Robust Subspace Clustering via Thresholding Ridge Regression

Xi Peng\(^1\), Zhang Yi\(^2,*,\) and Huajin Tang\(^1,2,*,\)

\(^1\) Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632
\(^2\) College of Computer Science, Sichuan University, Chengdu 610065, P.R. China.

A*SGAI@gmail.com, zhangyi@scu.edu.cn, htang@i2r.a-star.edu.sg.

In the follow analysis, \textbf{lower-case bold letters} represent column vectors and \textbf{upper-case bold ones} represent matrices. \(A^T\) and \(A^{-1}\) denote the transpose and pseudo-inverse of the matrix \(A\), respectively. \(I\) denotes the identity matrix.

Table 1 summarizes some notations used throughout the material.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>the size of the dictionary</td>
</tr>
<tr>
<td>(m)</td>
<td>the dimensionality of sample</td>
</tr>
<tr>
<td>(r)</td>
<td>the rank of a given matrix</td>
</tr>
<tr>
<td>(x \in \mathbb{R}^m)</td>
<td>a data point</td>
</tr>
<tr>
<td>(c \in \mathbb{R}^n)</td>
<td>a given dictionary</td>
</tr>
<tr>
<td>(D = [d_1, d_2, \ldots, d_n])</td>
<td>the representation of (x) over (D)</td>
</tr>
</tbody>
</table>

\[D_x \in \{S|S \cap D_x \neq \emptyset\}\] and \(D_{-x} \in \{S|S \cap D_{-x} = \emptyset\}\), respectively. Hence, there are only two possibilities for the location of \(x\), i.e., in the intersection between \(S_{D_x}\) and \(S_{D_{-x}}\) (denoted by \(x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}\)), or in \(S_{D_x}\) except the intersection (denoted by \(x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}\)).

\[c^*_x\] and \(c^*_{-x}\) be the optimal solutions of

\[
\min \|c\|_p \text{ s.t. } x = Dc, \tag{1}
\]

over \(D_x\) and \(D_{-x}\), respectively. \(\| \cdot \|_p\) denotes the \(\ell_p\)-norm and \(p = \{1, 2, \infty\}\). Let \(x \neq 0\) be a data point in the union of subspaces \(S_D\) that is spanned by \(D = [D_x, D_{-x}]\), where \(D_x\) and \(D_{-x}\) consist of the intra-cluster and inter-cluster data points, respectively. Note that, noise and outlier could be regarded as a kind of inter-cluster data point of \(x\). Without loss of generality, let \(S_{D_x}\) and \(S_{D_{-x}}\) be the subspace spanned by \(D_x\) and \(D_{-x}\), respectively. Hence, there are only two possibilities for the location of \(x\), i.e., in the intersection between \(S_{D_x}\) and \(S_{D_{-x}}\) (denoted by \(x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}\)), or in \(S_{D_x}\) except the intersection (denoted by \(x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}\)).

Let \(c^*_x\) and \(c^*_{-x}\) be the optimal solutions of

\[
\min \|c\|_p \text{ s.t. } x = Dc, \tag{1}
\]

over \(D_x\) and \(D_{-x}\), respectively. \(\| \cdot \|_p\) denotes the \(\ell_p\)-norm and \(p = \{1, 2, \infty\}\). Let \(x \neq 0\) be a data point in the union of subspaces \(S_D\) that is spanned by \(D = [D_x, D_{-x}]\), where \(D_x\) and \(D_{-x}\) consist of the intra-cluster and inter-cluster data points, respectively. Note that, noise and outlier could be regarded as a kind of inter-cluster data point of \(x\). Without loss of generality, let \(S_{D_x}\) and \(S_{D_{-x}}\) be the subspace spanned by \(D_x\) and \(D_{-x}\), respectively. Hence, there are only two possibilities for the location of \(x\), i.e., in the intersection between \(S_{D_x}\) and \(S_{D_{-x}}\) (denoted by \(x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}\)), or in \(S_{D_x}\) except the intersection (denoted by \(x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}\)).

In the following analysis, Lemma 1 and Lemma 3 show \([c^*_x]_{r_x,1} > [c^*_{-x}]_{1,1}\) when \(x \in \{S|S = S_{D_x} \setminus S_{D_{-x}}\}\) \(x \in \{S|S = S_{D_x} \cap S_{D_{-x}}\}\), respectively. And Lemma 2 is a preliminary step toward Lemma 3.

\(\ast\)Corresponding Authors

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Proof. \((\Leftarrow)\) We prove the result using contradiction. Assume \(c_{r,x}^* \neq 0\), then
\[
\mathbf{x} - \mathbf{D}_x c_{r,x}^* = \mathbf{D}_{-x} c_{r,x}^*.
\]
(5)

Note that, the left side and the right side of (5) correspond a data point from \(S_{D_x}\) and \(S_{D_{-x}}\), respectively. Then, we must have
\[
\mathbf{x} = \mathbf{D}_x c_{r,x}^* + \mathbf{D}_{-x} \mathbf{z}_e,
\]
and
\[
\mathbf{x} = \mathbf{D}_x c_{r,x}^* + \mathbf{D}_{-x} \mathbf{z}_e.
\]
(6)

