Journal of Machine Learning Research 18 (2017) 1-50 Submitted 8/16; Revised 5/17; Published 8/17

Robust Discriminative Clustering with Sparse Regularizers

Nicolas Flammarion[1] NICOLAS.FLAMMARION@ENS.FR
Balamurugan Palaniappan[2] PALANIAPPAN@TELECOM-PARISTECH.FR
Francis Bach[1] FRANCIS.BACH@ENS.FR
[1] *INRIA*
Département d'Informatique de l'ENS, École Normale Supérieure, CNRS, PSL Research University
Paris, France.

[2] *Signal, Statistique et Apprentissage (S^2A) Group, IDS Department, LTCI*
Telecom-ParisTech
Paris, France.

Editor: Sathiya Keerthi

Abstract

Clustering high-dimensional data often requires some form of dimensionality reduction, where clustered variables are separated from "noise-looking" variables. We cast this problem as finding a low-dimensional projection of the data which is well-clustered. This yields a one-dimensional projection in the simplest situation with two clusters, and extends naturally to a multi-label scenario for more than two clusters. In this paper, (a) we first show that this joint clustering and dimension reduction formulation is equivalent to previously proposed discriminative clustering frameworks, thus leading to convex relaxations of the problem; (b) we propose a novel sparse extension, which is still cast as a convex relaxation and allows estimation in higher dimensions; (c) we propose a natural extension for the multi-label scenario; (d) we provide a new theoretical analysis of the performance of these formulations with a simple probabilistic model, leading to scalings over the form $d = O(\sqrt{n})$ for the affine invariant case and $d = O(n)$ for the sparse case, where n is the number of examples and d the ambient dimension; and finally, (e) we propose an efficient iterative algorithm with running-time complexity proportional to $O(nd^2)$, improving on earlier algorithms for discriminative clustering with the square loss, which had quadratic complexity in the number of examples.

1. Introduction

Clustering is an important and commonly used pre-processing tool in many machine learning applications, with classical algorithms such as K-means (MacQueen, 1967), linkage algorithms (Gower and Ross, 1969) or spectral clustering (Ng et al., 2002). In high dimensions, these unsupervised learning algorithms typically have problems identifying the underlying optimal discrete nature of the data; for example, they are quickly perturbed by adding a few noisy dimensions. Clustering high-dimensional data thus requires some form of dimensionality reduction, where clustered variables are separated from non-informative "noise-looking" (e.g., Gaussian) variables.

Several frameworks aim at linearly separating noise from signal, that is finding projections of the data that extracts the signal and removes the noise. They differ in the ways signals and noise are defined. A line of work that dates back to projection pursuit (Friedman and Stuetzle, 1981) and independent component analysis (Hyvärinen et al., 2004) defines the noise as Gaussian while the signal is non-Gaussian (Blanchard et al., 2006; Le Roux and Bach, 2013; Diederichs et al., 2013). In this work, we follow De la Torre and Kanade (2006); Ding and Li (2007), along the alternative route where one defines the signal as being clustered while the noise is any non-clustered variable. In the simplest situation with two clusters, we may project the data into a one-dimensional subspace. Given a data matrix $X \in \mathbb{R}^{n\times d}$ composed of n d-dimensional points, the goal is to find a direction $w \in \mathbb{R}^d$ such that $Xw \in \mathbb{R}^n$ is well-clustered, e.g., by K-means. This is equivalent to identifying both a direction to project, represented as $w \in \mathbb{R}^d$ and the labeling $y \in \{-1, 1\}^n$ that represents the partition into two clusters.

Most existing formulations are non-convex and typically perform a form of alternating optimization (De la Torre and Kanade, 2006; Ding and Li, 2007), where given $y \in \{-1, 1\}^n$, the projection w is found by linear discriminant analysis (or any binary classification method), and given the projection w, the clustering is obtained by thresholding Xw or running K-means on Xw. As shown in Section 2, this alternating minimization procedure happens to be equivalent to maximizing the (centered) correlation between $y \in \{-1, 1\}^n$ and the projection $Xw \in \mathbb{R}^d$, that is

$$\max_{w\in\mathbb{R}^d, y\in\{-1,1\}^n} \frac{(y^\top \Pi_n X w)^2}{\|\Pi_n y\|_2^2 \, \|\Pi_n X w\|_2^2},$$

where $\Pi_n = I_n - \frac{1}{n} 1_n 1_n^\top$ is the usual centering projection matrix (with $1_n \in \mathbb{R}^n$ being the vector of all ones, and I_n the $n \times n$ identity matrix). This correlation is equal to one when the projection is perfectly clustered (independently of the number of elements per cluster). Existing methods are alternating minimization algorithms with no theoretical guarantees.

In this paper, we relate this formulation to discriminative clustering formulations (Xu et al., 2004; Bach and Harchaoui, 2007), which consider the problem

$$\min_{v\in\mathbb{R}^d,\ b\in\mathbb{R},\ y\in\{-1,1\}^n} \frac{1}{n}\|y - Xv - b1_n\|_2^2, \tag{1}$$

with the intuition of finding labels y which are easy to predict by an affine function of the data. In particular, we show that given the relationship between the number of positive labels and negative labels (i.e., the squared difference between the respective number of elements), these two problems are equivalent, and hence discriminative clustering explicitly performs joint dimension reduction and clustering.

While the discriminative framework is based on convex relaxations and has led to interesting developments and applications (Zhang et al., 2009; Li et al., 2009; Joulin et al., 2010a,b; Wang et al., 2010; Niu et al., 2013; Huang et al., 2015), it has several shortcomings when used with the square loss: (a) the running-time complexity of the semi-definite formulations is at least quadratic in n, and typically much more, (b) no theoretical analysis has ever been performed, (c) no convex sparse extension has been proposed to handle data with many irrelevant dimensions, (d) balancing of the clusters remains an issue, as it typically adds an extra hyperparameter which may be hard to set. In this paper, we focus on addressing these concerns.

When there are more than two clusters, one considers either the *multi-label* or the *multi-class* settings. The multi-class problem assumes that the data are clustered into distinct classes, i.e., a single class per observation, whereas the multi-label problem assumes the data share different labels, i.e., multiple labels per observation. We show in this work that discriminative clustering framework extends more naturally to multi-label scenarios and that this extension has the same convex relaxation.

A summary of the contributions of this paper follows:

- In Section 2, we relate discriminative clustering with the square loss to a joint clustering and dimension reduction formulation. The proposed formulation takes care of the balancing hyperparameter implicitly.
- We propose in Section 3, a novel sparse extension to discriminative clustering and show that it can still be cast through a convex relaxation.
- When there are more than two clusters, we extend naturally the sparse formulation to a multi-label scenario in Section 4.
- We then proceed to provide a theoretical analysis of the proposed formulations with a simple probabilistic model in Section 5, which effectively leads to scalings over the form $d = O(\sqrt{n})$ for the affine invariant case and $d = O(n)$ for the 1-sparse case.
- Finally, we propose in Section 6 efficient iterative algorithms with running-time complexity for each step equal to $O(nd^2)$, the first to be linear in the number of observations n for discriminative clustering with the square loss.

Throughout this paper we assume that $X \in \mathbb{R}^{n\times d}$ is *centered*, a common pre-processing step in unsupervised (and supervised) learning. This implies that $X^\top 1_n = 0$ and $\Pi_n X = X$.

2. Joint Dimension Reduction and Clustering

In this section, we focus on the single binary label case, where we first study the usual non-convex formulation, before deriving convex relaxations based on semi-definite programming. Some of the following results are already known in the literature; however, we state them here for completeness.

2.1 Non-convex formulation

Following De la Torre and Kanade (2006); Ding and Li (2007); Ye et al. (2008), we consider a cost function which depends on $y \in \{-1,1\}^n$ and $w \in \mathbb{R}^d$, which is such that alternating optimization is exactly (a) running K-means with two clusters on Xw to obtain y given w (when we say "running K-means", we mean solving the vector quantization problem exactly), and (b) performing linear discriminant analysis to obtain w given y.

Proposition 1 (Joint clustering and dimension reduction for two clusters) *Given $X \in \mathbb{R}^{n\times d}$ such that $X^\top 1_n = 0$ and X has rank d, consider the optimization problem*

$$\max_{w\in\mathbb{R}^d, y\in\{-1,1\}^n} \frac{(y^\top X w)^2}{\|\Pi_n y\|_2^2 \, \|Xw\|_2^2}. \tag{2}$$

Given y, the optimal w is obtained as $w = (X^\top X)^{-1} X^\top y$, while given w, the optimal y is obtained by running K-means on Xw.

