Supplementary File:
Data-Dependent Sparsity for Subspace Clustering

Bo Xin
Microsoft Research, Beijing

Yizhou Wang
Peking University

Wen Gao
Peking University

David Wipf
Microsoft Research, Beijing

1 INTRODUCTION

This document contains companion information regarding our UAI 2017 submission, including supporting experiments and technical proofs. Note that herein all equation numbers referencing back to the main submission document will be be prefixed with an 'M' to avoid confusion, i.e, (M.#) will refer to equation (#) from the main text. Similar notation differentiates sections, tables, and figures, e.g., Section M.#.

2 COMPARISONS WITH Aℓ_0-SSC USING SYNTHETIC DATA

The Aℓ_0-SSC algorithm (Yang et al., 2016) attempts to approximately solve (M.3) by first computing an ℓ_1 norm initialization and then later refining it via iterative hard thresholding (IHT) iterations (Blumensath and Davies, 2008). However, as we have discussed in Section M.4 and elsewhere in our submission, both the ℓ_1 norm and IHT iterations are quite sensitive to data correlations, and the latter may not be able to appreciable improve upon the former in recovering maximally sparse, subspace-aligned representations. To the extent that this claim is true, we would then expect ℓ_1-SSC and Aℓ_0-SSC to display similar performance in difficult practical situations where such correlation structure is unavoidable.

Indeed an existing experimental paradigm originally from (Soltanolkotabi and Candes, 2012), and investigated in Section M.4, serves to illustrate this point. In particular, Figure 1 represents a reproduction of Figure M.1 under the ideal linear setting (meaning no affine translations), but with the inclusion of results for Aℓ_0-SSC. Note that the released code from (Yang et al., 2016) does not directly address affine subspaces, and there is no standard way of adapting IHT iterations to include the required equality constraints used with existing methods (Elhamifar and Vidal, 2013) to handle the affine model. Hence we restrict our comparisons with Aℓ_0-SSC to ideal, linear subspaces using released code.[1] Additionally, we experiment with different threshold values for the IHT step; however, if this value is set too low, there is no improvement at all as IHT becomes provably stuck at the ℓ_1-norm based initialization, and if it is set too high, the performance can actually become *worse* than just ℓ_1-SSC itself. In contrast, for DD-SSC we do not tune any parameters, and just set $\alpha = 10^{-10}$ to closely enforce the constraint (see Section M.2), a rather arbitrary default setting near zero.[2]

From Figure 1 we observe that with Aℓ_0-SSC, the results are not much improved beyond that of ℓ_1-SSC, while DD-SSC produces considerably lower errors than both methods. For example, when the subspace intersection dimension (a measure of problem difficulty as described in Section M.4) is 8, the error rate of DD-SSC is less than half that of both ℓ_1-SSC and Aℓ_0-SSC. This reinforces our original claim that Aℓ_0-SSC may provide negligible gain over ℓ_1-SSC in challenging practical conditions, unlike DD-SSC which consistently supplies an advantage. And it should be emphasized that this experiment, which originated from (Soltanolkotabi and Candes, 2012), involves a somewhat ideal, uniform distribution of data points within subspaces. If we were to introduce further correlations within or between subspaces, both the ℓ_1 solution and IHT iterations can become even less reliable.

3 ADDITIONAL MOTION DATA EXPERIMENTS

Here we provide additional experiments that reflect real-world situations involving corrupted and outlying motion data.

[1] https://github.com/yingzhenyang/L0-SSC

[2] In fact any small value for α produces the same results, but $\alpha = 0$ can lead to some numerical instability upon nearing convergence because of a potentially ill-conditioned matrix inverse. With additional effort however, even this special case can be technically handled using judicious use of pseudoinverses.