Clearly, \([c_{r,x}^* + \mathbf{z}_e]_p \leq ||c_{r,x}^*||_p + ||\mathbf{z}_e||_p\), are feasible solutions of (4) over \(\mathbf{D}_x, \mathbf{D}_{-x}\). According to the triangle inequality and the condition \(||\mathbf{z}_e||_p < ||\mathbf{z}_e||_p\), we have
\[
\left\|\begin{bmatrix} c_{r,x}^* + \mathbf{z}_e \\ 0 \end{bmatrix}\right\|_p < \left\|\begin{bmatrix} c_{r,x}^* \\ 0 \end{bmatrix}\right\|_p + \left\|\begin{bmatrix} \mathbf{z}_e \\ 0 \end{bmatrix}\right\|_p.
\]
(8)

From (7), we have \(||\mathbf{z}_e||_p \leq ||c_{r,x}^*||_p\) as \(||\mathbf{z}_e||_p\) is the optimal solution of (4) over \(\mathbf{D}_{-x}\). Then, \(\left\|\begin{bmatrix} c_{r,x}^* + \mathbf{z}_e \\ 0 \end{bmatrix}\right\|_p < \left\|\begin{bmatrix} c_{r,x}^* \leq 0 \\ c_{r,x}^* \leq 0 \end{bmatrix}\right\|_p\). It contradicts the fact that \(\left\|\begin{bmatrix} c_{r,x}^* \leq 0 \\ c_{r,x}^* \leq 0 \end{bmatrix}\right\|_p\) is the optimal solution of (4) over \(\mathbf{D}_x\).

\((\Rightarrow)\) We prove the result using contradiction. For a nonzero data point \(\mathbf{x} \in \{\mathbf{S}|\mathbf{S} = S_{D_x} \cap S_{D_{-x}}\}\), assume \(||\mathbf{z}_e||_p \geq ||\mathbf{z}_e||_p\). Thus, for the data point \(\mathbf{x} = \mathbf{x}\), it is possible that (4) will only choose the points from \(S_{D_{-x}}\) to represent \(\mathbf{x}\). This contradicts the assumption that \(c_{r,x}^* \neq 0\) and \(c_{r,x}^* = 0\).

This completes the proof.

\textbf{Definition 1} (The First Principal Angle). Let \(\xi\) be a Euclidean vector-space, and consider the two subspaces \(\mathcal{W}, \mathcal{V}\) with \(\dim(\mathcal{W}) := m \leq \mathcal{V} := \nu\). There exists a set of angles \(\{\theta_i\}_{i=1}^\nu\) called the principal angles of the first kind being defined as:
\[
\theta_{\min} := \min \left\{ \arccos \left( \frac{\mu^T \nu}{||\mu||_2 ||\nu||_2} \right) \right\},
\]
(9)

where \(\mu \in \mathcal{W}\) and \(\nu \in \mathcal{V}\).

\textbf{Lemma 3.} Consider the nonzero data point \(\mathbf{x} \in \{\mathbf{S}|\mathbf{S} = S_{D_x} \cap S_{D_{-x}}\}\), where \(S_{D_x}\) and \(S_{D_{-x}}\) denote the subspaces spanned by \(\mathbf{D}_x\) and \(\mathbf{D}_{-x}\), respectively. The dimensionality of \(S_{D_x}\) is \(r_x\) and that of \(S_{D_{-x}}\) is \(r_{-x}\). Let \(c^*\) be the optimal solution of
\[
\min ||c||_p \text{ s.t. } \mathbf{x} = \mathbf{D} c
\]
(10)

over \(\mathbf{D} = [\mathbf{D}_x, \mathbf{D}_{-x}]\) and \(c^* = \left[ c_{r,x}^* \right.\right]
\[
\begin{bmatrix}
\end{bmatrix}
\]
be partitioned according to the sets \(\mathbf{D}_x\) and \(\mathbf{D}_{-x}\). If
\[
\sigma_{\min}^2(\mathbf{D}_x) \geq r_x \cos \theta_{\min} \parallel \mathbf{D}_{-x} \parallel_{1,2},
\]
(11)

then \(c_{r,x}^* \geq c_{r_{-x}}^*\). Here, \(\sigma_{\min}^2(\mathbf{D}_x)\) is the smallest nonzero singular value of \(\mathbf{D}_x\), \(\theta_{\min}\) is the first principal angle between \(\mathbf{D}_x\) and \(\mathbf{D}_{-x}\), \(\parallel \mathbf{D}_{-x} \parallel_{1,2}\) is the largest \(\ell_2\)-norm of the columns of \(\mathbf{D}_{-x}\) and \(c_{r,x}^*\) denotes the \(r\)-th largest absolute value of the entries of \(c\).
It should be pointed out that the above proofs are motivated by the theoretical analysis in (Elhamifar and Vidal 2013), but they are different. (Elhamifar and Vidal 2013) provides the conditions under which SSC (\(\ell_1\)-norm) will only choose the intra-subspace data points to represent the input when data lies onto the union of independent subspaces or disjoint subspaces, while our theoretical analyses investigate the conditions under which \(\ell_p\)-norm-based coefficients with small value correspond to the projections over the errors even though the data come from dependent subspaces, where \(p = 1, 2, \infty\). Moreover, the most different point between these two works may be that, we theoretically show that the effect of errors could be eliminated by removing trivial coefficients in the projection space, whereas (Elhamifar and Vidal 2013) aims to prove the success of \(\ell_1\)-norm-based representation in the input data space.

References