This equivalence might be straightforward, however it has not been precisely stated in the literature to the best of our knowledge.

Proof Given y, we need to optimize the Rayleigh quotient $\frac{w^\top X^\top y y^\top X w}{w^\top X^\top X w}$ with a rank-one matrix in the numerator, which leads to $w = (X^\top X)^{-1} X^\top y$. Given w, we show in Appendix A, that the averaged distortion measure of K-means once the means have been optimized is exactly equal to $(y^\top X w)^2 / \|\Pi_n y\|_2^2$. ■

Algorithm. The proposition above leads to an alternating optimization algorithm. Note that K-means in one dimension may be run *exactly* in $O(n \log n)$ (Bellman, 1973). After having optimized with respect to w in Eq. (2), we then need to maximize with respect to y the function $\frac{y^\top X (X^\top X)^{-1} X^\top y}{\|\Pi_n y\|_2^2}$, which happens to be exactly performing K-means on the whitened data (which is now in high dimension and not in 1 dimension). At first, it seems that dimension reduction is *simply* equivalent to whitening the data and performing K-means; while this is a formally correct statement, the resulting K-means problem is not easy to solve as the clustered dimension is hidden in noise; for example, algorithms such as K-means++ (Arthur and Vassilvitskii, 2007), which have a multiplicative theoretical guarantee on the final distortion measure, are not provably effective here because the minimal final distortion is not small (since the clusters are corrupted by some noisy dimensions), and the multiplicative guarantee is then meaningless.

2.2 Convex relaxation and discriminative clustering

The discriminative clustering formulation in Eq. (1) may be optimized for any $y \in \{-1, 1\}^n$ in closed form with respect to b as $b = \frac{1_n^\top (y - Xv)}{n} = \frac{1_n^\top y}{n}$ since X is centered. Substituting b in Eq. (1) leads us to

$$\min_{v \in \mathbb{R}^d} \frac{1}{n} \|\Pi_n y - Xv\|_2^2 = \frac{1}{n} \|\Pi_n y\|_2^2 - \max_{w \in \mathbb{R}^d} \frac{(y^\top X w)^2}{\|Xw\|_2^2}, \tag{3}$$

where v is obtained from any solution w as $v = w \frac{y^\top X w}{\|Xw\|_2^2}$. Thus, given

$$\frac{(y^\top 1_n)^2}{n^2} = \frac{1}{n^2} \big(\#\{i, y_i = 1\} - \#\{i, y_i = -1\} \big)^2 = \alpha \in [0, 1], \tag{4}$$

which characterizes the asymmetry between clusters and with $\|\Pi_n y\|^2 = n(1 - \alpha)$, we obtain from Eq. (3), an equivalent formulation to Eq. (2) (with the added constraint) as

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \|\Pi_n y - Xv\|_2^2 \ \text{ such that } \ \frac{(y^\top 1_n)^2}{n^2} = \alpha. \tag{5}$$

This is exactly equivalent to a discriminative clustering formulation with the square loss (Bach and Harchaoui, 2007) with an explicit cluster balance constraint. Consequently we have formally established that the discriminative clustering formulation in Eq. (5) is related to the joint clustering and dimension reduction formulation in Eq. (2). Following Bach and Harchaoui (2007), we may optimize Eq. (5) in closed form with respect to v as $v = (X^\top X)^{-1} X^\top y$. Substituting v in Eq. (5) leads us to

$$\min_{y \in \{-1,1\}^n} \frac{1}{n} y^\top \big(\Pi_n - X(X^\top X)^{-1} X^\top\big) y \ \text{ such that } \ \frac{(y^\top 1_n)^2}{n^2} = \alpha. \tag{6}$$

This combinatorial optimization problem is NP-hard in general (Karp, 1972; Garey et al., 1976). Hence in practice, it is classical to consider the following convex relaxation of Eq. (6) (Luo et al., 2010). For any admissible $y \in \{-1,+1\}^n$, the matrix $Y = yy^\top \in \mathbb{R}^{n \times n}$ is a rank-one symmetric positive semi-definite matrix with unit diagonal entries and conversely any such Y may be written in the form $Y = yy^\top$ such that y is admissible for Eq. (6). Moreover by rewriting Eq. (6) as

$$\min_{y \in \{-1,1\}^n} \frac{1}{n} \operatorname{tr} yy^\top \big(\Pi_n - X(X^\top X)^{-1} X^\top\big) \ \text{ such that } \ \frac{1_n^\top (yy^\top) 1_n}{n^2} = \alpha,$$

we see that the objective and constraints are linear in the matrix $Y = yy^\top$ and Eq. (6) is equivalent to

$$\min_{Y \succcurlyeq 0,\ \operatorname{rank}(Y)=1,\ \operatorname{diag}(Y)=1} \frac{1}{n} \operatorname{tr} Y \big(\Pi_n - X(X^\top X)^{-1} X^\top\big) \text{ such that } \frac{1_n^\top Y 1_n}{n^2} = \alpha.$$

Then dropping the non-convex rank constraint leads us to the following classical convex relaxation:

$$\min_{Y \succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{1}{n} \operatorname{tr} Y \big(\Pi_n - X(X^\top X)^{-1} X^\top\big) \text{ such that } \frac{1_n^\top Y 1_n}{n^2} = \alpha. \tag{7}$$

This is the standard (unregularized) formulation, which is cast as a semi-definite program. The complexity of interior-point methods is $O(n^7)$, but efficient algorithms in $O(n^2)$ for such problems have been developed due to the relationship with the max-cut problem (Journée et al., 2010; Wen et al., 2012). We note that convex relaxation techniques are also used for semi-supervised methods (De Bie and Cristianini, 2003).

Given the solution Y, one may traditionally obtain a candidate $y \in \{-1,1\}^n$ by running K-means on the largest eigenvector of Y or by sampling (Goemans and Williamson, 1995). In this paper, we show in Section 5 that it may be advantageous to consider the first two eigenvectors.

2.3 Unsuccessful full convex relaxation

The formulation in Eq. (7) imposes an extra parameter α that characterises the cluster imbalance. It is tempting to find a direct relaxation of Eq. (2). It turns out to lead to a trivial relaxation, which we outline below.

When optimizing Eq. (2) with respect to w, we obtain the following optimization problem

$$\max_{y \in \{-1,1\}^n} \frac{y^\top X (X^\top X)^{-1} X^\top y}{y^\top \Pi_n y},$$

leading to a quasi-convex relaxation as

$$\max_{Y \succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{\operatorname{tr} YX(X^\top X)^{-1}X^\top}{\operatorname{tr} \Pi_n Y},$$

whose solution is found by solving a sequence of convex problems (Boyd and Vandenberghe, 2004, Section 4.2.5). As shown in Appendix B, this may be exactly reformulated as a single convex problem:

$$\max_{M \succcurlyeq 0,\ \operatorname{diag}(M)=1+\frac{1^\top M1}{n^2}} \operatorname{tr} MX(X^\top X)^{-1}X^\top.$$

Unfortunately, this relaxation always leads to trivial solutions, and we thus need to consider the relaxation in Eq. (7) for several values of $\alpha = 1_n^\top Y 1_n / n^2$ (and then the non-convex algorithm can be run from the rounded solution of the convex problem, using Eq. (2) as a final objective). Alternatively, we may solve the following *penalized* problem for several values of $\nu \geqslant 0$:

$$\min_{Y \succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{1}{n} \operatorname{tr} Y \big(\Pi_n - X(X^\top X)^{-1}X^\top\big) + \frac{\nu}{n^2} 1_n^\top Y 1_n. \tag{8}$$

For $\nu = 0$, $Y = 1_n 1_n^\top$ is always a trivial solution. As outlined in our theoretical section and as observed in our experiments, it is sufficient to consider $\nu \in [0, 1]$.

By convex duality (Borwein and Lewis, 2000, Sec. 4.3), both constrained relaxation in Eq. (7) and penalized relaxation in Eq. (8) are formally equivalent for specific choices of constraint parameter α and penalization parameter ν. We will see in Section 6 that the formulation in Eq. (8) is more suitable for algorithmic design (Bach et al., 2012).

