
Sparse Representation Classification via Screening for Graphs

Cencheng Shen^{*1} Li Chen^{*2} Yuexiao Dong³ Carey Priebe⁴

Abstract

The sparse representation classifier (SRC) is shown to work well for image recognition problems that satisfy a subspace assumption. In this paper we propose a new implementation of SRC via screening, establish its equivalence to the original SRC under regularity conditions, and prove its classification consistency for random graphs drawn from stochastic blockmodels. The results are demonstrated via simulations and real data experiments, where the new algorithm achieves comparable numerical performance but significantly faster.

1. Introduction

Sparse coding is a useful and principled tool in machine learning, due to the theoretical advancement in regularized regression and ℓ_1 minimization (Osborne et al., 2000a;b; Donoho & Huo, 2001; Efron et al., 2004; Candes & Tao, 2005; Donoho, 2006; Candes & Tao, 2006; Candes et al., 2006), and it is also effective in numerous classification and clustering applications in computer vision and pattern recognition (Wright et al., 2009; 2010; Yin et al., 2012; Yang et al., 2013; Elhamifar & Vidal, 2013; Chen et al., 2016a).

Here we focus on the sparse representation classification (SRC), which is proposed by (Wright et al., 2009) to combine sparse encoding with computer vision. This classification scheme exhibits state-of-the-art performance for robust face recognition. It has straightforward implementation and works well for data satisfying the subspace assumption (e.g. face recognition, motion segmentation, and activity recognition). Furthermore SRC is a robust classifier against data

contamination and extensive to block-wise SRC and structured data sets (Eldar & Mishali, 2009; Eldar et al., 2010; Elhamifar & Vidal, 2012). In the task of face recognition, where a number of face images $\mathcal{X} = [x_1, \dots, x_n]$ with the corresponding class labels $\mathcal{Y} = [y_1, \dots, y_n]$ are considered the training data, the task is to classify a new testing image x . SRC identifies a small subset $\hat{\mathcal{X}}$ in the training data that best represent the testing image, calculates the least square regression coefficients, and computes the regression residual for classification.

The intrinsic mechanism of SRC is not well-understood. A number of literature suggest that neither ℓ_1 minimization nor the subspace assumption are necessary and sufficient conditions for SRC to achieve high classification accuracy (Rigamonti et al., 2011; Zhang et al., 2011; Shi et al., 2011; Chi & Porikli, 2013). This motivates us to further investigate SRC and the underlying theoretical properties. In this paper we propose a new implementation of SRC via screening, which is much faster than via ℓ_1 minimization with comparable numerical performance. We further prove its theoretical consistency properties on random graphs realized from latent position models, specifically the stochastic blockmodels. Our results make SRC more appealing in terms of theoretical foundation, computational complexity and general applicability.

Our contributions are summarized as follows:

- We are the first to propose sparse representation classification via screening to significantly reduce run time while maintaining classification efficacy and prove theoretical guarantee on its consistency for random graph models.
- We extend sparse representation classification beyond ℓ_1 minimization to achieve variable selection. We analyze the differences of SRC and SRC via screening and establish their equivalence under regularity conditions.
- We demonstrate in simulation and real world networks that SRC via screening has competitive vertex classification performance but significantly faster compared with SRC via ℓ_1 minimization.

^{*}Equal contribution ¹Department of Applied Economics and Statistics, University of Delaware ²Security and Privacy Research, Intel Labs, Hillsboro 97124 ³Department of Statistical Science, Temple University ⁴Department of Applied Mathematics and Statistics, Johns Hopkins University. Correspondence to: Cencheng Shen <shenc@udel.edu>, Li Chen <lichen.jhu1@gmail.com>.