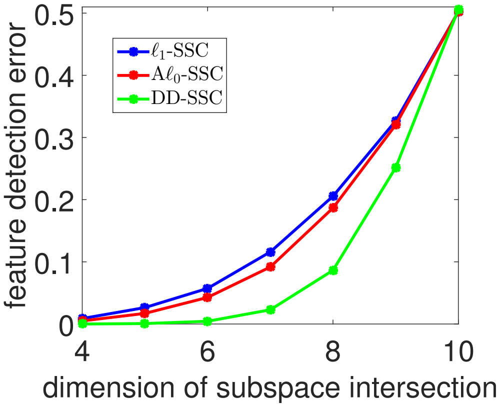

Figure 1: Feature detection error as a function of the subspace intersection dimension t. Any possible algorithm cannot have error below 0.5 on average once $t = 10$, at which point the two subspaces merge into one, the problem is no longer identifiable, and random guessing will achieve a 50% error rate. This plot corresponds to the ideal setting of Figure M.1 (the affine case is omitted here because the released code for Aℓ_0-SSC does not presently handle the additional embedded equality constraint).

3.1 Recovery of Partially Corrupted Trajectories

The Hopkins 155 motion data set contains 156 video sequences, each of which has 39-550 data points drawn from two or three motions (a motion corresponds to a subspace). Conventionally, subspace clustering works were evaluated on this data and the performance of most algorithms achieved nearly perfect segmentation accuracy (over 95%). Therefore using this dataset is difficult to differentiate the ability of different algorithms. However, as shown in (Rao et al., 2010), many additional confounding factors can be introduced to this data set to make it more challenging. These factors include missing or corrupted trajectories and outlying trajectories. We have discussed outlier detection on this dataset in the main text and will focus on dealing with corrupted trajectories here.

Following the protocol from (Rao et al., 2010), we partially corrupt 12 motion sequences from the Hopkins data with missing entries, including 9 with 2 motions, and 3 with 3 motions. The corruption rate is between 4% and 35% of the entries in the data matrix of trajectories. These entries were manually located and labeled. In the ground truth image of Figure 2, we illustrate the corrupted positions in a representative data matrix derived from this set.

For DD-SSC to incorporate such sparse element-wise corruptions, we simply need to concatenate an identity matrix of size $d+1$ to the original dictionary forming $\boldsymbol{X}^{++} \triangleq [\boldsymbol{X}^{+}, \boldsymbol{I}^{+}]$ (where $\boldsymbol{X}^{+} \triangleq [\boldsymbol{X}; \mathbf{1}_{(n)}^{\top}]$ and $\boldsymbol{I}^{+} \triangleq \left[[\boldsymbol{I}, \mathbf{0}_{(n-1)}]; \mathbf{0}_n^{\top}\right]$ were introduced in Section M.2) and update the dimensionality of the sparse prior (M.5) and hyperparameters accordingly. Nonzero entries in the expanded $\boldsymbol{\gamma}_i$ vector that correspond with this extra $\boldsymbol{I}^{+}$ factor will then reflect corrupted entries.

We compared this corruption estimation strategy using DD-SSC with a competing ℓ_1-SSC enhancement for dealing with missing/corrupted entries (Elhamifar and Vidal, 2013). On average, we achieve a feature detection error (defined in Section M.4) of 0.029 for DD-SSC on all these corrupted sequences. As a comparison, the average detection error of ℓ_1-SSC is 0.064. Moreover, in order to check the estimated positions of the corrupted entries with respect to the ground truth label, we define the estimation error as $\frac{\#\text{mismatch entries}}{\#\text{total entries}}$. The average error of DD-SSC is 11.4% whereas that of ℓ_1-SSC is 14.2%. In Figure 2, we illustrate one representative estimation example. We see that while DD-SSC can successfully detect most of the corruptions without many false positive, ℓ_1-SSC exhibits many false positives even after we truncate minor values (those less than 0.0001).

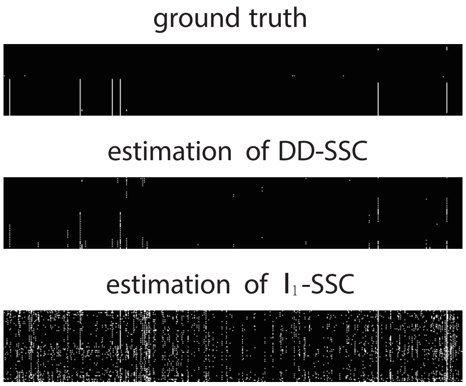

Figure 2: Estimates of the corrupted positions in a representative data matrix of Hopkins 155 data set. The data matrix $\boldsymbol{X}$ is presented as a black rectangle with white pixels indicating corrupted positions.

