Clustering with feature selection using alternating minimization
Application to computational biology

Cyprien Gilet, M. Deprez, J-B. Caillau and M. Barlaud

November 2017
Contents

1. Context
2. Our new method: K-sparse
3. Experiments on real single-cell biological data
4. Conclusion
Let $X \in \mathbb{R}^{m \times d}$ be the main database:

\[ d \text{ genes } \equiv \text{features} \]

\[
X = \begin{bmatrix}
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
\end{bmatrix} \\
\{ m \text{ cells} \}
\]

Objective: find $k \geq 2$ homogeneous clusters in $\mathbb{R}^{m \times d}$
Context

K-sparse clustering
Experiments on real single-cell biological data
Conclusion

$X = \begin{bmatrix}
\cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\end{bmatrix}$

$\begin{cases}
d \text{ genes } \equiv \text{ features} \\
m \text{ cells}
\end{cases}$

Difficulties:

- For genomic databases we have $d \gg m$
  $\rightarrow$ typically $d = 20,000$ genes and $m = 2,000$ cells

- A large set of features (genes) is noise
  $\rightarrow$ clustering errors
Solution:

Reduce clustering space dimensions
→ Project data into a low dimensional space

Main requirement:

Select only the most relevant features
→ Remove noisy features
State of the art for clustering

- **PCA kmeans** (frequently used in biology)
  → not the best space to discriminate clusters

- **Spectral clustering** (Ng et al 2001)
  → efficient but does not select significant features

- **Discriminative methods: Diffrac** (Bach, Harchaoui)
  → Does not select sparse features

- **Kernel methods: SIMLR** (Bach (2004), Wang et al (2017))
  → Does not select sparse features

- **Sparse kmeans clustering** **SPARCL** (D.Witten & R.Tibshirani 2010)
  → Select sparse features but does not project data onto Low-dimension space ($m \times m$).
Our new method: **K-sparse**

- **Project data** into a **low dimensional** space
- **Select** only the most **relevant features**
Let define the new space dimension $\bar{d}$

**Labels:** $Y \in \{0, 1\}^{m \times k}$

$Y = \begin{bmatrix} \ddots \end{bmatrix} \} m \text{ cells}$

$y_{ij} = \begin{cases} 1 & \text{if cell } i \in C_j \\ 0 & \text{otherwise} \end{cases}$

**Centroid matrix:** $\mu \in \mathbb{R}^{k \times \bar{d}}$

$\mu = \begin{bmatrix} \ddots \end{bmatrix} \} k \text{ centroids}$

**Sparse weights:** $W \in \mathbb{R}^{d \times \bar{d}}$, $W = \begin{bmatrix} \ddots \end{bmatrix} \} d \text{ genes}$
K-sparse : Optimization problem

\[
\min_{W, Y} \frac{1}{2} \| Y\mu - XW \|^2_F \quad \text{s.t.} \quad \| W \|_1 \leq \eta
\]

Alternating Minimization :

- **Y** labels are known, **compute sparse** **W** :
  - \( \rightarrow \) Remove noisy features (\( \ell_1 \) constraint).
  - \( \rightarrow \) gene \( i \) will be selected if \( \| W(i, :) \| > 0 \).

- **W** is known, **clustering step** :
  - Minimize the within-cluster sum of squares (WCSS) in \( XW \)
  - \( \rightarrow \) compute new clusters \( (Y \) and \( \mu) \) in \( XW \)
Compute sparse $W$

$$\min_{W} \frac{1}{2} \|Y\mu - XW\|_F^2 \quad \text{s.t.} \quad \|W\|_1 \leq \eta$$

**Algorithm 1** gradient-projection algorithm, $\ell^1$-constraint

1: **Input**: $X$, $Y$, $\mu$, $W_0$, $N$, $\gamma$, $\eta$
2: $W \leftarrow W_0$
3: for $n = 0, \ldots, N$ do
4: $V \leftarrow W - \gamma X^T(XW - Y\mu)$
5: $W \leftarrow P_{\eta}^1(V)$
6: end for
7: **Output**: $W$

**Theorem**: $\forall \gamma \in ]0, 2/\sigma_{\text{max}}^2(X)[$, with an exact projection on $\ell^1$-ball, we have convergence towards a solution.
Clustering step:

$$\min_Y \frac{1}{2} \| Y_\mu - XW \|_F^2$$

Compute new clusters with only significant features:

Minimize the within-cluster sum of squares (WCSS) on $XW$

$$[Y, \mu] = \text{kmeans}(XW)$$
K-sparse algorithm:

\[
\min_{W, Y} \frac{1}{2} \| Y\mu - XW \|_F^2 \quad \text{s.t.} \quad \| W \|_1 \leq \eta
\]

Algorithm 2 Alternating minimization.