2. Background and Definitions

Denote the training data matrix by $\mathcal{X} = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{m \times n}$, the class label vector by $\mathcal{Y} = [y_1, y_2, \dots, y_n] \in [K]^n$, where m is the feature dimension, n is the number of observations, and K is the number of classes and $[K] = [1, \dots, K]$. A common statistical framework assumes that $(x, y), (x_1, y_1), \dots, (x_n, y_n)$ are independent realizations from the distribution F_{XY} , where $(x, y) \in \mathbb{R}^m \times [K]$ is the testing pair and y is the true but unobserved label. A classifier $g_n(x, D_n)$ is a function that estimates the unknown label $y \in [K]$ based on the training pairs $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ and the testing observation x . For brevity, we always denote the classifier as $g_n(x)$, and the classifier is correct when $g_n(x) = y$. Throughout the paper, we assume all observations are of unit norm ($\|x_i\|_2 = 1$), because SRC scales all observations to unit norm by default.

The sparse representation selects a subset of the training data that best represents the testing observation. Suppose s is the sparsity level, we denote the subset of training data as $\hat{\mathcal{X}} = [x_{(1)}, x_{(2)}, \dots, x_{(s)}] \in \mathbb{R}^{m \times s}$. Once $\hat{\mathcal{X}}$ is determined, $\hat{\beta}$ is the $s \times 1$ least square regression coefficients between $\hat{\mathcal{X}}$ and x , and the regression residual equals $\|x - \hat{\mathcal{X}}\hat{\beta}\|_2$. For each $k \in [K]$ and a given $\hat{\mathcal{X}}$, we define

$$\begin{aligned}\hat{\mathcal{X}}_k &= \{x_{(i)} \in \hat{\mathcal{X}}, i = 1, \dots, s \mid y_{(i)} = k\} \\ \hat{\mathcal{X}}_{-k} &= \{x_{(i)} \in \hat{\mathcal{X}}, i = 1, \dots, s \mid y_{(i)} \neq k\}.\end{aligned}$$

Namely, $\hat{\mathcal{X}}_k$ is the subset of $\hat{\mathcal{X}}$ that contains all observations from class k , and $\hat{\mathcal{X}}_{-k} = \hat{\mathcal{X}} - \hat{\mathcal{X}}_k$. We further denote $\hat{\beta}_k$ as the regression coefficients of $\hat{\beta}$ corresponding to $\hat{\mathcal{X}}_k$, and $\hat{\beta}_{-k}$ as the regression coefficients corresponding to $\hat{\mathcal{X}}_{-k}$, i.e.,

$$\hat{\mathcal{X}}_k \hat{\beta}_k + \hat{\mathcal{X}}_{-k} \hat{\beta}_{-k} = \hat{\mathcal{X}} \hat{\beta}.$$

The original SRC makes use of the class-wise regression residual $\|x - \hat{\mathcal{X}}_k \hat{\beta}_k\|_2$ in Algorithm 1.

Sparse Representation Classification by ℓ_1

SRC consists of three steps: subset selection, least square regression, and classification. Algorithm 1 describes the original algorithm: Equation 1 identifies the sparse representation, and solves the least square regression coefficients $\hat{\beta}$ at the same time. Then Equation 3 assigns the class by minimizing the class-wise regression residual. Computation-wise, the ℓ_1 minimization step takes at least $O(mns)$, while the classification step is much cheaper and takes $O(msK)$.

The ℓ_1 minimization step is a crucial step to the success of SRC, but also the most computationally expensive step of SRC. There exists various iterative implementations of similar complexity, such as ℓ_1 homotopy method (Osborne et al.,

Algorithm 1 Sparse representation classification by ℓ_1 minimization

Input: The training data matrix \mathcal{X} with associated label vector \mathcal{Y} , and the testing observation x .

ℓ_1 Minimization: For each testing observation x , find $\hat{\mathcal{X}}$ and $\hat{\beta}$ that solves the ℓ_1 minimization problem:

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

Classification: Assign the testing observation by minimizing the class-wise residual, i.e.,

$$g_n^{\ell_1}(x) = \arg \min_{k \in [K]} \|x - \hat{\mathcal{X}}_k \hat{\beta}_k\|_2, \quad (2)$$

break ties deterministically.

Output: The estimated class label $g_n^{\ell_1}(x)$.