3.2 Outlier Detection ROC Curve

In Table M.2, we displayed the the performance of ℓ_1-SSC and DD-SSC using an outlier detection accuracy defined as $\frac{\#\text{ correctly found inliers}}{\#\text{ total inliers}}$. Following the literature, in Figure 3, we provide an additional, complementary ROC curve associated with the case where the number of inliers and outliers are equal (other ROC curves show a similar pattern).

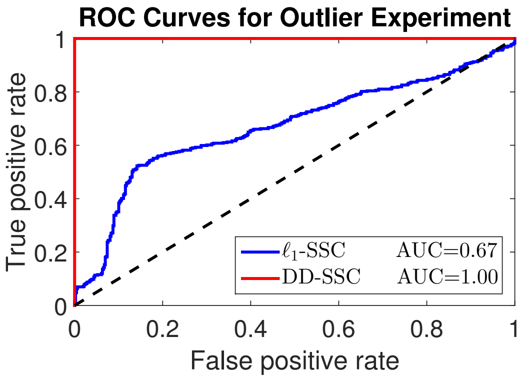

Figure 3: ROC curves for outlier detection.

4 TECHNICAL DETAILS REGARDING THEOREM 1

For the i-th data point, we begin from an initial weight vector $\boldsymbol{w}^{(0)} = \mathbf{1}$ and then proceed to the $(t+1)$-th iteration by computing

$$\begin{aligned} \boldsymbol{z}_i^{(t+1)} &\leftarrow \arg\min_{\boldsymbol{z}_i} \sum_j w_j^{(t)} |z_{ij}| \quad \text{s.t. } \boldsymbol{x}_i = \boldsymbol{X}_{\bar{i}} \boldsymbol{z}_i \\ \boldsymbol{w}^{(t+1)} &\leftarrow \eta_j(\boldsymbol{z}_i^{(t+1)}; \alpha, q), \end{aligned} \tag{1}$$

where the $|\cdot|$ operator is understood to apply element-wise and

$$\eta_j(\boldsymbol{z}_i; \alpha, q) \triangleq \left[\boldsymbol{x}_j^\top \left(\alpha I + \boldsymbol{X}_{\bar{i}} |Z_i|^2 \boldsymbol{X}_{\bar{i}}^\top \right)^{-1} \boldsymbol{x}_j \right]^q . \tag{2}$$

These updates comprise the generalized iterative reweighted ℓ_1 variant of DD-SSC.

Theorem 1 *The DD-SSC updates produced by (1) satisfy the following:*

1. *If at any iteration we compute a subspace optimal solution $\boldsymbol{z}_i^{(t)}$, then all subsequent iterations are guaranteed to be subspace optimal for any $\alpha \in (0, \alpha']$, provided α' is sufficiently small.*
2. *For any identifiable configuration of subspaces, some $q \geq 1/2$, $\alpha \in (0, \alpha']$, and α' sufficiently small, there will always exist configurations of points within each subspace such that the iterations are guaranteed to produce a subspace optimal solution for all i.*

Proof: For convenience we will adopt the notation $f(x) = \mathrm{O}(h(\epsilon))$ to indicate that $|f(x)| < C|h(\epsilon)|$ for some constant C independent of x or ϵ, and $f(x) = \Omega(h(\epsilon))$ to indicate that $|f(x)| > C|h(\epsilon)|$ for some constant C independent of x or ϵ. Likewise, we say $f(x) = \Theta(h(\epsilon))$ iff $f(x) = \mathrm{O}(h(\epsilon))$ and $f(x) = \Omega(h(\epsilon))$.