2.4 Equivalent relaxations

Optimizing Eq. (5) with respect to v in closed form as in Section 2.2 is feasible with no regularizer or with a quadratic regularizer. However, if one needs to add more complex regularizers, we need a different relaxation. Therefore, we now propose a new formulation of the discriminative clustering framework. We start from the penalized version of Eq. (5),

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \|\Pi_n y - Xv\|_2^2 + \nu \frac{(y^\top 1_n)^2}{n^2}, \tag{9}$$

which we expand as:

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \operatorname{tr} \Pi_n yy^\top - \frac{2}{n} \operatorname{tr} Xvy^\top + \frac{1}{n} \operatorname{tr} X^\top Xvv^\top + \nu \frac{(y^\top 1_n)^2}{n^2}, \tag{10}$$

and relax as, using $Y = yy^\top$, $P = yv^\top$ and $V = vv^\top$,

$$\min_{V,P,Y} \frac{1}{n} \operatorname{tr} \Pi_n Y - \frac{2}{n} \operatorname{tr} P^\top X + \frac{1}{n} \operatorname{tr} X^\top XV + \nu \frac{1_n^\top Y 1_n}{n^2} \text{ s.t. } \begin{pmatrix} Y & P \\ P^\top & V \end{pmatrix} \succcurlyeq 0,\ \operatorname{diag}(Y) = 1. \tag{11}$$

When optimizing Eq. (11) with respect to V and P, we get exactly Eq. (8). Indeed, the optimum is attained for $V = (X^\top X)^{-1} X^\top Y X (X^\top X)^{-1}$ and $P = YX(X^\top X)^{-1}$ as shown in Appendix C.1. Therefore, the convex relaxation in Eq. (11) is equivalent to Eq. (8).

However, we get an interesting behavior when optimizing Eq. (11) with respect to P and Y also in closed form. For $\nu = 1$, we obtain, as shown in Appendix C.2, the following closed form expressions:

$$\begin{aligned} Y &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} XVX^\top \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} \\ P &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} XV, \end{aligned}$$

leading to the problem:

$$\min_{V \succcurlyeq 0} \ 1 - \frac{2}{n}\sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} + \frac{1}{n}\operatorname{tr}(VX^\top X). \tag{12}$$

The formulation above in Eq. (12) is interesting for several reasons: (a) it is formulated as an optimization problem in $V \in \mathbb{R}^{d\times d}$, which will lead to algorithms whose running time will depend on n linearly (see Section 6), (b) it allows for easy adding of regularizers (see Section 3), which may be formulated as convex functions of $V = vv^\top$. At first sight this seems to be valid only for $\nu = 1$. However we now propose a reformulation which can handle all possible $\nu \in [0, 1)$ through a simple data augmentation.

Reformulation for any ν When $\nu \in [0, 1)$, we may reformulate the objective function in Eq. (9) as follows:

$$\begin{aligned} \frac{1}{n}\|\Pi_n y - Xv\|_2^2 + \nu\frac{(y^\top 1_n)^2}{n^2} &= \frac{1}{n}\|\Pi_n y - Xv + \nu\frac{y^\top 1_n}{n}1_n\|_2^2 - \big(\nu\frac{y^\top 1_n}{n}\big)^2 + \nu\big(\frac{y^\top 1_n}{n}\big)^2 \\ &= \frac{1}{n}\|y - Xv - (1-\nu)\frac{y^\top 1_n}{n}1_n\|_2^2 + \frac{\nu}{1-\nu}\big((1-\nu)\frac{y^\top 1_n}{n}\big)^2 \\ &= \min_{b\in\mathbb{R}} \frac{1}{n}\|y - Xv - b1_n\|_2^2 + \frac{\nu}{1-\nu}b^2, \end{aligned} \tag{13}$$

since $\frac{1}{n}\|y - Xv - b1_n\|_2^2 + \frac{\nu}{1-\nu}b^2$ can be optimized in closed form with respect to b as $b = (1-\nu)\frac{y^\top 1_n}{n}$. Note that the weighted imbalance ratio $(1-\nu)\frac{y^\top 1_n}{n}$ is made as an optimization variable in Eq. (13). Thus we have the following reformulation

$$\begin{aligned} &\min_{v\in\mathbb{R}^d,\ y\in\{-1,1\}^n} \frac{1}{n}\|\Pi_n y - Xv\|_2^2 + \nu\frac{(y^\top 1_n)^2}{n^2} \\ = &\min_{v\in\mathbb{R}^d,\ b\in\mathbb{R},\ y\in\{-1,1\}^n} \frac{1}{n}\|y - Xv - b1_n\|_2^2 + \frac{\nu}{1-\nu}b^2, \end{aligned} \tag{14}$$

which is a non-centered penalized formulation on a higher-dimensional problem in the variable $\binom{v}{b} \in \mathbb{R}^{d+1}$. In the rest of the paper, we will focus on the case $\nu = 1$ for ease of exposition. This enables the use of the formulation in Eq. (12), which is easier to optimize. It is worth noting that this is not an algorithmic restriction. Of course any problem with $\nu \in [0, 1)$ can be treated with equal ease by adding a constant term and a quadratic regularizer.

3. Regularization

There are several natural possibilities. We consider norms Ω such that $\Omega(w)^2 = \Gamma(ww^\top)$ for a certain convex function Γ; all norms have that form (Bach et al., 2012, Proposition 5.1). When $\nu = 1$, Eq. (12) then becomes

$$\max_{V \succcurlyeq 0} \frac{2}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - \frac{1}{n} \operatorname{tr}(VX^\top X) - \Gamma(V). \tag{15}$$

The quadratic regularizers $\Gamma(V) = \operatorname{tr} \Lambda V$ have already been tackled by Bach and Harchaoui (2007). They consider the regularized version of problem in Eq. (3)

$$\min_{v \in \mathbb{R}^d} \frac{1}{n} \|\Pi_n y - Xv\|_2^2 + v^\top \Lambda v, \tag{16}$$

optimize in closed form with respect to v as $v = (X^\top X + n\Lambda)^{-1} X^\top y$. Substituting v in Eq. (16) leads them to

$$\min_{Y \succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{1}{n} \operatorname{tr} Y \big(\Pi_n - X(X^\top X + n\Lambda)^{-1} X\big).$$

In this paper, we propose a novel sparse extension to discriminative clustering framework with the square loss. Specifically we formulate a non-trivial sparse regularizer which is a combination of weighted squared ℓ_1-norm and ℓ_2-norm. It leads to

$$\Gamma(V) = \operatorname{tr}[\operatorname{Diag}(a) V \operatorname{Diag}(a)] + \|\operatorname{Diag}(c) V \operatorname{Diag}(c)\|_1, \tag{17}$$

such that $\Gamma(vv^\top) = \sum_{i=1}^d a_i^2 v_i^2 + \big(\sum_{i=1}^d c_i |v_i|\big)^2$. This allows to treat all situations simultaneously, with $\nu = 1$ or with $\nu \in [0, 1)$. To be more precise, when $\nu \in [0, 1)$, we can consider in Eq. (14), a problem of size $d + 1$ with a design matrix $[X, 1_n] \in \mathbb{R}^{n \times (d+1)}$, a direction of projection $\binom{v}{b} \in \mathbb{R}^{d+1}$ and different weights for the last variable with $a_{d+1} = \frac{\nu}{1-\nu}$ and $c_{d+1} = 0$.

Note that the sparse regularizers on V introduced in this paper are significantly different when compared to the sparse regularizers on variable v in Eq. (3), for example, considered by Wang et al. (2013). A straightforward sparse regularizer on v in Eq. (3), despite leading to a sparse projection, does not yield natural generalizations of the discriminative clustering framework in terms of theory or algorithms.

In our analysis and experiments for the balanced clusters (when $\nu = 1$), the sparse regularization $\Gamma(\cdot) = \lambda \| \cdot \|_1$, for $\lambda \in \mathbb{R}$ will often be considered. This is equivalent to setting $a = 0_d$ and $c = \sqrt{\lambda} 1_d$ in Eq. (17). The problem in Eq. (15) then becomes

$$\max_{V \succcurlyeq 0} \frac{2}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - \frac{1}{n} \operatorname{tr}(VX^\top X) - \lambda \|V\|_1. \tag{18}$$

The sparse regularizers considered in this paper have a significant algorithmic appeal for certain applications in computer vision (Bojanowski et al., 2013; Alayrac et al., 2016), audio processing (Lajugie et al., 2016) and natural language processing (Grave, 2014). They also lead to robust cluster recovery under minor assumptions as will be illustrated on a simple example in Section 5. The practical benefits of the sparse regularizers will be further demonstrated using empirical evaluation on synthetic and real data sets in Section 7.