2000a;b; Efron et al., 2004), orthogonal matching pursuit (OMP) (Tropp, 2004; Tropp & Gilbert, 2007), augmented Lagrangian method (Yang et al., 2013), among many others. We use the homotopy algorithm for subsequent analysis and numerical comparison without delving into the algorithmic details.

Note that model selection of s is inherent to the ℓ_1 minimization problem. One need to either specify a tolerance noise level ϵ or a maximum sparsity level in order for the iterative algorithm to stop. The choice does not affect the theoretical result, but can impact the actual numerical performance and thus a separate topic for investigation (Zhang, 2009; Cai & Wang, 2011).

Stochastic Blockmodels

Definition (Directed and Unweighted Stochastic Blockmodel (SBM)). *Given the class membership \mathcal{Y} , a directed stochastic blockmodel generates an $n \times n$ binary adjacency matrix \mathcal{X} via a class connectivity matrix $V \in [0, 1]^{K \times K}$ by Bernoulli distribution $\mathbf{B}(\cdot)$:*

$$\mathcal{X}(i, j) = \mathbf{B}(V(y_i, y_j)).$$

From the definition, the adjacency matrix produced by SBM is a high-dimensional object that is characterized by a low-dimensional class connectivity matrix.

Contamination Scheme

Definition (Stochastic Blockmodel with Contamination). *Under the stochastic blockmodel, for each class k define V_k*

as the contamination vector of size $1 \times m$:

$$\begin{aligned} V_k(j) &= 1 \text{ when } j\text{th dimension is not contaminated,} \\ V_k(j) &= 0 \text{ when } j\text{th dimension is contaminated.} \end{aligned}$$

Then the contaminated random variable X is

$$X|Y = \text{diag}(V_Y)W_Y U,$$

where $\text{diag}(V_k)$ is a diagonal matrix of size $m \times m$ and $\text{diag}(V_k)(j, j) = V_k(j)$.

In Section 4, we use the above contamination scheme to contaminate both the simulated and the real world networks and show SRC via screening achieves the same level of robustness as the original SRC algorithm.

3. Main Results

In this section, we present our proposed SRC algorithm via screening, which differs from the original SRC algorithm mostly in the subset selection step and slightly in the classification step. We investigate the theoretical properties of the classification step by proving classification consistency for stochastic blockmodels.

3.1. SRC via Screening

Our proposed algorithm is shown in Algorithm 2, which replaces the ℓ_1 minimization step by screening, and minimizes the class-wise residual in angle in the classification step. Algorithm 2 obtains better computation complexity because the screening procedure becomes non-iterative and simply chooses s observations out of $\hat{\mathcal{X}}$ that are most correlated with the testing observation x as $\hat{\mathcal{X}}$. This step merely requires $O(mn + n \log(n))$ instead of $O(mns)$ for ℓ_1 .

Indeed, the screening procedure has recently gained popularity as a fast alternative of regularized regression for high-dimensional data inference. The speed advantage makes it a suitable candidate for efficient data extraction, and is shown to be equivalent to ℓ_1 and ℓ_0 minimization under various regularity conditions (Fan & Lv, 2008; Fan et al., 2009; Wasserman & Roeder, 2009; Fan et al., 2011; Genovese et al., 2012; Kolar & Liu, 2012). In particular, it is argued that the maximal sparsity level $s = \max\{n/\log(n), m\}$ works well for screening (Fan & Lv, 2008). Thus we also set the same sparsity level in our experiments in Section 4.

3.2. Consistency Under Stochastic Block-Model

Here we prove SRC consistency for the stochastic block-model (Holland et al., 1983; Sussman et al., 2012; Lei & Rinaldo, 2015), which is a popular network model commonly used for classification and clustering. Although the results are applicable to undirected, weighted, and other sim-

Algorithm 2 Sparse representation classification via screening for vertex classification in stochastic blockmodels

Input: The adjacency \mathcal{X} , the known label vector \mathcal{Y} , and the testing vertex x .

