

LOCAL PRINCIPAL COMPONENT PURSUIT FOR NONLINEAR DATASETS

Brendt Wohlberg, Rick Chartrand, and James Theiler

Los Alamos National Laboratory
Los Alamos, NM 87545, USA

ABSTRACT

A robust version of Principal Component Analysis (PCA) can be constructed via a decomposition of a data matrix into low rank and sparse components, the former representing a low-dimensional linear model of the data, and the latter representing sparse deviations from the low-dimensional subspace. This decomposition has been shown to be highly effective, but the underlying model is not appropriate when the data are not modeled well by a single low-dimensional subspace. We construct a new decomposition corresponding to a more general underlying model consisting of a union of low-dimensional subspaces, and demonstrate the performance on a video background removal problem.

Index Terms— Robust Principal Component Analysis, Low Rank, Group Sparse, Compressive Sensing

1. INTRODUCTION

Matrix completion, which attempts to reconstruct a matrix with only a small fraction of its entries known [1], is a recent branch of the field of compressive sensing. (The assumption that the matrix has a low rank plays a role analogous to that of sparsity in compressive sensing.) An extension of this problem seeks to decompose a matrix D of high-dimensional data into a sum of two components, one having low rank, the other being sparse. This can be expressed as the optimization

$$\min_{L,S} \text{rank}(L) + \lambda \|S\|_0, \text{ subject to } L + S = D, \quad (1)$$

where $\|\cdot\|_0$ counts the number of nonzero entries, and $\lambda > 0$ is a tuning parameter. We can regard L as a low-dimensional description of the data, while S consists of discrepancies from that model, which can be interesting in their own right. Applications considered thus far include automated background removal in video [2], text analysis [3], and image alignment [4].

We can compare (1) to Principal Component Analysis (PCA), which would compute the matrix L of desired rank that minimizes $\|D - L\|_2$, the entry-wise Euclidean norm of the residual. Because the second term of (1) penalizes

only the number of discrepancies and not their size, the low-dimensional model L will not be perturbed by outliers among the entries of D , and hence will provide a more robust description of most of the dataset. This connection between sparse optimization and “robust PCA” was made by Candès *et al.* [5], who also provided a tractable, convex approximation, which they called Principal Component Pursuit, of the NP-hard problem (1)

$$\min_{L,S} \|\sigma(L)\|_1 + N^{-1/2} \|S\|_1, \text{ subject to } L + S = D, \quad (2)$$

where D is $m \times n$ and $N = \max\{m, n\}$. The first term is the ℓ^1 norm of the vector $\sigma(L)$ of singular values of L , and is known as the *nuclear norm* of L .

This decomposition approach assumes that there is a single, low-dimensional model that describes most components of the elements of the dataset. In this work, we develop a more general method that is suitable, for instance, for data described by a *manifold* [6], except for a sparse set of possibly-large discrepancies. We will thus allow our low-dimensional description to vary across the dataset, while retaining the robustness given by having a second, sparse component. In the context of video background removal, this will allow us to handle the case of a moving camera, making the method suitable for a much larger class of surveillance problems.

2. LOCAL PRINCIPAL COMPONENT PURSUIT

The most obvious extension to the nonlinear case is to a union of low-dimensional subspaces, which is able to approximate data lying within a nonlinear manifold as a collection of tangent spaces. The geometric intuition motivating our approach is that if the data lie within a nonlinear manifold, then every sample in the manifold may be represented (assuming adequate sampling density) as a sparse linear combination of neighboring samples spanning an approximation to the local tangent plane. This idea can be implemented as the problem

$$\min_{U,S} \|U\|_1 + \alpha \|U\|_{2,1} + \beta \|S\|_1 \text{ such that } DU + S = D,$$

in which the explicit notion of low rank, and its nuclear-norm proxy, is replaced by representability of a matrix as a sparse representation on itself. (The subspace segmentation algorithm of Liu *et al.* [7], in contrast, combines the concept of