We begin with the first part of the theorem and the following qualification to ensure that $\eta_j(\boldsymbol{z}_i; \alpha, q)$ is well-defined even in the limit $\alpha \to 0$. Given that

$$\lim_{\epsilon \to 0} \boldsymbol{U}^\top \left(\epsilon \boldsymbol{I} + \boldsymbol{U}\boldsymbol{U}^\top \right)^{-1} \boldsymbol{U} = \boldsymbol{U}^\dagger \boldsymbol{U} = \boldsymbol{U}^\top \left(\boldsymbol{U}\boldsymbol{U}^\top \right)^\dagger \boldsymbol{U} \tag{3}$$

for any matrix $\boldsymbol{U}$, we have that

$$\lim_{\alpha \to 0} \eta_j(\boldsymbol{z}_i; \alpha, q) = \left[\boldsymbol{x}_j^\top \left(\boldsymbol{X}_{\bar{i}} |Z_i|^2 \boldsymbol{X}_{\bar{i}}^\top \right)^\dagger \boldsymbol{x}_j \right]^q \tag{4}$$

if $\boldsymbol{x}_j$ is in the span of the right singular vectors of $\boldsymbol{X}_{\bar{i}} |Z_i|$, and

$$\lim_{\alpha \to 0} \eta_j(\boldsymbol{z}_i; \alpha, q) \quad \triangleq \quad \infty \tag{5}$$

otherwise.

Now suppose that for some arbitrary point $\boldsymbol{x}_i \in \mathcal{S}_k$ we have computed a subspace optimal solution at any iteration t with $\alpha \in (0, \alpha']$. Then based on the above, the corresponding weight $w_j^{(t)} = \eta_j\left(\boldsymbol{z}_i^{(t)}; \alpha, q\right)$ will be $\Theta(1)$ if j indexes a point in the correct subspace, while $w_j^{(t)} = \Theta(\alpha^{-q})$ if j is not in the correct subspace (i.e., $j \notin \mathcal{S}_k$), where $\lim_{\alpha \to 0} w_j^{(t)} = \infty$. Therefore at the next iteration we must solve

$$\boldsymbol{z}_i^{(t+1)} \leftarrow \arg\min_{\boldsymbol{z}_i} \sum_j w_j^{(t)} |z_{ij}| \quad \text{s.t. } \boldsymbol{x}_i = \boldsymbol{X}_{\bar{i}} \boldsymbol{z}_i. \tag{6}$$

Based on (Rao and Kreutz-Delgado, 1999), any minimum of (6) can be achieved at a so-called basic feasible solution satisfying $\|\boldsymbol{z}_i\|_0 \leq d$ (Luenberger and Ye, 1984) (of which there exist a finite number with points in general position, although this is not a strict requirement). Let η^- denote the smallest value of $\sum_{j \notin C_k} |z_{ij}|$ for any non-subspace-optimal basic feasible solution, and define

$$\eta^+ = \min_{\boldsymbol{z}_i} \sum_{j \notin C_k} |z_{ij}| \quad \text{s.t. } \boldsymbol{x}_i = \boldsymbol{X}_{\bar{i}} \boldsymbol{z}_i, \;\; z_{ij} = 0, \; \forall j \notin C_k. \tag{7}$$

It then follows that any non-subspace optimal solution to (6) will have

$$\begin{aligned} \sum_j w_j^{(t)} |z_{ij}| \;&\geq\; \sum_{j \notin C_k} w_j^{(t)} |z_{ij}| \\ &\equiv\; \Theta(\alpha^{-q}) \sum_{j \notin C_k} |z_{ij}| \;>\; \Theta(\alpha^{-q}) \eta^- \;\equiv\; \Theta(\alpha^{-q}). \end{aligned} \tag{8}$$

Furthermore, any subspace optimal solution will satisfy

$$\sum_j w_j^{(t)} |z_{ij}| \;\equiv\; \Theta(1) \eta^+ \;\equiv\; \Theta(1) \;<\; \Theta(\alpha^{-q}) \tag{9}$$

if α' (and therefore any allowable α) are sufficiently small. This implies that a subspace optimal solution, with $\boldsymbol{z}_{ij}^{(t+1)} = 0$ for all $j \notin \mathcal{S}_k$, is preserved if α' (and therefore α) are sufficiently small, otherwise the weighted ℓ_1 norm will be driven to an arbitrarily large value, possibly infinity. Hence subsequent iterations can only change the support set *within* the optimal subspace; they cannot tarnish an existing subspace optimal solution.

