Sparse representation classification and positive $L1$ minimization

Cencheng Shen

Joint Work with Li Chen, Carey E. Priebe

Applied Mathematics and Statistics
Johns Hopkins University,

August 5, 2014

Overview

1. Introduction

2. Numerical experiments

3. Conclusion

Section 1

Introduction

Sparse representation classification?

- Our motivation comes from the sparse representation classification (SRC) proposed in *Wright et al. 2009* [1].
- It is a simple and intuitive classification procedure making use of $L1$ minimization, and argued to strike a balance between nearest-neighbor and nearest-subspace classifiers, while being more discriminative than both.
- Numerically shown to be a superior classifier for image data, robust against dimension reduction and data contamination.

The SRC Algorithm

- **Set-up**: An $m \times n$ training matrix $\mathcal{X}$, and the labels $y_i \in [1, \ldots, K]$ corresponding to each column x_i of $\mathcal{X}$. And an $m \times 1$ testing vector x for classification. All data are normalized to column-wise unit norm.
- **Find a sparse representation of x in terms of $\mathcal{X}$**: Solve

$$\hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } \|x - \mathcal{X}\beta\|_2 \leq \epsilon. \tag{1}$$

  We use homotopy by *Osborne et al. 2000* [2] and orthogonal matching pursuit (OMP) by *Tropp 2004* [3] to solve this, and bound the number of maximal iterations without using ϵ in our work.
- **Classify x by the sparse representation** $\hat{\beta}$:

$$g(x) = \arg\min_{k=1,\ldots,K} \|x - \mathcal{X}\hat{\beta}_k\|_2, \tag{2}$$

  where $\hat{\beta}_k$ is the class-conditional sparse representation with $\hat{\beta}_k(i) = \hat{\beta}(i)$ if $y_i = k$ and $\hat{\beta}_k(i) = 0$ otherwise. Break ties deterministically.

Theoretical guarantee?

- *Wright et al. 2009* [1] argues that SRC works well for the image data, because empirically different classes of images lie on different subspaces.
- Towards the same direction, *Elhamifar and Vidal 2013* [4] proves a sufficient condition for $L1$ minimization to only choose points from the same subspace, so that sparse representation can work optimally for spectral clustering on data from multiple subspaces.
- *Chen et al. 2013* [5] applies SRC to vertex classification using adjacency matrices and OMP, which exhibits robust performance on graph data, but not always the best classifier.
- But adjacency matrix does not enjoy the subspace property. Also adjacency matrix has $m = n$ such that the residual by $L1$ minimization is usually high at small sparsity limit.

Our questions on SRC and *L*1 minimization

Q1. Since many data do not have the subspace property, is SRC applicable beyond the subspace property?

Q2. The key step of SRC is the *L*1 minimization step (also widely known as Lasso by *Tibshirani 1996* [6]). Since real data is usually noisy and may be high-dimensional (like the (dis)similarity matrices which we care a lot), and a good residual cut-off is hard to estimate, is there a better way to stop the *L*1 minimization without explicit model selection?

(e.g., *Efron et al. 2004* [7] uses Mallows selection criteria for Lasso, *Wright et al. 2009* [1] uses a simple cut-off $\epsilon = 0.05$, *Elhamifar and Vidal 2013* [4] assumes perfect recovery for their theorem.)

Q3. As a greedy algorithm that is very easy to implement, OMP is very popular to give an approximate solution of the exact *L*1 minimization, and a suitable tool for large data processing. Is there any guarantee on its equivalence with *L*1 minimization? (This is discussed by both *Efron et al. 2004* [7] and *Donoho and Tsaig 2006* [8])

A simple guarantee on SRC performance

- In our working paper *Shen et al. 2014* [9], we provide a very coarse error bound of SRC based on within-class principal angles and between-class principal angles. In short, if the former is “smaller” than the latter, SRC may succeed.
- This can help us find meaningful models that can work with SRC beyond the subspace property.
- For example, we further prove that SRC is a consistent classifier for degree-corrected SBM (under one mild condition) applied on the adjacency matrix.
- It is conceptually similar to the condition in *Elhamifar and Vidal 2013* [4], where they also impose a condition so that data on the same subspace is sufficiently close comparing to data of different subspaces. But there are intrinsic differences in the assumption, condition and the proof.