This research was supported by the U.S. Department of Energy through the LANL/LDRD Program.

self-representability with a continuation of the explicit low-rank formulation.) A simple 1-norm notion of sparsity is insufficient here, since it is essential to exclude the identity as a coefficient matrix, which we avoid by employing the 2, 1-norm, defined as $\|U\|_{2,1} = \sum_i \sqrt{\sum_j u_{ij}^2}$. This norm encourages most rows of U to be zero, but does not discourage nonzero values among the entries of a nonzero row [8]. Using this norm also helps take advantage of *group-sparsity* structure, such as can arise when points of the dataset are near to each other.

To better handle noisy data, we replace the constrained form with a penalized form, and add a total variation penalty on the sparse deviations (for cases when we expect these deviations to form contiguous regions), giving the problem

$$\min_{U,S} \frac{1}{2} \|AU + S - D\|_2^2 + \alpha \|U\|_1 + \beta \|U\|_{2,1} + \gamma \|S\|_1 + \delta \|\nabla S\|_1, \quad (3)$$

where the dictionary A is derived from the data D (by mean-subtraction and scaling), and ∇S is a vector-valued discretization of the 3-D gradient of S , interpreted as an image cube.

Eq. (3) can be solved efficiently using the Split Bregman method [9]. We introduce variables P , Q , and R , which are auxiliary versions of U , S , and ∇S , respectively. We add terms relaxing the equality constraints of each quantity and its auxiliary variable, and in order to enforce equality at convergence, we introduce Bregman variables B_p , B_q , and B_r [9]:

$$\begin{aligned} \min_{U,S,P,Q,R} \frac{1}{2} \|AU + S - D\|_2^2 + \alpha \|P\|_1 + \beta \|P\|_{2,1} \\ + \gamma \|Q\|_1 + \delta \|R\|_1 + \frac{\lambda}{2} \|P - U - B_p\|_2^2 \\ + \frac{\mu}{2} \|Q - S - B_q\|_2^2 + \frac{\nu}{2} \|R - \nabla S - B_r\|_2^2. \end{aligned} \quad (4)$$

This allows the problem to be split into an alternating minimization of the following subproblems:

$$\min_U \frac{1}{2} \|AU - (D - S)\|_2^2 + \frac{\lambda}{2} \|U - (P - B_p)\|_2^2, \quad (5)$$

$$\begin{aligned} \min_S \frac{1}{2} \|S - (D - AU)\|_2^2 + \frac{\mu}{2} \|S - (Q - B_q)\|_2^2 \\ + \frac{\nu}{2} \|\nabla S - (R - B_r)\|_2^2, \end{aligned} \quad (6)$$

$$\min_P \frac{\lambda}{2} \|P - (U + B_p)\|_2^2 + \alpha \|P\|_1 + \beta \|P\|_{2,1}, \quad (7)$$

$$\min_Q \frac{\mu}{2} \|Q - (S + B_q)\|_2^2 + \gamma \|Q\|_1, \text{ and} \quad (8)$$

$$\min_R \frac{\nu}{2} \|R - (\nabla S + B_r)\|_2^2 + \delta \|R\|_1. \quad (9)$$

Subproblems (5) and (6) are simple ℓ^2 problems, and can be solved by standard techniques for solving linear systems

(e.g., conjugate gradient). The other three subproblems can be solved very cheaply using *shrinkage*. Subproblems (8) and (9) use standard shrinkage, also known as soft thresholding:

$$\text{shrink}(T, \zeta) = \text{sign}(T) \max\{0, |T| - \zeta\}, \quad (10)$$

where the operations are to be understood entrywise. Subproblem (7), which contains both the ℓ^1 and $\ell^{2,1}$ norms, uses a generalized shrinkage, defined row-wise by

