Clustering with sparse feature selection using alternating minimization and an exact projection-gradient splitting method

Cyprien Gilet gilet@i3s.unice.fr
I3S, Univ. Côte d'Azur & CNRS, F-06900 Sophia Antipolis, France.
Marie Deprez deprez@ipmc.cnrs.fr
IPMC, Univ. Côte d'Azur & CNRS, F-06560 Sophia Antipolis, France.
Jean-Baptiste Caillau caillau@unice.fr
LJAD, Univ. Côte d'Azur & CNRS/Inria, F-06108 Nice, France.
Michel Barlaud barlaud@i3s.unice.fr
I3S, Univ. Côte d'Azur & CNRS, F-06900 Sophia Antipolis, France.

Abstract

This paper deals with unsupervised clustering with feature selection on public single-cell RNA-seq datasets. The problem is to estimate both labels Y and a sparse projection matrix W. To address this combinatorial non-convex problem maintaining a strict control on the sparsity of W, we follow an alternating minimization of the Frobenius norm criterion. We provide a new efficient algorithm named `K-sparse` which alternates a `k-means` step and a projection-gradient step. The projection-gradient step is a method of splitting type, with exact projection on the ℓ^1 ball to promote sparsity. The convergence of the gradient-projection step is addressed,and a preliminary analysis of the alternating minimization is made. The Frobenius norm criterion converges as the number of iterates in Algorithm `K-sparse` goes to infinity. Experiments on scRNA-seq datasets show that our method outperforms `PCA k-means`, `spectral clustering`, `SIMLR`, and `Sparcl` methods, and achieves a selection of genes. Complexity of `K-sparse` is linear with the number of samples (cells) so it scales up to large datasets.

1 Introduction

As a relatively new technology elected "method of the year" in 2013 by the Journal *Nature Methods*, single-cell RNA-sequencing represents a computational and analytical challenge in terms of machine learning applications in biology. Single-cell RNA-sequencing (scRNA-seq) measures gene expression at single cell level thus providing a full characterization of the transcriptional landscape of individual cells. One of the main use and challenge for scRNA-seq is the characterization of cell types by selecting informative features/genes from high dimension expression profiles which requires robust and accurate clustering methods. State of the art in computational biology uses methods such as `PCA k-means`, kernel based spectral clustering. Some public datasets are now available with ground truth label annotations to test and improve those methods. However, clustering in high dimension suffers from the curse of dimensionality: as dimensions increase, vectors become indiscernible and the predictive power of the aforementioned methods is drastically reduced. In order to overcome this issue, a popular approach for high-dimensional data is to perform *Principal Component Analysis* (PCA) prior to clustering using `k-means`. This approach is however difficult to justify in general. An alternative approach proposed in [Ding and Li(2007)] is to combine clustering and dimension reduction by means of *Linear Discriminant Analysis* (LDA). The heuristic used in [Ding and Li(2007)] is based on alternating minimization, which consists in iteratively computing a projection subspace by LDA, using the labels at the current iteration and then running `k-means` on the projection of the data onto the subspace. Another efficient approach is the spectral clustering where main tools are graph Laplacian matrices [Von Luxburg(2007)]. However PCA, LDA and Spectral do not provide sparsity. A popular approach for selecting sparse features in regression is the *Least Absolute Shrinkage and Selection Operator* (LASSO) formulation [Tibshirani(1996)]. The LASSO uses the ℓ^1 norm instead of ℓ^0 [Candès(2008)], as an added penalty term (λ being a hyper-parameter tuning sparsity). A main issue is that optimizing the values of the Lagrangian parameter λ [Witten and Tibshirani(2010)] is computationally expensive [Mairal and Yu(2012)]. Note that all these methods [Ding and Li(2007), Von Luxburg(2007), Witten and Tibshirani(2010)] require

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

a k-means heuristic to retrieve the labels. We propose an alternating scheme, K-sparse, combining k-means clustering, dimension reduction and feature selection using an ℓ^1 sparsity inducing constraint.