And...

- **Q1** partly solved?! But finite-sample performance is not necessarily optimal.
- What about **Q2** and **Q3**?
- Let us use positive $L1$ minimization!

Positive $L1$ minimization

- Instead of the usual $L1$ minimization, we add one more constraint

$$\hat{\beta} = \arg\min \|\beta\|_1 \text{ subject to } \|x - \mathcal{X}\beta\|_2 \leq \epsilon \text{ and } \beta \geq 0_{n\times 1}, \tag{3}$$

  where the $\geq$ sign is entry-wise. The positive constraint can be easily added to homotopy and OMP with no extra computation.
- It is briefly mentioned in the Lasso implementation using homotopy in *Efron et al. 2004* [7], and called positive Lasso.
- So far we cannot find any other investigation on positive $L1$ minimization, in spite of the rich literature in $L1/L0$ area.

Impact on SRC?

- It usually stops much earlier than usual $L1$ minimization.
- And we prove that OMP is more likely to be equivalent to $L1$ or the true model under the positive constraint.
- It is a bias-variance trade-off?

Section 2

Numerical experiments

Numerical experiments

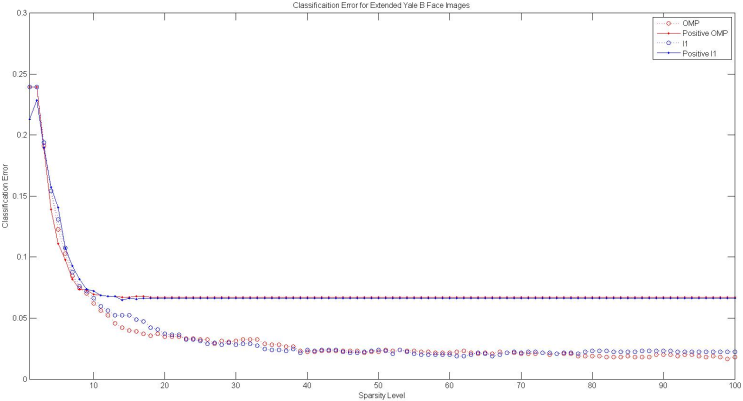
- For all the data, we randomly split half for training and the other half for testing, and plot the hold-out SRC error against the sparsity level, with iteration limit being 100.
- Then we plot the sparsity level histogram of usual/positive OMP/homotopy.
- In order to show that OMP and $L1$ is more likely to be equivalent, we plot the histogram of the following matching statistic

$$p = \sum_{i=1}^{n} I_{\hat{\beta}(i)>0} I_{\beta(i)>0} / \min\{\sum I_{\hat{\beta}(i)>0}, \sum I_{\beta(i)>0}\}. \tag{4}$$

  So if $\hat{\beta}$ and β have nonzero entries at same positions (or a subset of another), $p = 1$; and increasing mismatch will degrade the p towards 0.
- We also show the residual histogram of usual/positive $L1$ minimization.

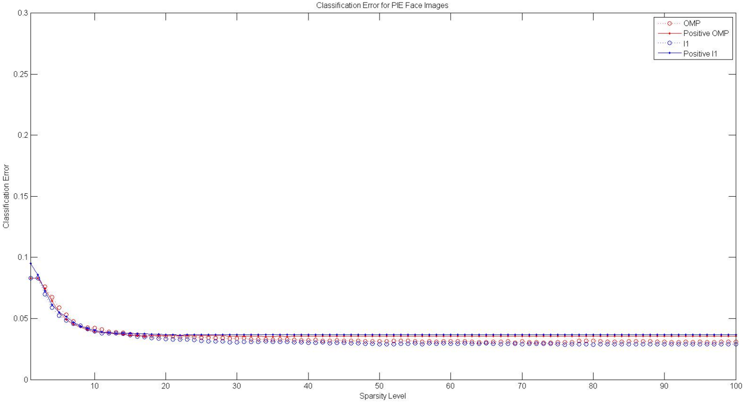
SRC errors for Extended Yale B Images

Extended Yale B database has 2414 face images of 38 individuals under various poses and lighting conditions. So $m = 1024$, $n = 1207$, and $K = 38$. SRC under positive constraint is roughly worse by 0.04.