Screening: Calculate $\Omega = \{x_1^T x, x_2^T x, \dots, x_n^T x\}$ (T is the transpose), and sort the elements by decreasing order. Take $\hat{\mathcal{X}} = \{x_{(1)}, x_{(2)}, \dots, x_{(s)}\}$ with $s = \min\{n/\log(n), m\}$, where $x_{(i)}$ is the i th largest element in Ω .

Regression: Solve the ordinary least square problem between $\hat{\mathcal{X}}$ and x . Namely, compute $\beta = \hat{\mathcal{X}}^{-1}x$ where $\hat{\mathcal{X}}^{-1}$ is the Moore-Penrose inverse.

Classification: Assign the testing vertex by

$$g_n^{\text{scr}}(x) = \arg \min_{k \in [K]} \theta(x, \hat{\mathcal{X}}_k \hat{\beta}_k), \quad (3)$$

where θ denotes the principal angle. Break ties deterministically.

Output: The estimated class label $g_n^{\text{scr}}(x)$.

ilar graph models, for ease of presentation we concentrate on the directed and unweighted SBM.

Theorem 1. Denote $\rho \in [0, 1]^K$ as the prior probability of each class, SRC by screening is consistent for vertex classification under SBM when

$$\frac{\rho \circ V(Y, :) V(Y, :)^T}{\|\rho \circ V(Y, :)\|_1^{1/2}} > \frac{\rho \circ V(Y', :) V(Y', :)^T}{\|\rho \circ V(Y', :)\|_1^{1/2}} \quad (4)$$

holds for $Y' \neq Y$, where \circ denotes the entry-wise product.

The condition also guarantees data of the same class to be more similar in angle than data of different classes, thus inherently the same as the principal angle condition for latent subspace mixture model.

4. Numerical Experiments

In this section we apply the SRC screening algorithm to simulated and real world network graphs. The evaluation criterion is the leave-one-out error for vertex classification: within each vertex set, one vertex is held out for testing and the remaining are used for training, do the classification, and repeat until each observation in the given data is hold-out once. SRC is shown to be a robust vertex classifier in (Chen et al., 2016a) with superior performance to other classifiers for both simulated and real networks. We use the SRC algorithm's performance for vertex classification as a baseline for comparison.

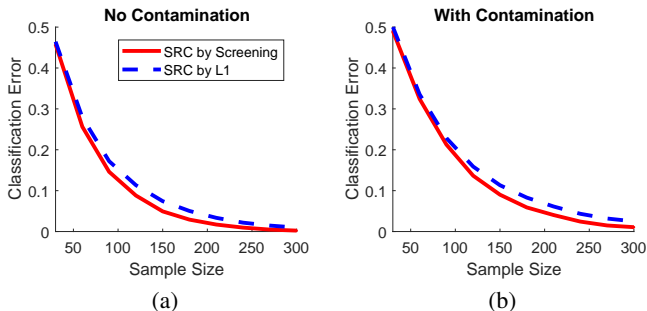


Figure 1. SRC screening achieves the same classification error as SRC for vertex classification on simulated graphs.

Table 1. Leave-one-out error and run time (s) comparison

Algorithm	Metrics	Wikipedia	<i>C.elegans</i>
SRC by Screening	Error	32.27%	42.29%
SRC by ℓ_1	Error	29.31%	48.62%
SRC by Screening	Runtime	32.3	0.3
SRC by ℓ_1	Runtime	573.7	9.1

Our simulation and real data experiments show that SRC is consistent under stochastic blockmodel and robust against contamination on graphs. Overall, we observe that SRC by screening performs very similar to SRC without screening, but with much faster run time for training.

4.1. Stochastic Blockmodel Simulation

We simulate graphs from the stochastic blockmodel with number of blocks $K = 3$, membership prior ($\rho_1 = \rho_3 = 0.3, \rho_2 = 0.4$) and the class connectivity matrix V where

$$V = \begin{bmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.3 \end{bmatrix}.$$

The class connectivity matrix satisfies the condition in Theorem 1. From the SBM, we simulate graphs with varied sizes for $n = 30, 60, \dots, 300$, compute the leave-one-out error, then repeat for 100 Monte-Carlo replicates and plot the average errors in Figure 1.