2 Constrained unsupervised classification

Let $X(\neq 0)$ be the $m \times d$ matrix made of m line samples $x_1, \dots, x_m$ belonging to the d-dimensional space of features. Let $Y \in \{0,1\}^{m \times k}$ be the label matrix where $k \geqslant 2$ is the number of clusters. Each line of Y has exactly one nonzero element equal to one, $y_{ij} = 1$ indicating that the sample x_i belongs to the j-th cluster. Let $W \in \mathbb{R}^{d \times \bar{d}}$ be the projection matrix, $\bar{d} \ll d$, and let μ be the $k \times \bar{d}$ matrix of centroids: Each line $\mu(j,:)$, $j = 1, \dots, k$, is the model for all samples x_i belonging to the j-th cluster ($y_{ij} = 1$). For a given W, the clustering criterion can be cast as

$$\frac{1}{2}\|Y\mu - XW\|_F^2 \to \min, \tag{1}$$

where $\|.\|_F$ is the Frobenius norm induced by the Euclidean structure on $m \times \bar{d}$ matrices. In contrast with the Lagrangian LASSO formulation, we want to have a direct control on the value of the ℓ^1 constraint, so we constrain W according to $\|W\|_1 \leqslant \eta$ ($\eta > 0$), where $\|.\|_1$ is the ℓ^1 norm of the vectorized $d \times \bar{d}$ matrix of weights. The problem is to estimate both labels and centroids (Y, μ), together with the sparse projection matrix W. A naive joint minimization of (1) in (Y, μ, W) under the previous ℓ^1 constraint and the nonconvex constraints on Y

$$y_{ij} \in \{0,1\}, \quad i = 1, \dots, m, \quad j = 1, \dots, \bar{d}, \quad \sum_{j=1} y_{ij} = 1, \quad i = 1, \dots, m, \tag{2}$$

admits trivial solutions consisting in $k-1$ empty clusters and $W = 0$. To overcome this issue, we follow an alternating scheme as in [Ding and Li(2007), Witten and Tibshirani(2010)]. (Another option would be to devise a global convexification; see, *e.g.*, [Flammarion et al.(2016)Flammarion, Palaniappan, and Bach].) The first convex subproblem finds the best projection from dimension d to dimension $\bar{d}$ for a given clustering, while the second one is the standard k-means on the projected data.

Problem 1 *For a fixed a clustering (Y, μ) (and a given $\eta > 0$),*

$$\frac{1}{2}\|Y\mu - XW\|_F^2 \to \min, \quad \|W\|_1 \leqslant \eta.$$

Problem 2 *For a fixed projection matrix W,*

$$\frac{1}{2}\|Y\mu - XW\|_F^2 \to \min, \quad \mu \in \mathbf{R}^{k \times \bar{d}} \textit{and constraints defined by (2).}$$

Gradient-projection splitting method. To solve Problem 1, we use a gradient-projection approach. This method belongs to the class of splitting methods [Combettes and Pesquet(2011), Combettes and Wajs(2005), Mosci et al.(2010)Mosci, Rosasco, Santoro, Verri, and Villa, Sra et al.(2012)Sra, Nowozin, and Wright] and is designed to solve minimization problems of the form

$$\varphi(W) \to \min, \quad W \in C, \tag{3}$$

using separately the convexity properties of the function φ on one hand, and of the convex set C on the other. We use the following forward-backward scheme to generate a sequence of iterates:

$$V_n := W_n + \gamma_n \nabla\varphi(W_n), \quad W_{n+1} := P_C(V_n) + \varepsilon_n, \tag{4}$$

where P_C denotes the projection on the convex set C (a subset of some Euclidean space). Under standard assumptions on the sequence of gradient steps $(\gamma_n)_n$, and on the sequence of projection errors $(\varepsilon_n)_n$, convergence holds (see, *e.g.*, [Bauschke and Combettes(2011)]).