1: **Input**: $X$, $Y_0$, $\mu_0$, $W_0$, $L$, $N$, $k$, $\gamma$, $\eta$
2: $Y \leftarrow Y_0$, $W \leftarrow W_0$, $\mu \leftarrow \mu_0$
3: **for** $l = 0, \ldots, L$ **do**
4: **for** $n = 0, \ldots, N$ **do**
5: \[ V \leftarrow W - \gamma X^T(XW - Y\mu) \]
6: \[ W \leftarrow P^1_\eta(V) \]
7: **end for**
8: $[Y, \mu] \leftarrow \text{kmeans}(XW, k)$
9: **end for**
10: **Output**: $Y$, $W$
3D animation

**Figure** – 3D animation
Single-cell RNA sequencing

- **New technology** elected method of the year in 2013 by Nature Methods.
- Provides a high resolution of cellular genes expression.
Biological databases with ground true labels:

- **Patel**:
  - $m = 430$ cells, $d = 5,948$ expressed genes, $k = 5$ Clusters

- **Usoskin**:
  - $m = 622$ cells, $d = 9,195$ expressed genes, $k = 4$ Clusters

- **Klein**:
  - $m = 2,717$ cells, $d = 10,322$ expressed genes, $k = 4$ Clusters

- **Zeisel**:
  - $m = 3,005$ cells, $d = 7,364$ expressed genes, $k = 9$ Clusters
Comparison between methods

**Clustering accuracy**

<table>
<thead>
<tr>
<th>Methods</th>
<th>PCA</th>
<th>Spectral</th>
<th>SIMLR</th>
<th>Sparcl</th>
<th>K-sparse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patel</td>
<td>76.04</td>
<td>80.46</td>
<td>97.21</td>
<td>94.18</td>
<td>98.37</td>
</tr>
<tr>
<td>Klein</td>
<td>68.50</td>
<td>63.31</td>
<td>99.12</td>
<td>65.11</td>
<td>99.26</td>
</tr>
<tr>
<td>Zeisel</td>
<td>39.60</td>
<td>59.30</td>
<td>71.85</td>
<td>65.23</td>
<td>83.42</td>
</tr>
<tr>
<td>Usoskin</td>
<td>54.82</td>
<td>60.13</td>
<td>76.37</td>
<td>57.24</td>
<td>95.98</td>
</tr>
</tbody>
</table>

K-sparse significantly improves other methods in terms of accuracy.
Comparison between methods

**Computational time**

<table>
<thead>
<tr>
<th>Methods</th>
<th>PCA</th>
<th>Spectral</th>
<th>SIMLR</th>
<th>Sparcl</th>
<th>K-sparse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patel</td>
<td>0.81</td>
<td>0.46</td>
<td>8.0</td>
<td>1,027</td>
<td>10.0</td>
</tr>
<tr>
<td>Klein</td>
<td>10.91</td>
<td>20.81</td>
<td>511.49</td>
<td>30,384</td>
<td>101.40</td>
</tr>
<tr>
<td>Zeisel</td>
<td>11</td>
<td>23</td>
<td>464</td>
<td>28,980</td>
<td>74</td>
</tr>
<tr>
<td>Usoskin</td>
<td>1.06</td>
<td>0.91</td>
<td>15.67</td>
<td>1,830</td>
<td>53.61</td>
</tr>
</tbody>
</table>

**Table** – Comparison between methods : Time (s).

Ksparse computational time is **linear with the number of cells** $m$  
→ scales up to large databases.
Frobenius norm evolution:

\[
\min_{W, Y} \frac{1}{2} \| Y\mu - XW \|_F^2 \quad \text{s.t.} \quad \| W \|_1 \leq \eta
\]

![Graph showing the Frobenius norm evolution with different methods over the number of loops.](image-url)
Accuracy evolution:

![Graphs showing accuracy evolution for different methods.](image)
gene \( i \) will be selected if \( \|W(i,:}\| > 0 \)
To conclude:

**Conclusion**

- **New clustering method (k-sparse):**
  - Projects data into a lower space dimensions
  - Selects relevant features which:
    - Bring together the cells of a same cluster
    - Discriminate clusters

- **Experiments on scRNA-seq databases:**
  - Ksparse significantly improves other methods in terms of accuracy.
  - Computational time linear with the number of cells $m$
    - Scales up to large datasets.
To conclude:

**Current work**

- Application to very large scRNA-seq datasets containing $m = 68,000$ and $m = 1,000,000$ cells.

- Biological evaluations on other real genomic databases.
References


Find our article in arXiv.org: 1711.02974

Thank you for your attention