We now turn to the second part of the theorem. Our strategy will be to demonstrate that for certain selections of α and q, we can guarantee that reweighted ℓ_1 using $\eta_j(\boldsymbol{z}_i; \alpha, q)$ is guaranteed to produce a subspace optimal solution if the columns of $\boldsymbol{X}$ associated with each subspace $\mathcal{S}_k$ are sufficiently close together. More precisely, suppose every column of $\boldsymbol{X}$ originally drawn from $\mathcal{S}_k$ can be expressed as

$$\boldsymbol{x}_i = \boldsymbol{\mu}_k + \mathrm{O}\left(\delta\right), \tag{10}$$

where $\boldsymbol{\mu}_i \in \mathcal{S}_k$. In other words, every column within the k-th subspace can be expressed as a small offset $\mathrm{O}\left(\delta\right)$ from some mean vector $\boldsymbol{\mu}_k$ that can be made arbitrarily small as $\delta \rightarrow 0$, which is similar in spirit to a correlated dictionary model from (Wu and Wipf, 2012). We also assume that $\|\boldsymbol{x}_i\|_2 = 1$ for all i. This decision is not limiting given that we are allowed to choose points within each subspace per the theorem statement; however, it is nonetheless not a requirement. This selection merely simplifies the exposition.

The subspace support $\Psi_{\boldsymbol{z}_i} \subset \{1, 2, \ldots, m\}$ is defined as the set of subspace indices whereby some arbitrary $\boldsymbol{z}_i$ has at least one associated nonzero element. We also let $C_k \subset \{1, 2, \ldots, n\}$ denote the set of dictionary column indices associated with $\mathcal{S}_k$.

Now consider the i-th data point $\boldsymbol{x}_i$ assumed to be drawn from $\mathcal{S}_k$ without loss of generality. If δ is sufficiently small, then after the first iteration (with unit weights) $k \in \Psi_{\boldsymbol{z}_i^{(1)}}$, i.e., the optimal subspace support will be a subset of the current subspace support. Moreover, it will be the case that

$$\sum_{j \in C_k} z_{ij}^{(1)} = 1 + \mathrm{O}(\delta). \tag{11}$$

These points are a consequence of the well-known stability of the ℓ_1 norm solution to small errors (Candes et al., 2006) (we may consider the variability $\mathrm{O}(\delta)$ within subspaces as a form of error) and the fact that $\Psi_{\boldsymbol{z}_i^{(1)}} = k$ when hypothetically $d_k = 1$ (the k-th subspace has dimension one) and the subspaces are unique/identifiable. In other words, the ℓ_1 norm will always locate the maximally sparse solution if the cardinality of such a solution is one, and will accurately approximate such solutions that are sufficiently close.

However with multiple non-degenerate points per subspace, it remains likely that a few extraneous subspaces are also selected with small compensatory coefficients. These coefficients will be of order $\mathrm{O}(\delta)$. To see this in the simplest terms, assume that we have adopted the data corruption model from Section 3.1 such that an additional identity $\boldsymbol{I}$ has been appended to $\boldsymbol{X}$, in which case there will always exist such feasible solutions.[3] Therefore any other feasible minimum ℓ_1 norm solution certainly could not involve larger order coefficients.

For the second iteration we must solve

$$\boldsymbol{z}_i^{(2)} \leftarrow \arg\min_{\boldsymbol{z}_i} \sum_j w_j^{(1)} |z_{ij}| \quad \text{s.t. } \boldsymbol{x}_i = \boldsymbol{X}_{\bar{i}} \boldsymbol{z}_i \tag{12}$$

where

$$w_j^{(1)} \triangleq \left[\boldsymbol{x}_j^\top \left(\alpha I + \boldsymbol{X}_{\bar{i}} \left| Z_i^{(1)} \right|^2 \boldsymbol{X}_{\bar{i}}^\top \right)^{-1} \boldsymbol{x}_j \right]^q . \tag{13}$$