4. Extension to Multiple Labels

The discussion so far has focussed on two clusters. Yet it is key in practice to tackle more clusters. It is worth noting that the discrete formulations in Eq. (2) and Eq. (5) extend directly to more than two clusters. However two different extensions of the initial problems Eq. (2) or Eq. (5) are conceivable. They lead to problems with different constraints on different optimization domains and, consequently, to different relaxations. We discuss these possibilities next.

One extension is the *multi-class* case. The multi-class problem which is dealt with by Bach and Harchaoui (2007) assumes that the data are clustered into K classes and the various partitions of the data points into clusters are represented by the K-class indicator matrices $y \in \{0,1\}^{n\times K}$ such that $y1_K = 1_n$. The constraint $y1_K = 1_n$ ensures that one data point belongs to only one cluster. However as discussed by Bach and Harchaoui (2007), by letting $Y = yy^\top$, it is possible to lift these K-class indicator matrices into the outer convex approximations $\mathcal{C}_K = \{Y \in \mathbb{R}^{n\times n} : Y = Y^\top, \mathrm{diag}(Y) = 1_n, Y \succcurlyeq 0, Y \preccurlyeq \frac{1}{K}1_n1_n^\top\}$ (Frieze and Jerrum, 1995), which is different for all values of K. Note that letting $K = 2$ corresponds to the previous sections.

In this paper, we consider a different novel extension for discriminative clustering to the *multi-label* case. The multi-label problem assumes that the data share k labels and the data-label membership is represented by matrices $y \in \{-1,+1\}^{n\times k}$. In other words, the multi-class problem embeds the data in the extreme points of a simplex, while the multi-label problem does so in the extreme points of the hypercube.

The discriminative clustering formulation of the multi-label problem is

$$\min_{v\in\mathbb{R}^{d\times k},\ y\in\{-1,1\}^{n\times k}} \frac{1}{n}\|\Pi_n y - Xv\|_F^2, \tag{19}$$

where the Frobenius norm is defined for any vector or rectangular matrix as $\|A\|_F^2 = \mathrm{tr}\, AA^\top = \mathrm{tr}\, A^\top A$. Letting $k = 1$ here corresponds to the previous sections. The discrete ensemble of matrices $y \in \{-1,+1\}^{n\times k}$ can be naturally lifted into $\mathcal{D}_k = \{Y \in \mathbb{R}^{n\times n} : Y = Y^\top, \mathrm{diag}(Y) = k1_n, Y \succcurlyeq 0\}$, since $\mathrm{diag}(Y) = \mathrm{diag}(yy^\top) = \sum_{i=1}^k y_{i,i}^2 = k$. As the optimization problems in Eq. (7) and Eq. (8) have linear objective functions, we can change the variable from Y to $\tilde{Y} = Y/k$ to change the constraint $\mathrm{diag}(Y) = k1_n$ to $\mathrm{diag}(\tilde{Y}) = 1_n$ without changing the optimizer of the problem. Thus the problems can be solved over the relaxed domain $\mathcal{D} = \{Y \in \mathbb{R}^{n\times n} : Y = Y^\top, \mathrm{diag}(Y) = 1_n, Y \succcurlyeq 0\}$ which is independent of k.

Note that the domain $\mathcal{D}$ is similar to that considered in the problems in Eq. (8) and Eq. (11) and these convex relaxations are the same regardless of the value of k. Hence the multi-label problem is a more natural extension of the discriminative framework, with a slight change in how the labels y are recovered from the solution Y (we discuss this in Section 5.3).

5. Theoretical Analysis

In this section, we provide the first theoretical analysis for the discriminative clustering framework with the square loss. We start with the 2-clusters situation: the non-sparse case is considered first and analysis is provided for both balanced and imbalanced clusters. Our study for the sparse case currently only provides results for the simple 1-sparse solution. However, the analysis also yields valuable insights on the scaling between n and d. We then derive results for multi-label situation.

For ease of analysis, we consider the constrained problem in Eq. (7), the penalized problem in Eq. (8) or their equivalent relaxations in Eq. (12) or Eq. (18) under various scenarios, for which we use the same proof technique. We first try to characterize the low-rank solutions of these relaxations and then show in certain simple situations the uniqueness of such solutions, which are then non-ambiguously found by convex optimization. Perturbation arguments could extend these results by weakening our assumptions but are not within the scope of this paper, and hence we do not investigate them further in this section.

5.1 Analysis for two clusters: non-sparse problems

In this section, we consider several noise models for the problem, either adding irrelevant dimensions or perturbing the label vector with noise. We consider these separately for simplicity, but they could also be combined (with little extra insight).

5.1.1 Irrelevant dimensions

We consider an "ideal" design matrix $X \in \mathbb{R}^{n\times d}$ such that there exists a direction v along which the projection Xv is perfectly clustered into two distinct real values c_1 and c_2. Since Eq. (2) is invariant by affine transformation, we can rotate the design matrix X to have $X = [y, Z]$ with $y \in \{-1, 1\}^n$, which is clustered into $+1$ or -1 along the direction $v = \left(\begin{smallmatrix} 1 \\ 0_{d-1} \end{smallmatrix}\right)$. Then after being centered, the design matrix is written as $X = [\Pi_n y, Z]$ with $Z = [z_1, \dots, z_{d-1}] \in \mathbb{R}^{n\times(d-1)}$. The columns of Z represent the noisy irrelevant dimensions added on top of the signal y.

5.1.2 Balanced problem

When the problem is well balanced ($y^\top 1_n = 0$), y is already centered and $\Pi_n y = y$. Thus the design matrix is represented as $X = [y, Z]$. We consider here the penalized formulation in Eq. (8) with $\nu = 1$ which is the only scenario where we are able to provide a theoretical analysis.

Let us assume that the columns $(z_i)_{i=1,\dots,d-1}$ of Z are i.i.d. with symmetric distribution z, with $\mathbb{E}z = \mathbb{E}z^3 = 0$ and such that $\|z\|_\infty$ is almost surely bounded by $R \geq 0$. We denote by $\mathbb{E}z^2 = m$ its second moment and by $\mathbb{E}z^4/(\mathbb{E}z^2)^2 = \beta$ its (unnormalized) kurtosis.

Surprisingly the clustered vector y happens to generate a solution $yy^\top$ of the relaxation Eq. (8) for all possible values of Z (see Lemma 11 in Appendix D.2). However the problem in Eq. (8) should have a *unique* solution in order to always recover the correct assignment y. Unfortunately the semidefinite constraint $Y \succcurlyeq 0$ of the relaxation makes the second-order information arduous to study. Due to this reason, we consider the other equivalent relaxation in Eq. (12) for which $V_* = vv^\top$ is also solution with $v \propto (X^\top X)^{-1} X^\top y$ (see Lemma 12 in Appendix D.3). Fortunately the semidefinite constraint $V \succcurlyeq 0$ of the problem in Eq. (12) may be ignored since the second-order information in V of the objective function already provides unicity for the unconstrained problem. Hence we are able to ensure the uniqueness of the solution with high probability.

Proposition 2 *Let us assume $d \geq 3$, $\beta > 1$ and $m^2 \geq \frac{\beta-3}{2(d+\beta-4)}$:*
(a) If $n \geq d^2R^4\frac{1+(d+\beta)m^2}{m^2(\beta-1)}$, V_* is the unique solution of the problem in Eq. (12) with high probability.
(b) If $n \geq \frac{d^2R^4}{\min\{m^2(\beta-1), 2m^2, 2m\}}$, v is the principal eigenvector of any solution of the problem in Eq. (12) with high probability.

Let us make the following observations:

– **Proof technique**: The proof relies on a computation of the Hessian of $f(V) = \frac{2}{n}\sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \frac{1}{n}\operatorname{tr} X^\top XV$ which is the objective function in Eq. (12). We first derive the expectation of $\nabla^2 f(V)$ with respect to the distribution of X. By the law of large numbers, it amounts to have n going to infinity in $\nabla^2 f(V)$. Then we expand the spectrum of this operator $\mathbb{E}\nabla^2 f(V)$ to lower-bound its smallest eigenvalue. Finally we use concentration theory on matrices, following Tropp (2012), to bound the Hessian $\nabla^2 f(V)$ for finite n.

– **Effect of kurtosis**: We remind that $\beta \geqslant 1$, with equality if and only if z follows a Rademacher law ($\mathbb{P}(z = +1) = \mathbb{P}(z = -1) = 1/2$). Thus, if the noisy dimensions are clustered, then unsurprisingly, our guarantee is meaningless. Note that the constant β behaves like a distance of the distribution z to the Rademacher distribution. Moreover, $\beta = 3$ if z follows a standard normal distribution.