Theorem 1 *Assume that (3) has a solution. Assume that φ is convex, differentiable, and that $\nabla\varphi$ is β-Lipschitz, $\beta > 0$. Assume finally that C is convex and that*

$$\sum_n |\varepsilon_n| < \infty, \quad \inf_n \gamma_n > 0, \quad \sup_n \gamma_n < 2/\beta.$$

Then the sequence of iterates of the forward-backward scheme (4) converges, whatever the initialization. If moreover $(\varepsilon_n)_n = 0$ (exact projections), there exists a rank N and a positive constant K such that for $n \geqslant N$

$$\varphi(W_n) - \inf_C \varphi \leqslant K/n. \tag{5}$$

In our case, $\nabla\varphi$ is Lipschitz with constant $\sigma^2_{\max}(X)$ since $\nabla\varphi(W) = X^T(XW - Y\mu)$, and the following holds.

Corollary 1 *For any fixed step $\gamma \in (0, 2/\sigma^2_{\max}(X))$, the forward-backward scheme applied to the Problem 1 with an exact projection on ℓ^1 balls converges with a linear rate towards a solution, and the estimate (5) holds.*

The $O(1/n)$ convergence rate of the algorithm can be speeded up to $O(1/n^2)$ using a FISTA step [Beck and Teboulle(2009)]. In practice we use a modified version [Chambolle and Dossal(2015)] which ensures convergence of the iterates.

Exact ℓ^1 projection. We denote by $P^1_\eta(W)$ the (reshaped as a $d \times \bar{d}$ matrix) projection of the vectorized matrix $W(:)$. An important asset of the method is that it takes advantage of the availability of efficient methods [Condat(2016)] to compute the ℓ^1 projection. For $\eta > 0$, denote $B^1(0,\eta)$ the closed ℓ_1 ball of radius η in the space $\mathbb{R}^{d\bar{d}}$ centered at the origin, and Δ_η the simplex $\{w \in \mathbb{R}^{d\bar{d}} \mid w_1 + \cdots + w_{d\bar{d}} = 1,\ w_1 \geqslant 0, \dots, w_{d\bar{d}} \geqslant 0\}$. Let $w \in \mathbb{R}^{d\bar{d}}$, and let v denote the projection on Δ_η of $(|w_1|, \dots, |w_{d\bar{d}}|)$. It is well known that the projection of w on $B^1(0,\eta)$ is

$$(\varepsilon_1(v_1), \dots, \varepsilon_{kd}(v_{d\bar{d}})), \quad \varepsilon_j := \operatorname{sign}(w_j), \quad j = 1, \dots, d\bar{d}, \tag{6}$$

and the fast method described in [Condat(2016)] is used to compute v. The complexity of the gradient part is $O(m \times d)$, while complexity of the projection is $O(d \times \bar{d})$. Thus our algorithm is scalable and is tractable for large data sets.

Clustering algorithm. The resulting alternating minimization is described by Algorithm 1. Labels Y are for instance initialized by spectral clustering on X, while the `k-means` computation relies on standard methods such as `k-means++`[Arthur and Vassilvitski(2007)]. Similarly to the approaches advocated in [Bach and Harchaoui(2008), De la Torre and Kanade(2006), Ding and Li(2007), Witten and Tibshirani(2010)], our method involves non-convex `k-means` optimization for which convergence towards local minimizers only can be proved [Bottou and Bengio(1995)]. In practice, we use `k-means++` with several replicates to improve each clustering step. We note for further research that there have been recent attempts to convexify `k-means`. As each step of the alternating minimization scheme decreases the Frobenius norm in (1), which is nonnegative, one has

Proposition 1 *The norm $\|Y\mu - XW\|_F$ converges as the number of iterates L in Algorithm 1 goes to infinity.*

Algorithm 1 Alternating minimization clustering.

```
Input: $X, Y_0, \mu_0, W_0, L, N, k, \gamma, \eta$
$Y \leftarrow Y_0, \quad \mu \leftarrow \mu_0, \quad W \leftarrow W_0$
for $l = 0, \dots, L$ do
    for $n = 0, \dots, N$ do
        $V \leftarrow W - \gamma X^T(XW - Y\mu)$
        $W \leftarrow P^1_\eta(V)$
    end for
    $[Y, \mu] \leftarrow$ k-means$(XW, k)$
end for
Output: $Y, W$
```