$$\text{shrink}_{2,1}(T, \zeta, \eta)^i = \frac{\text{shrink}(T^i, \zeta)}{1 + \eta / \text{shrink}(\|\text{shrink}(T^i, \zeta)\|_2, \eta)}, \quad (11)$$

with the convention that $1/(1 + \eta/0) = 0$. The algorithm consists of iteratively solving the main variables and updating the Bregman variables as follows:

$$\begin{aligned} U^{(k+1)} &= (A^T A + \lambda I)^{-1} (A^T (D - S^{(k)}) + \lambda (P^{(k)} - B_p^{(k)})), \\ S^{(k+1)} &= ((1 + \mu)I + \nu \nabla^T \nabla)^{-1} ((D - AU^{(k+1)}) \\ &\quad + \mu (Q^{(k)} - B_q^{(k)}) + \nu \nabla^T (R^{(k)} - B_r^{(k)})), \\ P^{(k+1)} &= \text{shrink}_{2,1}(U^{(k+1)} + B_p^{(k)}, \alpha/\lambda, \beta/\lambda), \\ Q^{(k+1)} &= \text{shrink}(S^{(k+1)} + B_q^{(k)}, \gamma/\mu), \\ R^{(k+1)} &= \text{shrink}(\nabla S^{(k+1)} + B_r^{(k)}, \delta/\nu), \\ B_p^{(k+1)} &= B_p^{(k)} + U^{(k+1)} - P^{(k+1)}, \\ B_q^{(k+1)} &= B_q^{(k)} + S^{(k+1)} - Q^{(k+1)}, \text{ and} \\ B_r^{(k+1)} &= B_r^{(k)} + \nabla S^{(k+1)} - R^{(k+1)}. \end{aligned}$$

3. ADAPTIVE, OUTLIER-REMOVED DICTIONARY

We can regard (3) as approximating $D - S$ with the local sparse representation AU . Since the dictionary A is simply a shifted and scaled version of D , we can expect using $A - S$ as an adaptive dictionary to allow U to be even sparser. This gives us the modified problem

$$\begin{aligned} \min_{U,S} \frac{1}{2} \|(A - S)U + S - D\|_2^2 + \alpha \|U\|_1 \\ + \beta \|U\|_{2,1} + \gamma \|S\|_1 + \delta \|\nabla S\|_1. \end{aligned} \quad (12)$$

This problem can be minimized as before, the only changes being to the subproblems for U and S :

$$\begin{aligned} \min_U \frac{1}{2} \|(A - S)U - (D - S)\|_2^2 + \frac{\lambda}{2} \|U - (P - B_p)\|_2^2, \\ \min_S \frac{1}{2} \|S(I - U) - (D - AU)\|_2^2 + \frac{\mu}{2} \|S - (Q - B_q)\|_2^2 \\ + \frac{\nu}{2} \|\nabla S - (R - B_r)\|_2^2, \end{aligned}$$

with solutions given by the linear systems

$$\begin{aligned} ((A - S)^T(A - S) + \lambda I)U \\ &= (A - S)^T(D - S) + \lambda(P - B_p), \\ S(I - U)(I - U)^T + (\mu I + \nu \nabla^T \nabla)S \\ &= (D - AU)(I - U)^T + \mu(Q - B_q) + \nu \nabla^T(R - B_r). \end{aligned}$$

4. RESULTS

We test our algorithm on the video background removal problem addressed by Wright *et al.* [2], using a 288-frame traffic video sequence from the Lankershim Boulevard Dataset [10, camera 4, 8:45–9:00 AM]. (This problem provides a convenient comparison between these two general data decomposition techniques, but while the performance of our method is subjectively quite good, we do not claim that it is competitive when compared with application-specific algorithms for this problem.) We use the modified-dictionary form (12) of the algorithm since it gives better results. (Comparison omitted due to space constraints.) A and D were both constructed from the data by subtracting the mean from each column and scaling so that the maximum value was 1.

