956	0.8509	0.8427

Table 2: Three best results on *Yale* at 60% noise.

At the 60% noise level, our method scored markedly higher than its competitors on image quality metrics, as seen both in Figure 6 and in Table 2. Visualizing the re-

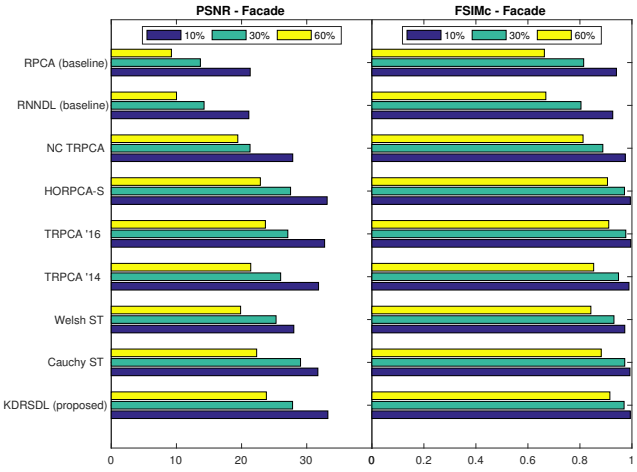


Figure 8: PSNR and FSIMc of all methods on the Facade benchmark at noise levels 10%, 30%, and 60%.

constructions confirms the difference: the image recovered by KDRSDL at the 60% noise level is comparable to the output of competing algorithms at the 30% noise level.

Color image denoising Our benchmark is the *Facade* image [10]: the rich details and lighting makes it interesting to assess fine reconstruction. The geometric nature of the building's front wall, and the strong correlation between the RGB bands indicate the data can be modeled by a low-rank 3-way tensor where each frontal slice is a color channel.

At the 10% noise level, KDRSDL attained the highest PSNR, and among the highest FSIMc values. Most methods provided excellent reconstruction, in agreement with the high values of the metrics shown in Figure 8. As in the previous benchmark, the results are omitted because of the space constraints (*c.f.* supplementary material). At the 30% noise level, Cauchy ST exhibited the highest PSNR, while TRPCA '16 scored best on the FSIMc metric. KDRSDL had the second highest PSNR and among the highest FSIMc. Details of the results are provided in Figure 9. Visually, clear differences are visible, and are best seen on the fine details of the picture, such as the black iron ornaments, or the light coming through the window. Our method best preserved the dynamics of the lighting, and the sharpness of the details, and in the end provided the reconstruction visually closest to original. Competing models tend to oversmooth the image, and to make the light dimmer; indicating substantial losses of high-frequency and dynamic information. KDRSDL appears to also provide the best color fidelity.

In the gross-corruption case, KDRSDL was the best performer on both PSNR and FSIMc. As seen on Figure 3, KDRSDL was the only method with TRPCA '16 and HORPCA-S to provide a reconstruction with distinguishable details, and did it best.

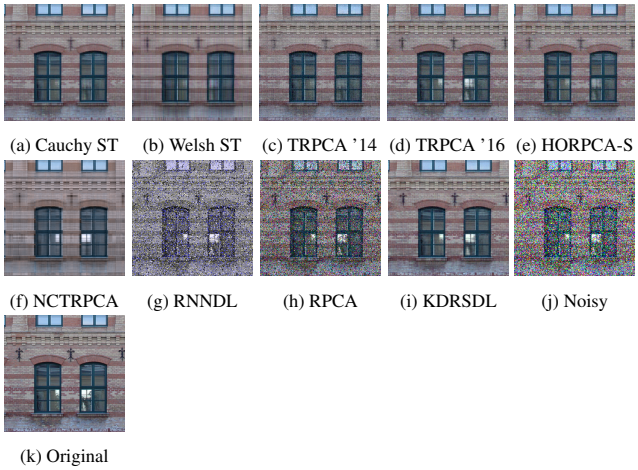


Figure 9: Results on the Facade benchmark with 30% noise.

Noisy	KDRSDL	TRPCA '16	HORPCA-S
PSNR	23.8064	23.6552	22.8811
FSIMc	0.9152	0.9109	0.9060

Table 3: Three best results on *Facade* at 60% noise.

5. Conclusion

The method we propose combines aspects from Robust Principal Component Analysis, and Sparse Dictionary Learning. As in K-SVD, our algorithm learns iteratively and alternatively both the dictionary and the sparse representations. Similarly to Robust PCA, our method assumes the data decompose additively in a sparse and low-rank component, and is able to separate the two signals. We introduce a two-level structure in the dictionary to allow for both scalable training, and robustness to outliers. Imposing this structure exhibits links with tensor factorizations and allows us to better model spatial correlation in images than classical matrix methods. These theoretical advantages translate directly in the experimental performance, as our method exhibits desirable scalability properties, and matches or outperforms the current state of the art in low-rank modeling.

6. Acknowledgements

The work of Y. Panagakis has been partially supported by the European Community Horizon 2020 [H2020/2014-2020] under Grant Agreement No. 645094 (SEWA). S. Zafeiriou was partially funded by EPSRC Project EP/N007743/1 (FACER2VM).