3 Application to single cell RNA-seq datasets

To compare our results we report accuracy, *Adjusted Rank Index* (ARI) [Lawrence and Phipps(1985)], *Normalized Mutual Information* (NMI) and tSNE for 2D-visual evaluation for the five following methods: `PCA k-means`, spectral clustering [Von Luxburg(2007), Ng et al.(2002)Ng, Jordan, and Weiss], `SIMLR` [Wang et al.(2017)Wang, Zhu, Pierson, Ramazzotti, and Batzoglou], and `Sparcl` (we have used the *R* software provided by [Witten and Tibshirani(2010)]), and our method. Moreover we report the processing times using a 2.5 GHz *Macbook Pro* with an $i7$ processor. We compare algorithms on four scRNA-seq datasets: **Patel** et al. [Patel(2014)] characterizes intra-tumoral heterogeneity and redundant transcriptional pattern in glioblastoma tumors, **Klein** et al. [Kiselev(2017), Klein(2015)] characterized the transcriptome of $2,717$ cells (*Mouse Embryonic Stem Cells*, mESCs) across various conditions), **Zeisel** et al. [Kiselev(2017), Zeisel(2015)] collected 3005 cells from the primary somatosensory cortex (S1) and the hippocampal CA1 region, **Usoskin** et al. [Usoskin(2014)] collected 622 cells from the mouse dorsal root ganglion.

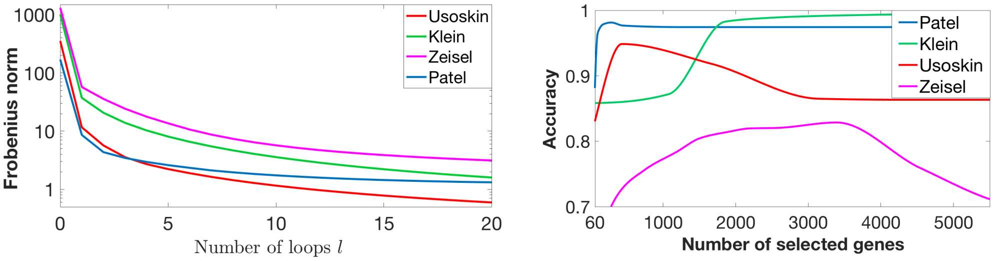

Figure 1: Left: the decay of the Frobenius norm for the four data sets versus the number of loops of the alternating minimization scheme emphasizes the fast and smooth convergence of our algorithm. Right: these results show that a minimum number of genes is required to get the best possible clustering accuracy.

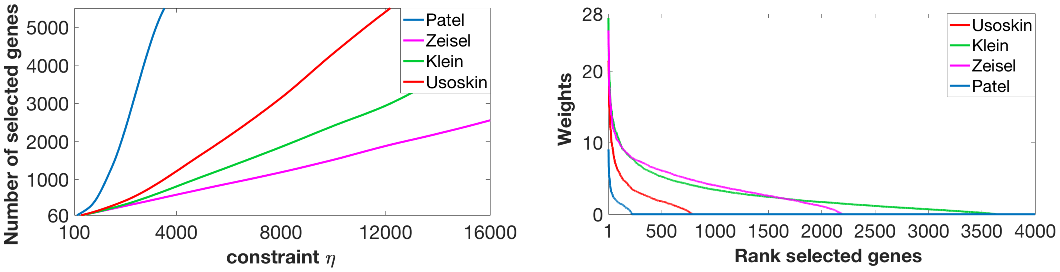

Figure 2: Left: The The evolution of the number of selected genes versus the constraint η is a smooth monotonous function. Right: ranked weight $\|W(j,:)\|$ of selected genes.