4.2. Network Connectivity

Next we apply SRC to vertex classification on real-world networks. The first graph is collected from Wikipedia article hyperlinks (Priebe et al., 2013). A total of 1382 English documents based on the 2-neighborhood of the English article “algebraic geometry” are collected, and the adjacency matrix is formed via the documents’ hyperlinks. This is a directed, unweighted, and sparse graph without self-loop, where the graph density is 1.98%. There are five classes based on article contents: 119 articles in category class,

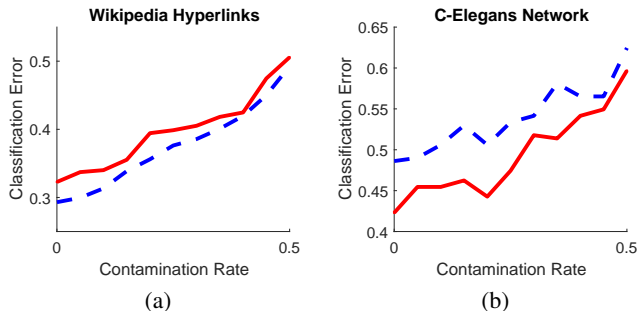


Figure 2. SRC on contaminated real networks. SRC via screening and SRC have similar robustness against contamination. On the *C.elegans* neuro-connectome, SRC via screening maintains more robustness than SRC in terms of lower error, as the contamination rate increases.

372 articles about people, 270 articles about locations, 191 articles on date, and 430 articles are real math.

The second graph data is the electric neural connectome of *Caenorhabditis elegans* (*C.elegans*) (Hall & Russell, 1991; Varshney et al., 2011; Chen et al., 2016b). The hermaphrodite *C.elegans* somatic nervous system has over two hundred neurons, classified into 3 classes: motor neurons, interneurons, and sensory neurons. The adjacency matrix is also undirected, unweighted, and sparse with density 1.32%.

The leave-one-out errors and run times are reported in Table 1. On the Wikipedia graph, SRC with screening has slightly higher error rate at 32.27% than SRC without screening at 29.31%. On the *C.elegans* neuro-connectome, the SRC with screening is much lower at 42.29% than SRC without screening at 48.62%. We further conducted contamination analysis on these two real networks. The contaminated classification performance are shown in the bottom panels of Figure 2, where the solid lines are performance by SRC via screen and the dotted lines are performance by the original SRC. SRC via screening demonstrated more robustness in terms of lower classification error, compared with SRC without screening on the *C.elegans* neuro-connectome.

5. Conclusion

In this paper, we propose sparse representation classification via screening for random graphs distributed as stochastic blockmodels. We prove the theoretical consistency and show in simulation and real world experiments the effectiveness and robustness of our proposed algorithm. We continue our research of extending SRC via screening to latent subspace mixture models and quantifying robustness against contamination.