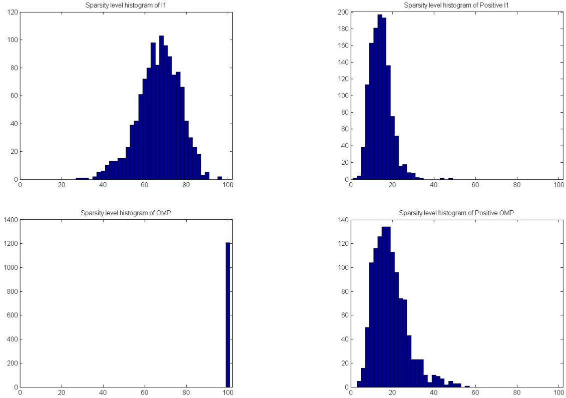


SRC errors on CMU PIE Images

The CMU PIE database has 11554 images of 68 individuals under various poses, illuminations and expressions. $m = 1024$, $n = 5777$, and $K = 68$. SRC under positive constraint is roughly worse by less than 0.01.

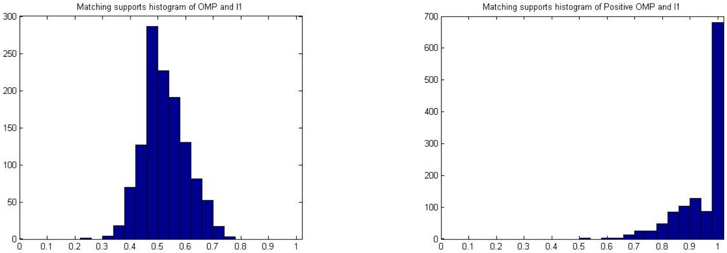


*L*1 comparison in sparsity level for Yale Image



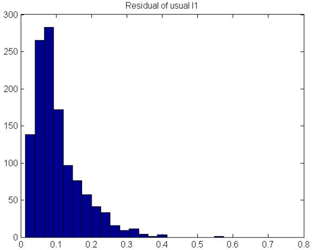
The left side is the number of selected data by usual homotopy/OMP, the right side is that for positive homotopy/OMP.

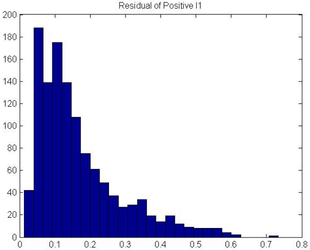
OMP $L1$ equivalence for Yale Image



The left is OMP and homotopy equivalence without positive constraint, the right is with positive constraint.