– **Scaling between d and n**: If the noisy variables are not evenly clustered between the same clusters $\{\pm 1\}$ (i.e., $\beta > 1$), we recover a rank-one solution as long as $n = O(d^3)$; while, as long as $n = O(d^2)$, the solution is not unique but its principal eigenvector recovers the correct clustering. Moreover, as explained in the proof, its spectrum would be very spiky.

– The assumption $m^2 \geq \frac{\beta-3}{2(d+\beta-4)}$ is generally satisfied for large dimensions. Note that $m^2 d$ is the total variance of the irrelevant dimensions, and when it is small, i.e., when $m^2 \leq \frac{\beta-3}{2(d+\beta-4)}$, the problem is particularly simple, and we can also show that V_* is the unique solution of the problem in Eq. (12) with high probability if $n \geq \frac{d^2R^4}{m^2}$. Finally, note that for sub-Gaussian distributions (where $\beta \leq 3$), the extra constraint is vacuous, while for super-Gaussian distributions (where $\beta \geq 3$), this extra constraint only appears for small m.

– This result provides the first guarantee for discriminative clustering. However similar theoretical results have been derived for K-means by Ostrovsky et al. (2006) and Gaussian mixtures by Kalai et al. (2010); Moitra and Valiant (2010), where separation conditions between the two clusters are derived, under which the clustering problem is efficiently solved. It would be of great interest to relate these separation conditions to our condition on n and d but this is outside the scope of this work.

5.1.3 Noise robustness for the one dimensional balanced problem

We assume now that the data are one-dimensional and are perturbed by some noise $\varepsilon \in \mathbb{R}^n$ such that $X = y + \varepsilon$ with $y \in \{-1, 1\}^n$. The solution of the relaxation in Eq. (8) recovers the correct y in this setting only when each component of y and $y + \varepsilon$ have the same sign (this is shown in Appendix D.5). This result comes out naturally from the information on whether the signs of y and $y + \varepsilon$ are the same or not. Further if we assume that y and ε are independent, this condition is equivalent to $\|\varepsilon\|_\infty < 1$ almost surely.

5.1.4 Unbalanced problem

When the clusters are imbalanced ($y^\top 1_n \neq 0$), the natural rank-one candidates $Y_* = yy^\top$ and $V_* = vv^\top$ are no longer solutions of the relaxations in Eq. (8) (for $\nu = 1$) and Eq. (12), as proved in Appendix D.6. Nevertheless we are able to characterize some solutions of the penalized relaxation in Eq. (8) for $\nu = 0$.

Lemma 3 *For $\nu = 0$ and for any non-negative $a, b \in \mathbb{R}$ such that $a + b = 1$,*

$$Y = ayy^\top + b1_n 1_n^\top$$

is solution of the penalized relaxation in Eq. (8).

Hence any eigenvector of this solution Y would be supported by the directions y and 1_n. Moreover when the value $\alpha_* = (\frac{1_n^\top y}{n})^2$ is known, it turns out that we can characterize some solutions of the constrained relaxation in Eq. (7), as stated in the following lemma.

Lemma 4 *For $\alpha \geq \alpha_*$,*

$$Y = \frac{1 - \alpha}{1 - \alpha_*} yy^\top + \Big(1 - \frac{1 - \alpha}{1 - \alpha_*}\Big) 1_n 1_n^\top$$

is a rank-2 solution of the constrained relaxation in Eq. (7) with constraint parameter α.

The eigenvectors of Y enable to recover y for $\alpha_* \leq \alpha < 1$. We conjecture (and checked empirically) that this rank-2 solution is unique under similar regimes to those considered for the balanced case. The proof would be more involved since, when $\nu \neq 1$, we are not able to derive an equivalent problem in V for the penalized relaxation in Eq. (8) similar to Eq. (12) for the balanced case. We also note that Lemmas 3 and 4 will be direct consequences of Lemma 8 in Section 5.3.

Thus Y being rank-2, one should really be careful and consider the first two eigenvectors when recovering y from a solution Y. This can be done by rounding the principal eigenvector of $\Pi_n Y \Pi_n = \frac{1-\alpha}{1-\alpha_*} \Pi_n y (\Pi_n y)^\top$ as discussed in the following lemma.

Lemma 5 *Let y_{ev} be the principal eigenvector of $\Pi_n Y \Pi_n$ where Y is defined in Lemma 4, then*

$$\mathrm{sign}(y_{ev}) = y.$$

Proof By definition of Y, $y_{ev} = \sqrt{\frac{1-\alpha}{1-\alpha_*}} \Pi_n y$ thus $\mathrm{sign}(y_{ev}) = \mathrm{sign}(\Pi_n y)$ and since $\alpha \leq 1$ then $\mathrm{sign}(\Pi_n y) = \mathrm{sign}(y - \sqrt{\alpha} 1_n) = y$. ■

In practice, contrary to the standard procedure, we should, for any ν, solve the penalized relaxation in Eq. (8) and then do K-means on the principal eigenvector of the centered solution $\Pi_n Y \Pi_n$ instead of the solution Y to recover the correct y. This procedure is followed in our experiments on real-world data in Section 7.2.

5.2 Analysis for two clusters: one-sparse problems

We assume here that the direction of projection v (such that $Xv = y$) is l-sparse (by l-sparse we mean $\|v\|_0 = l$). The ℓ_1-norm regularized problem in Eq. (18) is no longer invariant by affine transformation and we cannot consider that $X = [y, Z]$ without loss of generality. Yet the relaxation Eq. (18) seems experimentally to only have rank-one solutions for the simple $l = 1$ situation. Hence we are able to derive some theoretical analysis only for this case. It is worth noting the $l = 1$ case is simple since it can be solved in $O(d)$ by using K-means separately on all dimensions and ranking them. Nonetheless the proposed scaling also holds in practice for $l \geqslant 1$ (see Figure 1b).

Thereby we consider data $X = [y, Z]$ with $y \in \{-1, 1\}^n$ and $Z \in \mathbb{R}^{n \times (d-1)}$ which are clustered in the direction $v = [1, 0, \ldots, 0]^\top \in \mathbb{R}^d$. When adding a ℓ_1-penalty, the initial problem in Eq. (5) for $\alpha = 0$ is

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \|y - Xv\|_2^2 + \lambda \|v\|_1^2. \tag{20}$$

When optimizing in v this problem is close to the Lasso (Tibshirani, 1996) and a solution is known to be $v_i^* = (y^\top y + n\lambda)^{-1} y^\top y = \frac{1}{1+\lambda}$, $\forall i \in J$ and $v_i^* = 0$, $\forall i \in \{1, 2, \ldots, d\} \setminus J$, where J is the support of v^*. The candidate $V_* = v^* v^{*\top}$ is still a solution of the relaxation in Eq. (18) (see Lemma 15 in Appendix E.1) and we will investigate under which conditions on X this solution is unique. Let us assume as before $(z_i)_{i=1,\ldots,d}$ are i.i.d. with distribution z symmetric with $\mathbb{E}z = \mathbb{E}z^3 = 0$, and denote by $\mathbb{E}z^2 = m$ and $\mathbb{E}z^4 / (\mathbb{E}z^2)^2 = \beta$. We also assume that $\|z\|_\infty$ is almost surely bounded by $0 \leq R \leq 1$. We are able to ensure the uniqueness of the solution with high-probability.

Proposition 6 *Let us assume $d \geq 3$.*
(a) If $n \geq dR^2 \frac{1+(d+\beta)m^2}{m^2(\beta-1)}$, V_* is the unique solution of the problem Eq. (12) with high probability.
(b) If $n \geq \frac{dR^2}{m^2(\beta-1)}$, v^* is the principal eigenvector of any solution of the problem Eq. (12) with high probability.

The proof technique is very similar to the one of Proposition 2. With the function $g(V) = \frac{2}{n} \sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \lambda \|V\|_1 - \frac{1}{n} \operatorname{tr} X^\top X V$, we can certify that g will decrease around the solution V_* by analyzing the eigenvalues of its Hessian.

The rank-one solution V_* is recovered by the principal eigenvector of the solution of the relaxation Eq. (18) as long as $n = O(d)$. Thus we have a much better scaling when compared to the non-sparse setting where $n = O(d^2)$. We also conjecture a scaling of order $n = O(ld)$ for a projection in a l-sparse direction (see Figure 1b for empirical results).

The proposition does not state any particular value for the regularizer parameter λ. This makes sense since the proposition only holds for the simple situation when $l = 1$. We propose to use $\lambda = 1/\sqrt{n}$ by analogy with the Lasso.