When we choose $\alpha = \Theta(\delta)$, then for all j within C_k, it follows that

$$w_j^{(1)} = \Theta(1). \tag{14}$$

To see this, note that under the stated conditions $\boldsymbol{X}_{\bar{i}} \left| Z_i^{(1)} \right|^2 \boldsymbol{X}_{\bar{i}}^\top = \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top + \mathrm{O}(\delta)$. In contrast, for all $j \notin C_k$, we have

$$w_j^{(1)} = \Theta(\delta^{-q}). \tag{15}$$

This occurs since $\boldsymbol{x}_j = \boldsymbol{\mu}_{\bar{k}} + \mathrm{O}\left(\delta\right)$ with $\bar{k} \neq k$ for all $j \notin C_k$ and

$$\begin{aligned} &\boldsymbol{\mu}_{\bar{k}}^\top \left(\alpha I + \boldsymbol{X}_{\bar{i}} \left| Z_i^{(1)} \right|^2 \boldsymbol{X}_{\bar{i}}^\top \right)^{-1} \boldsymbol{\mu}_{\bar{k}} \\ &\quad = \boldsymbol{\mu}_{\bar{k}}^\top \left(\alpha I + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top + \mathrm{O}(\delta) \right)^{-1} \boldsymbol{\mu}_{\bar{k}} = \Theta(\delta^{-1}), \end{aligned} \tag{16}$$

noting that for the subspaces to be identifiable, there will always be distinct selections such that $\boldsymbol{\mu}_k \neq \boldsymbol{\mu}_{\bar{k}}$.

We will now compute the weighted ℓ_1 norm of a subspace optimal solution and compare it to any other non-subspace-optimal feasible solution, demonstrating that it is necessarily smaller. We can always choose points within $\mathcal{S}_k$, consistent with the above stipulations and arguments, such that a representation with $|z_{ij}| = \mathrm{O}\left(\delta^{-1}\right)$ for all j exists and is computable using the appropriate pseudo-inverse. Therefore a subspace optimal solution will always exist such that

$$\sum_j w_j^{(1)} |z_{ij}| \;=\; \sum_{j \in C_k} w_j^{(1)} |z_{ij}| \;=\; \mathrm{O}\left(\delta^{-1}\right). \tag{17}$$

[3] This assumption can be relaxed, but we avoid additional details here for the sake of simplicity.

Any other feasible solution, which again will be a basic feasible solution, must necessarily have nonzero coefficients in subspaces outside of $\mathcal{S}_k$. Let $\eta^-(\delta)$ denote the δ-dependent smallest value of $\sum_{j \notin C_k} |z_{ij}|$ for any non-subspace-optimal basic feasible solution. Similar to before, it follows that any such solution will satisfy

$$\begin{aligned} \sum_j w_j^{(1)} |z_{ij}| \; &> \; \sum_{j \notin C_k} w_j^{(1)} |z_{ij}| \\ &\equiv \; \Theta(\delta^{-q}) \sum_{j \notin C_k} |z_{ij}| \; > \; \Theta(\delta^{-q}) \eta^-(\delta). \end{aligned} \tag{18}$$

Because we may always pick some q sufficiently large such that the lower bound from (18) is larger than the upper bound from (17), a subspace optimal solution will be produced via (12). Similar analysis applies at subsequent updates to ensure that the solution will not change, completing the proof.

To summarize, the intuition behind the proof is relatively simple at a high level. If we assume that the data points are sufficiently clustered within each subspace, then at the first iteration the regular ℓ_1 norm solution selects the correct subspace support with its largest magnitude coefficients, while the second, weighted iteration prunes away small spurious coefficients associated with other subspaces. Note that the proposed algorithm accomplishes this without any particular tuning to such a clustered data model and thus similar results likely hold in many other cases as well. ■

References

T. Blumensath and M.E. Davies. Iterative thresholding for sparse approximations. *J. Fourier Analysis and Applications*, 14(5), 2008.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2765–2781, 2013.

David G Luenberger and Yinyu Ye. *Linear and nonlinear programming*, volume 2. Springer, 1984.

Bhaskar D Rao and Kenneth Kreutz-Delgado. An affine scaling methodology for best basis selection. *Signal Processing, IEEE Transactions on*, 47(1):187–200, 1999.

S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1832–1845, 2010.

M. Soltanolkotabi and E. Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40 (4):2195–2238, 2012.

Yi Wu and David P Wipf. Dual-space analysis of the sparse linear model. In *Advances in Neural Information Processing Systems*, pages 1745–1753, 2012.

Y. Yang, J. Feng, N. Jojic, J. Yang, and T. Huang. ℓ_0-sparse subspace clustering. In *Computer Vision–ECCV 2016*. Springer, 2016.