Residuals for Yale Image

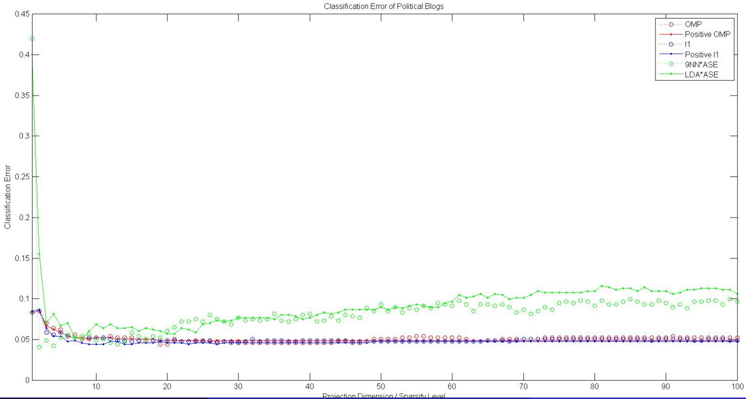




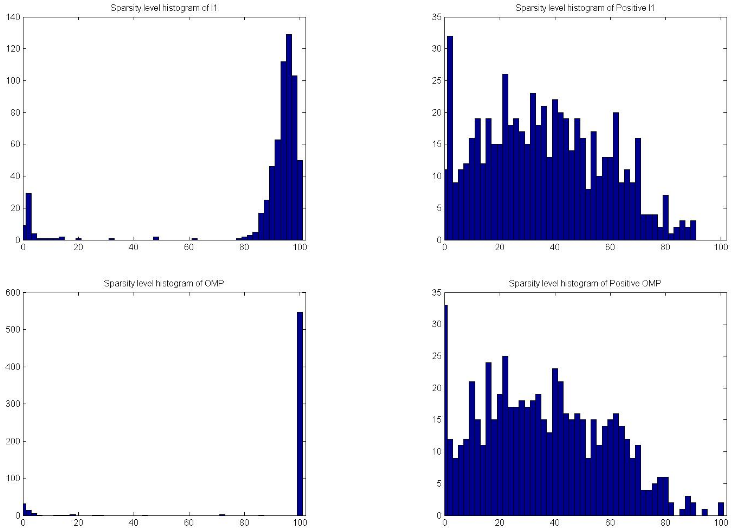
The left is the residual of usual homotopy, the right is the residual of positive homotopy. CMU PIE dataset has similar plots too!

SRC errors on Political Blogs Network

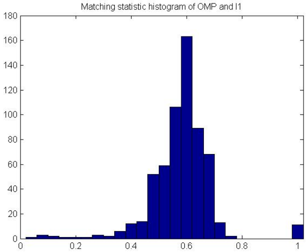
The Political Blogs data is a directed graph of 1490 blogs on conservatives and libertarians, so we have a 1490×1490 adjacency matrix. Among which 1224 vertices have edges, so $m = 1224$, $n = 612$ and $K = 2$. The data can be modeled by DC-SBM. We also add LDA/9NN $\circ$ ASE for comparison.

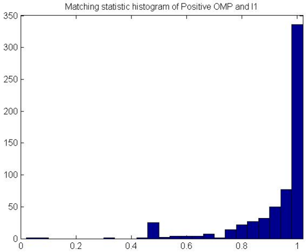


$L1$ comparison in sparsity level for PolBlogs Network

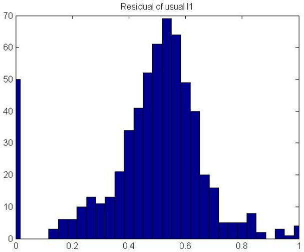


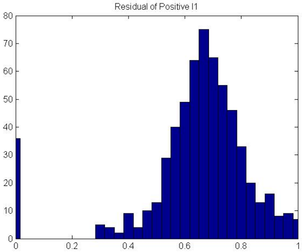
OMP *L*1 equivalence for PolBlogs Network