5.3 Analysis for the multi-label extension

In this section, the signals share k labels which are corrupted by some extra noisy dimensions. We assume the centered design matrix to be $X = [\Pi_n y, Z]$ where $y \in \{-1, +1\}^{n\times k}$ and $Z \in \mathbb{R}^{n\times(d-k)}$. We also assume that y is full-rank[1]. We denote by $y = [y_1, \ldots, y_k]$ and $\alpha_i = \left(\frac{y_i^\top 1_n}{n}\right)^2$ for $i = 1, \cdots, k$. We consider the discrete constrained problem

$$\min_{v\in\mathbb{R}^{d\times k},\ y\in\{-1,1\}^{n\times k}} \frac{1}{n}\|\Pi_n y - Xv\|_F^2 \text{ such that } \frac{1_n^\top y y^\top 1_n}{n^2} = \alpha^2, \tag{21}$$

and the discrete penalized problem for $\nu = 0$

$$\min_{v\in\mathbb{R}^{d\times k},\ y\in\{-1,1\}^{n\times k}} \frac{1}{n}\|\Pi_n y - Xv\|_F^2. \tag{22}$$

As explained in Section 4, these two discrete problems admit the same relaxations in Eq. (7) and Eq. (8) we have studied for one label. We now investigate when the solution of the problems in Eq. (21) and in Eq. (22) generate solutions of the relaxations in Eq. (7) and Eq. (8).

By analogy with Lemma 3, we want to characterize the solutions of these relaxations which are supported by the constant vector 1_n and the labels $(y_1, \ldots, y_k)$. Their general form is $Y = \tilde{y} A \tilde{y}^\top$ where $A \in \mathbb{R}^{k\times k}$ is symmetric semi-definite positive and $\tilde{y} = [1_n, y]$. However the initial y is easily recovered from the solution Y only when A is diagonal. To that end the following lemma derives some condition under which the only matrix A such that the corresponding Y satisfies the constraint of the relaxations in Eq. (7) and Eq. (8) is diagonal.

Lemma 7 *The solutions of the matrix equation* $\operatorname{diag}(\tilde{y} A \tilde{y}^\top) = 1_n$ *with unknown variable A are diagonal if and only if the family $\{1_n, (y_i)_{1\le i\le k}, (y_i \odot y_j)_{1\le i<j\le k}\}$ is linearly independent where we denoted by $\odot$ the Hadamard (i.e., pointwise) product between matrices.*

In this way we are able to characterize the solution of relaxations in Eq. (7) and Eq. (8) with the following result:

Lemma 8 *Let us assume that the family $\{1_n, (y_i)_{1\le i\le k}, (y_i \odot y_j)_{1\le i<j\le k}\}$ is linearly independent. If $\alpha \ge \alpha_{\min} = \min_{1\le i\le k}\{\alpha_i\}$ with $(\alpha_i)_{1\le i\le k}$ defined above Eq. (21), the solutions of the constrained relaxation in Eq. (7) supported by the vectors $(1_n, y_1, \cdots, y_k)$ are of the form:*

$$Y = a_0^2 1_n 1_n^\top + \sum_{i=1}^k a_i^2 y_i y_i^\top,$$

where $(a_i)_{0\le i\le k}$ satisfies $\sum_{i=0}^k a_i^2 = 1$ and $a_0^2 + \sum_{i=1}^k a_i^2 \alpha_i = \alpha$.

[1] This assumption is fairly reasonable since the probability of a matrix whose entries are i.i.d. Rademacher random variables to be singular is conjectured to be $1/2 + o(1)$ (Bourgain et al., 2010).

Moreover the solutions of the penalized relaxation in Eq. (8) for $\nu = 0$ which are supported by the vectors $(1_n, y_1, \cdots, y_k)$ are of the form:

$$Y = a_0^2 1_n 1_n^\top + \sum_{i=1}^k a_i^2 y_i y_i^\top,$$

where $(a_i)_{0 \le i \le k}$ satisfies $\sum_{i=0}^k a_i^2 = 1$.

In the *multi-label* case, some combinations of the constant matrix $1_n 1_n^\top$ and the rank-one matrices $y_i y_i^\top$ are solutions of constrained or penalized relaxations. Furthermore, under some assumptions on the labels $(y_i)_{1 \le i \le k}$, these combinations are the only solutions which are supported by the vectors $(1_n, y_1, \cdots, y_k)$. And we conjecture (and checked empirically) that under assumptions similar to those made for the balanced one-label case, all the solutions of the relaxation are supported by the family $(1_n, y_1, \cdots, y_k)$ and consequently share the same form as in Lemma 8. Thus the eigenvector of the solution Y would be in the span of the directions $(1_n, y_1, \cdots, y_k)$.

Let us consider an eigenvalue decomposition of $Y = FF^\top = \sum_{i=0}^k \lambda_i e_i e_i^\top$ and denote by $M = [a_0 1_n, a_1 y_1, \cdots, a_k y_k]$ where $(a_i)_{0 \le i \le k}$ are defined in Lemma 8. Since $MM^\top = FF^\top$, there is an orthogonal transformation R such that $FR = M$. We also denote the product FR by $FR = [\xi_0, \cdots, \xi_K]$. We propose now an alternating minimization procedure to recover the labels $(y_1, \cdots, y_k)$ from M.

Lemma 9 *Consider the optimization problem*

$$\min_{M \in \mathcal{M},\ R \in \mathbb{R}^{k \times k}:\ R^\top R = I_k} \|FR - M\|_F^2,$$

where $\mathcal{M} = \{[a_0 1_n, a_1 y_1, \cdots, a_k y_k], a \in \mathbb{R}^{k+1} : \|a\|_2 = 1, y_i \in \{\pm 1\}^n\}$.

Given M, the problem is equivalent to the orthogonal Procrustes problem (Schönemann, 1966). Denote by $U \Delta V^\top$ a singular value decomposition of $F^\top M$. The optimal R is obtained as $R = UV^\top$. While given R, the optimal M is obtained as

$$M = \frac{1}{\sqrt{\|\xi_1\|_1^2 + \|\xi_2\|_1^2 + \ldots + \|\xi_k\|_1^2}} [\|\xi_0\|_1 \operatorname{sign}(\xi_0), \cdots, \|\xi_k\|_1 \operatorname{sign}(\xi_k)].$$

Proof We give only the argument for the optimization problem with respect to M. Given R, the optimization problem in M is equivalent to $\max_{a \in \mathbb{R}^{k+1}: \|a\|_2 = 1,\ y \in \{-1,1\}^{n \times k}} \operatorname{tr}(FR)^\top M$ and $\operatorname{tr}(FR)^\top M = a_0 \xi_0^\top 1_n + \sum_{i=1}^k a_i \xi_i^\top y_i$. Thus by property of the dual norms the solution is given by $y_i = \operatorname{sign}(\xi_i)$ and $a_i = \frac{\|\xi_i\|_1}{\sqrt{\|\xi_1\|_1^2 + \|\xi_2\|_1^2 + \ldots + \|\xi_k\|_1^2}}$. ■

The minimization problem in Lemma 9 is non-convex; however we observe that performing few alternating optimizations is sufficient to recover the correct $(y_1, \ldots, y_k)$ from M.

5.4 Discussion

In this section we studied the tightness of convex relaxations under simple scenarios where the relaxed problem admits low-rank solutions generated by the solution of the original non-convex problem. Unfortunately the solutions lose the characterized rank when the initial problem is slightly perturbed since the rank of a matrix is not a continuous function. Nevertheless, the spectrum of the new solution is really spiked, and thus these results are quite conservative. We empirically observe that the principal eigenvectors keep recovering the correct information outside these scenarios. However this simple proof mechanism is not easily adaptable to handle perturbed problems in a straightforward way since it is difficult to characterize the properties of eigenvectors of the solution of a semi-definite program. Hence we are able to derive a proper theoretical study only for these simple models.

6. Algorithms

In this section, we present an optimization algorithm which is adapted to large n settings, and avoids the n-dimensional semidefinite constraint.

6.1 Reformulation

We aim to solve the general regularized problem which corresponds to Eq. (15)

$$\max_{V \succcurlyeq 0} \frac{2}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - \frac{1}{n} \operatorname{tr} V(X^\top X + n \operatorname{Diag}(a)^2) - \| \operatorname{Diag}(c) V \operatorname{Diag}(c) \|_1. \tag{23}$$

We consider a slightly different optimization problem:

$$\max_{V \succcurlyeq 0} \frac{1}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - \| \operatorname{Diag}(c) V \operatorname{Diag}(c) \|_1 \quad \text{s.t. } \operatorname{tr} V(\frac{1}{n} X^\top X + \operatorname{Diag}(a)^2) = 1. \tag{24}$$

When c is equal to zero, then Eq. (24) is exactly equivalent to Eq. (23); when c is small (as will typically be the case in our experiments), the solutions are very similar—in fact, one can show by Lagrangian duality that by a sequence of problems in Eq. (24), one may obtain the solution to Eq. (23).