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 1, 6

[2] A. Anandkumar, P. Jain, Y. Shi, and U. N. Niranjan. Tensor vs Matrix Methods: Robust Tensor Decomposition under Block Sparse Perturbations. In A. Gretton and R. C. Christian, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, Cadiz, Spain, 2016. JMLR.org. 6

[3] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2 2003. 7

[4] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 2

[5] S. Boyd. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2010. 2

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 5 2011. 1, 4, 6

[7] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2 2006. 1

[8] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 12 2005. 1

[9] T. F. Chan. An Improved Algorithm for Computing the Singular Value Decomposition. *ACM Transactions on Mathematical Software*, 8(1):72–83, 1982. 5

[10] X. Chen, Z. Han, Y. Wang, Q. Zhao, D. Meng, and Y. Tang. Robust Tensor Factorization with Unknown Noise. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5213–5221, 2016. 8

[11] G. M. Davis, S. G. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *Optical Engineering*, 33(7):2183–2191, 1994. 2

[12] J. Dean and S. Ghemawat. MapReduce: Simplied Data Processing on Large Clusters. *Proceedings of 6th Symposium on Operating Systems Design and Implementation*, pages 137–149, 2004. 5

[13] D. L. Donoho. For most large underdetermined systems of equations, the minimal L1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934, 2006. 1

[14] M. Elad and M. Aharon. Image denoising via sparse and redundant representation over learned dictionaries. *IEEE Transations on Image Processing*, 15(12):3736–3745, 12 2006. 1

[15] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. 6

[16] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2973–2981, 2016. 2

[17] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 6 2001. 7

[18] D. Goldfarb and Z. T. Qin. Robust Low-Rank Tensor Recovery: Models and Algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 3 2014. 6

[19] G. Golub, S. Nash, and C. Van Loan. A Hessenberg-Schur method for the problem AX + XB= C. *IEEE Transactions on Automatic Control*, 24(6):909–913, 12 1979. 4

[20] N. Goyette, P. M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. changedetection.net: A new change detection benchmark dataset. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2012. 6

[21] B. Haeffele, E. Young, and R. Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In T. Jebara and E. P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 2007–2015. JMLR Workshop and Conference Proceedings, 2014. 5

[22] B. D. Haeffele and R. Vidal. Global Optimality in Tensor Factorization, Deep Learning, and Beyond. *CoRR*, abs/1506.0, 2015. 5

[23] S. Hawe, M. Seibert, and M. Kleinsteuber. Separable Dictionary Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3 2013. 2, 6

[24] M. Hong, Z. Q. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3836–3840, 2015. 5

[25] S. H. Hsieh, C. S. Lu, and S. C. Pei. 2D sparse dictionary learning via tensor decomposition. *2014 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2014*, pages 492–496, 2014. 2, 6

[26] T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 8 2009. 2

[27] L. Li, W. Huang, I.-H. Gu, and Q. Tian. Statistical Modeling of Complex Backgrounds for Foreground Object Detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 11 2004. 6

[28] X. Li, Z. Wang, J. Lu, R. Arora, J. D. Haupt, H. Liu, and T. Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *CoRR*, abs/1612.09296, 2016. 2

[29] Z. Lin, M. Chen, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *arXiv:1009.5055*, page 23, 9 2010. 5, 6

[30] Lin Zhang, Lei Zhang, Xuanqin Mou, and D. Zhang. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 8 2011. 7

[31] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Tensors via Convex Optimization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 6

[32] J. Mairal, M. Elad, and G. Sapiro. Sparse learned representations for image restoration. *Sparse learned representations for image restoration*, (December):1–10, 2008. 1

[33] P. Netrapalli, U. N. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex Robust PCA. *Advances in Neural Information Processing Systems 27 (Proceedings of NIPS)*, pages 1–9, 10 2014. 6

[34] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997. 1

[35] Q. Pan, D. Kong, C. Ding, and B. Luo. Robust Non-negative Dictionary Learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 2027–2033. AAAI Press, 2014. 6

[36] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding Low-Rank Solutions via Non-Convex Matrix Factorization, Efficiently and Provably. *ArXiv e-prints*, 2016. 2

[37] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44, 11 1993. 2

[38] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, volume 64, pages 497–500, New York, New York, USA, 2001. ACM Press. 7

[39] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *Optimization Online*, 52(3):1–33, 6 2007. 3

[40] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 6 2010. 1, 2

[41] E. Schwab, R. Vidal, and N. Charon. Efficient Global Spatial-Angular Sparse Coding for Diffusion MRI with Separable Dictionaries. *ArXiv e-prints*, Dec. 2016. 2

[42] E. Schwab, R. Vidal, and N. Charon. Spatial-Angular Sparse Coding for HARDI. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III*, pages 475–483. Springer International Publishing, Cham, 2016. 2

[43] Y. Wang, W. Yin, and J. Zeng. Global Convergence of ADMM in Nonconvex Nonsmooth Optimization. *ArXiv e-prints*, 2015. 5

[44] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian Optimization for Low Rank Matrix Completion. *ArXiv e-prints*, 2016. 2

[45] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse Representation for Computer Vision and Pattern Recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 6 2010. 1

[46] Y. Yang, Y. Feng, and J. A. K. Suykens. Robust Low-Rank Tensor Recovery With Regularized Redescending M-Estimator. *IEEE Transactions on Neural Networks and Learning Systems*, 27(9):1933–1946, 9 2015. 6

[47] X. Zhang, L. Wang, and Q. Gu. A Nonconvex Free Lunch for Low-Rank plus Sparse Matrix Recovery. *ArXiv e-prints*, Feb. 2017. 2

[48] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-SVD. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3842–3849, 2014. 2, 6

[49] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, and S.-I. Amari. Bayesian Robust Tensor Factorization for Incomplete Multiway Data. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):736–748, 4 2016. 6

[50] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global Optimality in Low-rank Matrix Optimization. *ArXiv e-prints*, Feb. 2017. 2