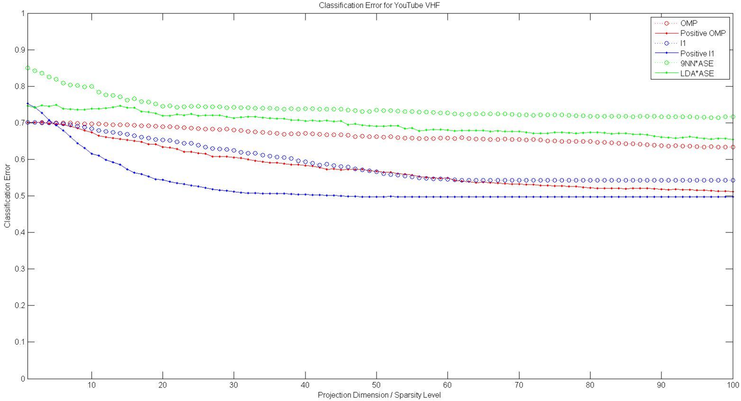
Residuals for PolBlogs Network



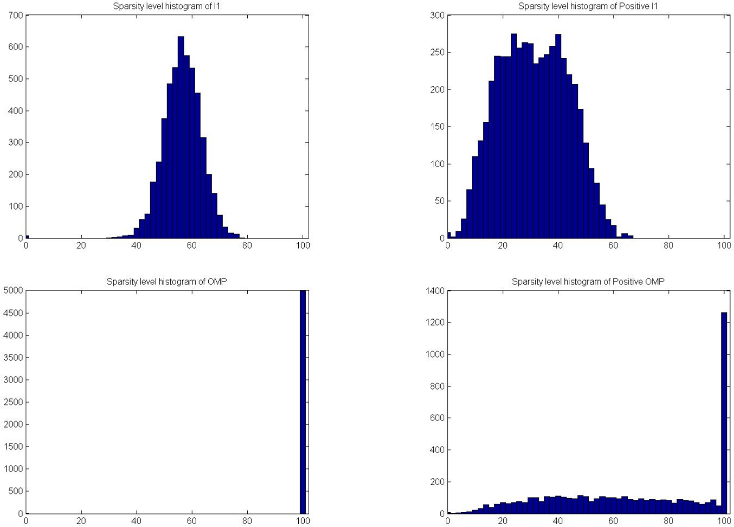


SRC errors on YouTube Video

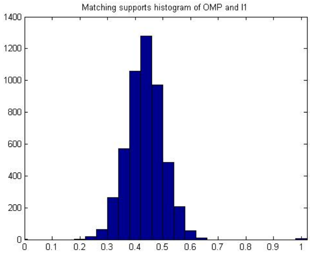
This is a dataset on YouTube game videos containing 12000 videos with 31 game genres. We randomly use 10000 videos and vision hog feature, where we have $m = 650$, $n = 5000$, and $K = 31$. We also add LDA/9NN $\circ$ PCA for comparison.

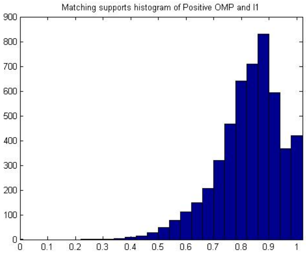


$L1$ comparison in sparsity level for YouTube Video

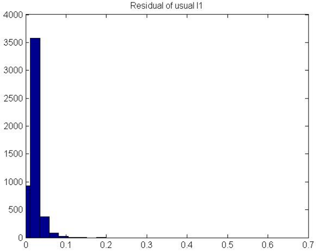


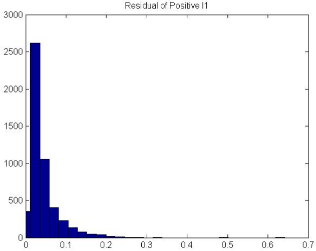
OMP $L1$ equivalence for YouTube Video





Residuals for YouTube Video





Section 3

Conclusion

Conclusion

In this talk, we find partial solutions to our three questions.

- **Q1** We extend SRC beyond the subspace property and generalize it to the graph data theoretically. We also argue that SRC with positive constraint is reasonable.
- **Q2** We show that positive $L1$ minimization terminates much earlier and yield a more parsimonious solution than usual $L1$ minimization (though mostly numerically). This is achieved without any additional model selection, at the cost of slightly larger residual.
- **Q3** From an algorithmic point of view, we show that OMP is more likely to be equivalent to the exact $L1$ minimization/true model under the positive constraint. The improvement is very significant in all our experiments for the equivalence of OMP and homotopy.

However, there are still many unknowns...

References I

J. Wright, A. Y. Yang, A. Ganesh, S. Shankar, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, pp. 389–404, 2000.

J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

L. Chen, J. Vogelstein, and C. E. Priebe, "Robust vertex classification," *submitted, on arxiv*, 2013.

R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

D. Donoho and Y. Tsaig, "Fast solution of l1-norm minimization problems when the solution may be sparse," *preprint*, 2006.

C. Shen, L. Chen, and C. E. Priebe, "Sparse representation classification and positive l1 minimization," *to be submitted*, 2014.

Thank you!