6.2 Smoothing

By letting $A = \frac{X^\top X}{n} + \operatorname{Diag}(a)^2$, we consider a strongly-convex approximation of Eq. (24) as:

$$\begin{aligned} \max_{V \succcurlyeq 0} \quad & \frac{1}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - \| \operatorname{Diag}(c) V \operatorname{Diag}(c) \|_1 - \varepsilon \operatorname{tr}[(A^{\frac{1}{2}} V A^{\frac{1}{2}}) \log(A^{\frac{1}{2}} V A^{\frac{1}{2}})] \\ \text{s.t.} \quad & \operatorname{tr}(A^{\frac{1}{2}} V A^{\frac{1}{2}}) = 1, \end{aligned} \tag{25}$$

where $-\operatorname{tr} M \log(M)$ is a spectral convex function called the von-Neumann entropy (von Neumann, 1927). The difference in the two problems is known to be $\varepsilon \log(d)$ (Nesterov, 2007). As shown in Appendix G.1, the dual problem is

$$\min_{u \in \mathbb{R}_+^n, C \in \mathbb{R}^{d \times d} : |C_{ij}| \leqslant c_i c_j} \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{u_i} + \phi^\varepsilon \big(A^{-\frac{1}{2}} \big(\frac{1}{2n} X^\top \operatorname{Diag}(u) X - C \big) A^{-\frac{1}{2}} \big), \tag{26}$$

where $\phi^\varepsilon(M)$ is an ε-smooth approximation to the maximal eigenvalue of the matrix M.

6.3 Optimization algorithm

In order to solve Eq. (26), we split the objective function into a smooth part $F(u,C) = \phi^\varepsilon\big(A^{-\frac{1}{2}}\big(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C\big)A^{-\frac{1}{2}}\big)$ and a non-smooth part $H(u,C) = \mathbb{I}_{|C_{ij}|\leqslant c_i c_j} + \frac{1}{2n}\sum_{i=1}^n \frac{1}{u_i}$. We may then apply FISTA (Beck and Teboulle, 2009) updates to the smooth function $\phi^\varepsilon(A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}})$, along with a proximal operator for the non-smooth terms $\mathbb{I}_{|C_{ij}|\leqslant c_i c_j}$ and $\frac{1}{2n}\sum_{i=1}^n \frac{1}{u_i}$, which may be computed efficiently. See details in Appendix G.2.

Running-time complexity. Since we need to project on the SDP cone of size d at each iteration, the running-time complexity per iteration is $O(d^3 + d^2 n)$; given that often $n \geqslant d$, the dominating term is $O(d^2 n)$. It is still an open problem to make this linear in d. Our function being $O(1/\varepsilon)$-smooth, the convergence rate is of the form $O(1/(\varepsilon t^2))$. Since we stop when the duality gap is $\varepsilon \log(d)$ (as we use smoothing, it is not useful to go lower), the number of iterations is of order $1/(\varepsilon\sqrt{\log(d)})$. The proposed algorithm is a clear improvement over the existing approach by Bach and Harchaoui (2007) which is quadratic in n.

7. Experiments

We implemented the proposed algorithm in Matlab. The code has been made available in `https://drive.google.com/uc?export=download&id=0B5Bx9jrp7celMk5pOFI4UGtOZEk`. Two sets of experiments were performed: one on synthetically generated data sets and the other on real-world data sets. The details about experiments follow.

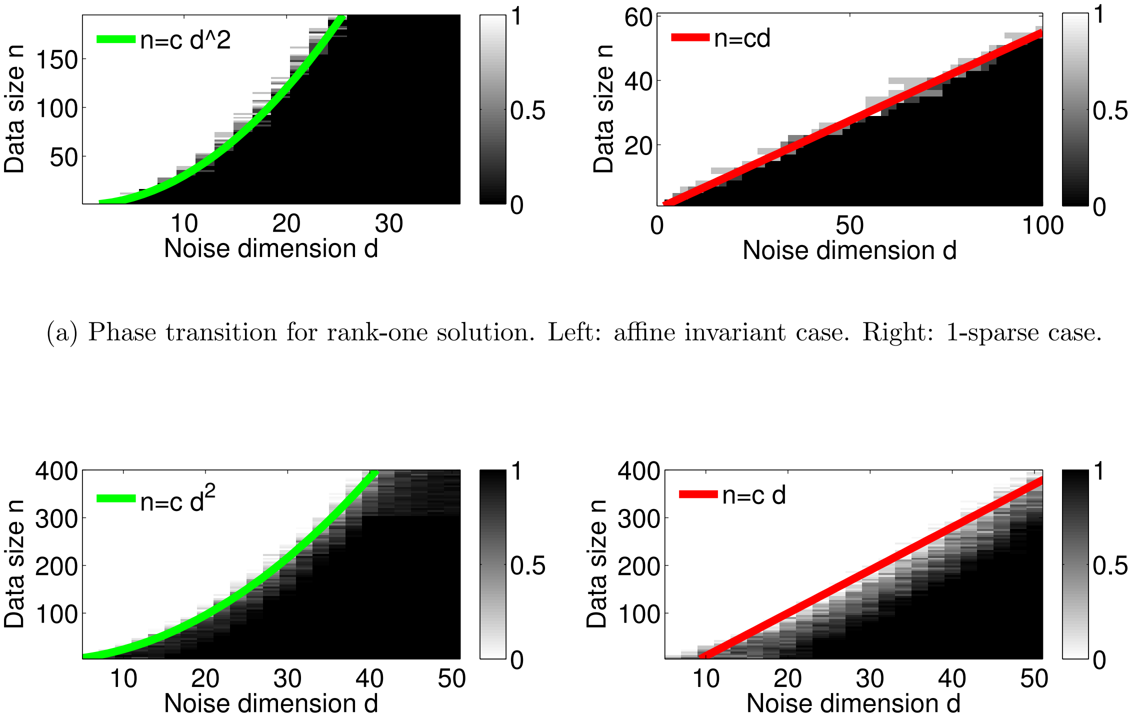
7.1 Experiments on synthetic data

In this section, we illustrate our theoretical results and algorithms on synthetic examples. The synthetic data were generated by assuming a fixed clustering with $\alpha_* \in [0,1]$, along a single direction and the remaining variables were whitened. We consider clustering error defined for a predictor $\bar{y}$ as $1 - (\bar{y}^\top y/n)^2$, with values in $[0,1]$ and equal to zero if and only if $y = \bar{y}$.

Phase transition. We first illustrate our theoretical results for the balanced case in Figure 1. We solve the relaxation in Eq. (12) and Eq. (18) for a large range of d and n using the `cvx` solver (Grant and Boyd, 2008, 2014). We show the results averaged over 4 replications and take $\lambda = 1/\sqrt{n}$ for the sparse problems. In Figure 1a, we investigate whether `cvx` finds a rank-one solution for a problem of size (n,d) (the value is 1 if the solution is rank-one and 0 otherwise). We compare the performance of the algorithms without ℓ_1-regularization in the affine invariant case and with ℓ_1-regularization in the 1-sparse case. We observe a phase transition with a scaling over the form $n = O(d^2)$ for the affine invariant case and $n = O(d)$ for the 1-sparse case. This is better than what is expected by the theory and corresponds rather to the performance of the principal eigenvector of the solution. It is worth noting that it may be uncertain to really distinguish between a rank-one solution and a spiked solution.

We also solve the relaxation for 4-sparse problems of different sizes d and n and plot the clustering error. We compare, in Figure 1b, the performance of the formulation in Eq. (12)

(without ℓ_1-regularization) which corresponds to the affine invariant case, against the ℓ_1-regularized formulation in Eq. (18). We notice a phase transition of the clustering error with a scaling over the form $n = O(d^2)$ for the affine invariant case and $n = O(d)$ for the 4-sparse case. It supports our conjecture on the scaling of order $n = O(ld)$ for l-sparse problems. Comparing left plots of Figure 1a and Figure 1b, we observe that the two phase-transitions occur at the same scaling between n and d. Thus there are few values of (n, d) for which the cvx solver finds a solution whose rank is strictly larger than one and whose principal eigenvector has a low clustering error. This illustrates, in practice, this solver aims to find a rank-one solution under the improved scaling $n = O(d^2)$.