The first test sequence is a reduced-resolution (240×320 pixel frames) version of the data, with each frame of the video being a column of D , giving us a 76800×288 matrix. Because the traffic camera is stationary, this dataset is well-modeled by a single low-dimensional subspace. Our algorithm gives a decomposition (see Fig. 1) that is visually almost indistinguishable from the result (omitted here due to space constraints) obtained by solving (2), using the algorithm of [11].

Our second test sequence is constructed by taking a moving 240×320 pixel cropping window within the original sequence to simulate a panning camera. This window moves slowly to the left, and then back to the original position, at a rate of 1/4 pixel/frame. In this case the background is poorly approximated by any single low-dimensional subspace, but since the background motion is slow with respect to the foreground motion, a locally low-dimensional model provides a much better approximation. A comparison of a single frame of the sparse components of different methods applied to this data is provided in Fig. 2. The local sparse component clearly has far less residual background than the “global” sparse component (and in fact exhibits slightly less than the reference sparse component).

5. CONCLUSION

We have proposed a new decomposition, together with a Split Bregman type algorithm, for high-dimensional data, generalizing the Robust PCA of Candès *et al.* [5] to certain nonlinear data. The ability of this generalization to model data that does not conform to the globally low-dimensional restriction has

been demonstrated on the video background removal problem. Future work will include development of automatic parameter selection methods, and application of the decomposition to additional problems in which the relaxed constraints on the data can be expected to provide an advantage.

6. REFERENCES

- [1] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Found. Comput. Math.*, vol. 9, pp. 717–772, 2009.
- [2] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, “Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization,” in *Adv. in Neural Inf. Proc. Sys. (NIPS) 22*, 2009, pp. 2080–2088.
- [3] K. Min, Z. Zhang, J. Wright, and Y. Ma, “Decomposing background topics from keywords by principal component pursuit,” in *Proc. ACM Intl. Conf. Inf. Knowl. Manage. (CIKM)*, 2010, pp. 269–278.
- [4] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, “RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, June 2010, pp. 763–770.
- [5] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *J. ACM*, vol. 58, pp. 11:1–11:37, June 2011.
- [6] G. Peyré, “Manifold models for signals and images,” *Comp. Vis. Image Understand.*, vol. 113, no. 2, pp. 249–260, 2009.
- [7] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation.” in *Intl. Conf. Mach. Learn. (ICML)*, 2010, pp. 663–670.
- [8] E. van den Berg and M. P. Friedlander, “Theoretical and empirical results for recovery from multiple measurements,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2516–2527, May 2010.
- [9] T. Goldstein and S. J. Osher, “The split Bregman method for l1-regularized problems,” *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 323–343, 2009.
- [10] “Lankershim Boulevard dataset,” U.S. Department of Transportation Publication FHWA-HRT-07-029, Jan. 2007, data available from <http://ngsim-community.org/>.
- [11] Z. Lin, M. Chen, and Y. Ma, “The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices,” University of Illinois-Urbana Champaign, Tech. Rep. UILU-ENG-09-2214, 2010.



Fig. 1. Results for frame 166 from the stationary test video sequence. The decomposition was computed using our algorithm with parameters $\alpha = 1.0 \times 10^{-5}$, $\beta = 1.0 \times 10^{-2}$, $\gamma = 3.0 \times 10^{-5}$, and $\delta = 1.0 \times 10^{-4}$.