Table 1: Ksparse outperforms SIMLR in term of accuracy on Usoskin, and Zeisel, Ksparse outperforms SIMLR in term of computational time on large data bases Klein and Zeisel

	Accuracy (%)					**Time** (s)		
Methods:	PCA	SIMLR [2]	Sparcl [3]	K-sparse		PCA	SIMLR [2]	K-sparse
Patel (430 cells, $k=5$)	76.04	97.21	94.18	**98.37**		**0.81**	8	10
Klein (2,717 cells, $k=4$)	68.50	99.12	65.11	**99.26**		**11**	511	101
Zeisel (3,005 cells, $k=9$)	39.60	71.85	65.23	**83.42**		**11**	464	74
Usoskin (622 cells, $k=4$)	54.82	76.37	57.24	**95.98**		**1.06**	15	53

4 Conclusion

The proposed method significantly improves the results of `Sparcl` and `SIMLR` in terms of accuracy. We also note that `K-sparse` and `Sparcl` provide built-in feature selection while `SIMLR` requires supplementary processing. Complexity of `K-sparse` is linear with the number of samples, so it scales up to large datasets, contrary to `SIMLR` and `Sparcl` (Computing the kernel for `SIMLR` or optimizing the values of the Lagrangian parameter using permutations for `Sparcl` are computationally expensive).

References

[Arthur and Vassilvitski(2007)] D. Arthur and S. Vassilvitski. k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007.

[Bach and Harchaoui(2008)] F. R. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Pro-*

cessing Systems 20, pages 49–56. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3269-diffrac-a-discriminative-and-flexible-framework-for-clustering.pdf.

[Bauschke and Combettes(2011)] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011.

[Beck and Teboulle(2009)] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[Bottou and Bengio(1995)] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 585–592. MIT Press, 1995.

[Candès(2008)] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Acad Sciences Paris*, 346(1):589–592, 2008.

[Chambolle and Dossal(2015)] A. Chambolle and C. Dossal. On the convergence of the iterates of "fista". *Journal of Optimization Theory and Applications, Springer Verlag*, (166), 2015.

[Combettes and Pesquet(2011)] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[Combettes and Wajs(2005)] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[Condat(2016)] L. Condat. Fast projection onto the simplex and the l1 ball. *Mathematical Programming Series A*, 158(1): 575–585, 2016.

[De la Torre and Kanade(2006)] F. De la Torre and T. Kanade. Discriminative cluster analysis. *ICML 06 Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, USA*, 2006.

[Ding and Li(2007)] C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 521–528, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273562. URL http://doi.acm.org/10.1145/1273496.1273562.

[Flammarion et al.(2016)Flammarion, Palaniappan, and Bach] N. Flammarion, B. Palaniappan, and F. Bach. Robust discriminative clustering with sparse regularizers. *arXiv preprint arXiv:1608.08052*, 2016.

[Kiselev(2017)] V. Y. e. a. Kiselev. Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods*, 2017.

[Klein(2015)] A. M. e. a. Klein. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 2015.

[Lawrence and Phipps(1985)] H. Lawrence and A. Phipps. Comparing partitions. *Journal of Classification*, 1985.

[Mairal and Yu(2012)] J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 353–360, 2012.

[Mosci et al.(2010)Mosci, Rosasco, Santoro, Verri, and Villa] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa. Solving structured sparsity regularization with proximal methods. In *Machine Learning and Knowledge Discovery in Databases*, pages 418–433. Springer, 2010.

[Ng et al.(2002)Ng, Jordan, and Weiss] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002.

[Patel(2014)] A. P. e. a. Patel. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science 344*, 2014.

[Sra et al.(2012)Sra, Nowozin, and Wright] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. MIT Press, 2012.

[Tibshirani(1996)] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Usoskin(2014)] D. e. a. Usoskin. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature Neuroscience*, 2014.

[Von Luxburg(2007)] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007.

[Wang et al.(2017)Wang, Zhu, Pierson, Ramazzotti, and Batzoglou] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods*, (14), 2017.

[Witten and Tibshirani(2010)] D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[Zeisel(2015)] A. e. a. Zeisel. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 2015.