References

- Cai, T. and Wang, L. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.
- Candes, E. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- Candes, E. and Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.
- Candes, E., Romberg, J., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1233, 2006.
- Chen, L., Shen, C., Vogelstein, J. T., and Priebe, C. E. Robust vertex classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):578–590, 2016a.
- Chen, L., Vogelstein, J. T., Lyzinski, V., and Priebe, C. E. A joint graph inference case study: the c.elegans chemical and electrical connectomes. *Worm*, 5(2):1, 2016b.
- Chi, Y. and Porikli, F. Classification and boosting with multiple collaborative representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1519–1531, 2013.
- Donoho, D. For most large underdetermined systems of linear equations the minimal ℓ_1 -norm near solution approximates the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(10):907–934, 2006.
- Donoho, D. and Huo, X. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47:2845–2862, 2001.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.
- Eldar, Y. and Mishali, M. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.
- Eldar, Y., Kuppinger, P., and Bolcskei, H. Compressed sensing of block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010.
- Elhamifar, E. and Vidal, R. Block-sparse recovery via convex optimization. *IEEE Transactions on Signal Processing*, 60(8):4094–4107, 2012.
- Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.
- Fan, J. and Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911, 2008.
- Fan, J., Samworth, R., and Wu, Y. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.
- Fan, J., Feng, Y., and Song, R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.
- Genovese, C., Lin, J., Wasserman, L., and Yao, Z. A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, 13:2107–2143, 2012.
- Hall, D. H. and Russell, R. The posterior nervous system of the nematode *Caenorhabditis elegans*: serial reconstruction of identified neurons and complete pattern of synaptic interactions. *The Journal of Neuroscience*, 11(1):1–22, 1991.
- Holland, P., Laskey, K., and Leinhardt, S. Stochastic block-models: First steps. *Social Networks*, 5(2):109–137, 1983.
- Kolar, M. and Liu, H. Marginal regression for multitask learning. *Journal of Machine Learning Research W & CP*, 22:647–655, 2012.
- Lei, J. and Rinaldo, A. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.
- Osborne, M., Presnell, B., and Turlach, B. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000a.
- Osborne, M., Presnell, B., and Turlach, B. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000b.
- Priebe, C. E., Marchette, D. J., Ma, Z., and Adali, S. Manifold matching: Joint optimization of fidelity and commensurability. *Brazilian Journal of Probability and Statistics*, 27(3):377–400, 2013.
- Rigamonti, R., Brown, M., and Lepetit, V. Are sparse representations really relevant for image classification? In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

Shi, Q., Eriksson, A., Hengel, A., and Shen, C. Is face recognition really a compressive sensing problem? In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

Sussman, D., Tang, M., Fishkind, D., and Priebe, C. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

Tropp, J. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Tropp, J. and Gilbert, A. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., and Chklovskii, D. B. Structural properties of the caenorhabditis elegans neuronal network. *PLoS computational biology*, 7(2):e1001066, 2011.

Wasserman, L. and Roeder, K. High dimensional variable selection. *Annals of statistics*, 37(5A):2178–2201, 2009.

Wright, J., Yang, A. Y., Ganesh, A., Shankar, S., and Ma, Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., and Yan, S. Sparse representation for computer vision and pattern recognition. *Proceedings of IEEE*, 98(6): 1031–1044, 2010.

Yang, A., Zhou, Z., Ganesh, A., Sastry, S., and Ma, Y. Fast l_1 -minimization algorithms for robust face recognition. *IEEE Transactions on Image Processing*, 22(8):3234–3246, 2013.

Yin, J., Liu, Z., Jin, Z., and Yang, W. Kernel sparse representation based classification. *Neurocomputing*, 77(1): 120–128, 2012.

Zhang, L., Yang, M., and Feng, X. Sparse representation or collaborative representation: which helps face recognition?. In *International Conference on Computer Vision (ICCV)*, 2011.

Zhang, T. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

Appendix: Sketch of proof

Theorem 1

Proof Sketch. Denote x as the testing adjacency vector of size $1 \times n$, and x_1 as a training adjacency vector of size $1 \times n$. Then

$$\begin{aligned} \cos \theta(x, x_1) &= \frac{\sum_{j=1}^n \mathcal{B}(V(y, Y_j)V(y_1, Y_j))}{\sqrt{\sum_{j=1}^n \mathcal{B}(V(y, Y_j)) \sum_{j=1}^n \mathcal{B}(V(y_1, Y_j))}} \\ &\stackrel{n \rightarrow \infty}{\rightarrow} \frac{E(V(y, Y)V(y_1, Y))}{E(V(y, Y))E(V(y_1, Y))} \\ &= \frac{\sum_{k=1}^K \rho_k V(y, k)V(y_1, k)}{\sqrt{\sum_{k=1}^K \rho_k V(y, k) \sum_{k=1}^K \rho_k V(y_1, k)}} \\ &= \frac{\rho \circ V(y, :)V(y_1, :)^T}{\sqrt{\|\rho \circ V(y, :)\|_1 \|\rho \circ V(y_1, :)\|_1}} \end{aligned}$$

It follows that when Equation 4 holds, $\cos \theta(X, X') > \cos \theta(X, X_1)$ always holds asymptotically for $Y = Y' \neq Y_1$. \square