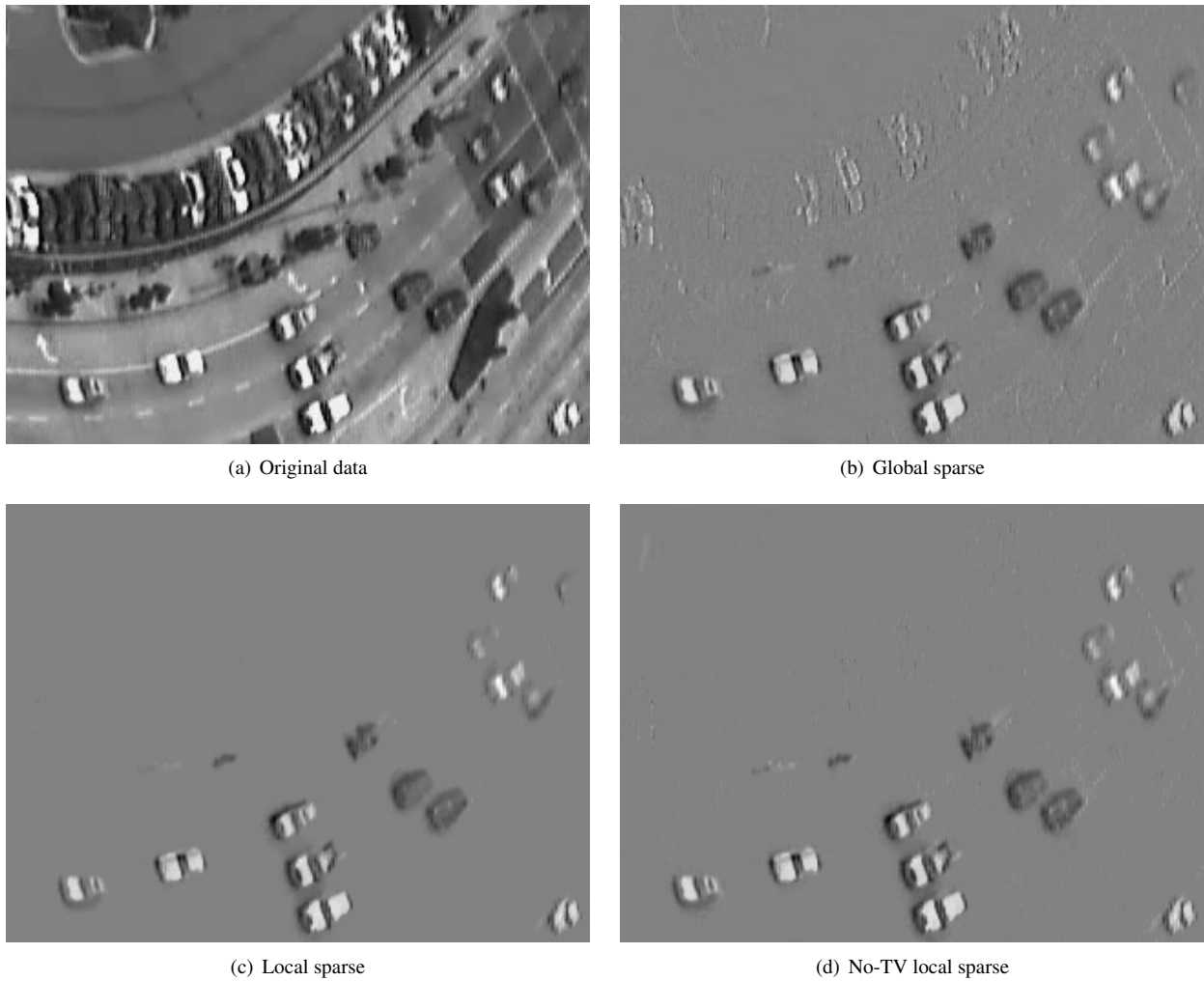


Fig. 2. Results for frame 166 from the slowly-panning test video sequence. The global sparse component (b) is obtained using decomposition (2), and the local sparse component (c) is generated by our algorithm with parameters $\alpha = 4.0 \times 10^{-3}$, $\beta = 8.0 \times 10^{-2}$, $\gamma = 5.0 \times 10^{-4}$, and $\delta = 3.0 \times 10^{-4}$. Component (d) is generated in the same way as (c), except that $\delta = 0$, so that there is no TV regularization. This result demonstrates the performance advantage of our algorithm is primarily due to the local-linear model, and not the TV regularization.