(a) Phase transition for rank-one solution. Left: affine invariant case. Right: 1-sparse case.

(b) Phase transition for clustering error. Left: affine invariant case . Right: 4-sparse case.

Figure 1: Phase transition plots.

Unbalanced case. We generate an unbalanced problem for $d = 10$, $n = 80$ and $\alpha_* = 0.25$ and we average the results over 10 replications. We compare the clustering error for the constrained and the penalized relaxations in Eq. (7) and Eq. (8) when we consider the sign of the first or second eigenvector and when we use projection technique defined as $(\Pi_n Y_{(2)} \Pi_n)_{(1)}$ where $Y_{(k)}$ is the best rank-k approximation of Y, to extract the information of y. We see in Figure 2 that (a) for the constrained case, the range of α such that the sign of y is recovered is cut in two parts where one eigenvector is correct, whereas the projection method performs well on the whole set. (b) For the penalized case, the correct sign is

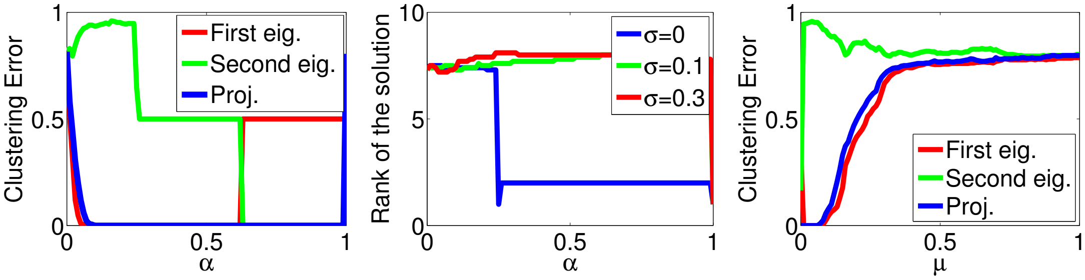


Figure 2: Unbalanced problem for $n = 80$, $d = 10$ and $\alpha_* = 0.25$. Left: Clustering error for the constrained relaxation. Middle: Rank of the solution for different level of noise σ. Right: Clustering error for the penalized relaxation.

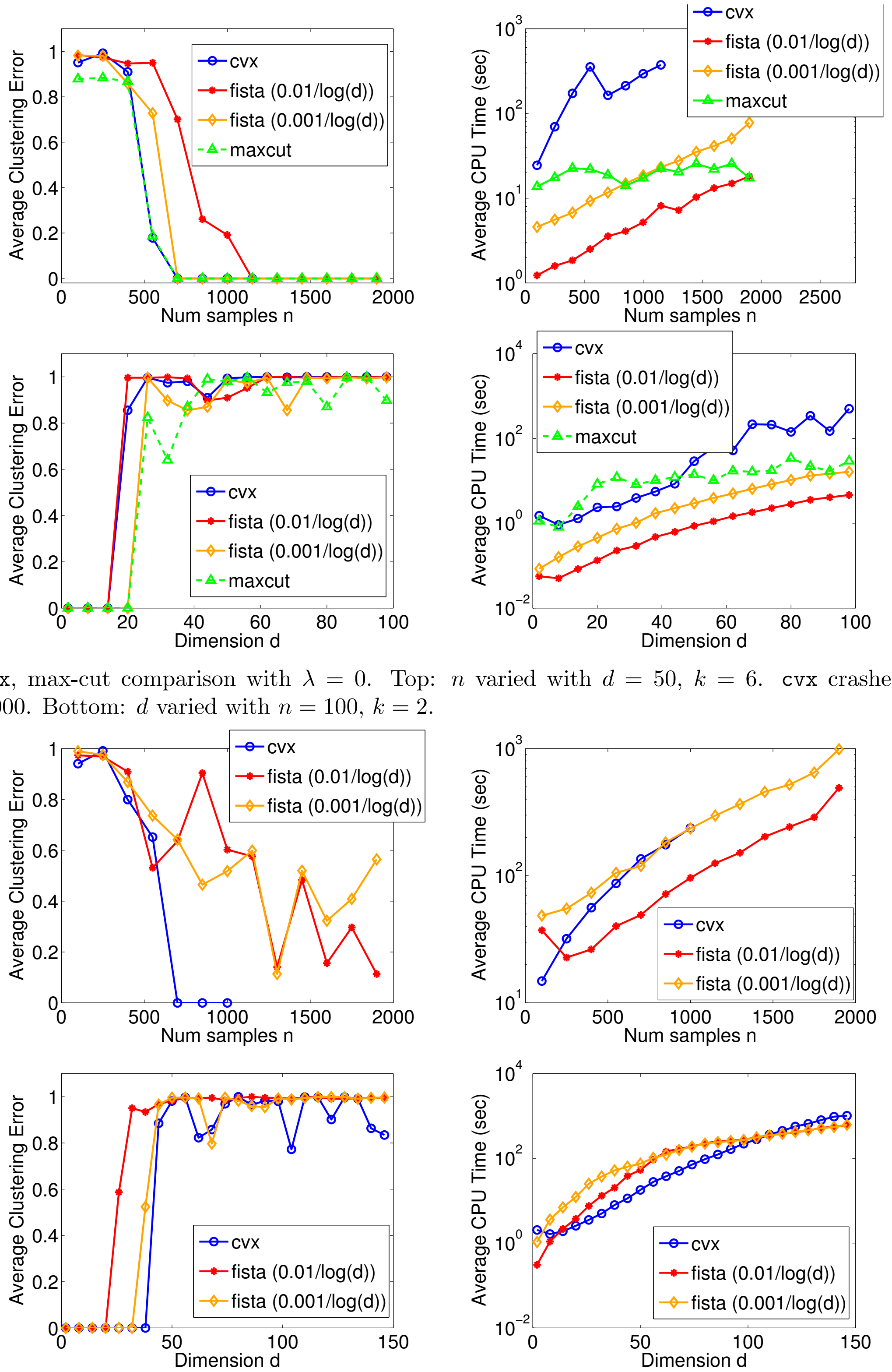
recovered for ν close to 0 by the first eigenvector and the projection method whereas the second one performs always badly. (c) When there is zero noise the rank of the solution is one for $\alpha \in \{\alpha_*, 1\}$, two for $\alpha \in (\alpha_*, 1)$ and greater otherwise. These findings confirm our analysis. However, when y is corrupted by some noise this result is no longer true.

Runtime experiments. We generated data with a k-sparse direction of projection v by adding $d - k$ noise variables to a randomly generated and rotated k-dimension data. The scalability of the FISTA based optimization algorithm illustrated in Section 6.3 to solve Eq. (24) (with $c = \sqrt{\lambda} 1_d$, $a = 0_d$) was compared against a benchmark `cvx` solver (which solves Eq. (18)). Experiments were performed for $\lambda = 0$ and $\lambda = 0.001$, the coefficient associated with the sparse $\|V\|_1$ term. For a fixed d, `cvx` breaks down for large n values (typically $n \geqslant 1000$). Similarly, the runtime required by `cvx` is generally high for $\lambda = 0$ and is comparable to our method for $\lambda = 0.001$. This behavior is illustrated in Figure 3.

When $\lambda = 0$, the problem reduces exactly to the original Diffrac problem (Bach and Harchaoui, 2007). In the plots in Figure 3a our implementation using FISTA is compared to the baseline Diffrac which is solved with max-cut SDP (Boumal et al., 2014). We observed that our method is comparable in terms of runtime and clustering performance of low-rank methods for max-cut. However, for $\lambda > 0$, the equivalence with max-cut disappears.

The plots in these figures show the behavior of FISTA for two different stopping criteria: $\varepsilon = 10^{-2}/\log(d)$ and $\varepsilon = 10^{-3}/\log(d)$. It is observed that the choice $10^{-3}/\log(d)$ gives a better accurate solution at the cost of more number of iterations (and hence higher runtime). For sparse problems in Figure 3b, we see that cvx gets a better clustering performance (while crashing for large n); the difference would be reduced with a smaller duality gap for FISTA.

Clustering performance. Experiments comparing the proposed method (Eq. (24) with $c = \sqrt{\lambda} 1_d$ and $a = 0_d$ solved using FISTA based optimization algorithm, and Eq. (18) solved using benchmark `cvx` solver) with K-means and alternating optimization are given in Figure 4. K-means is run on the whitened variables in $\mathbb{R}^d$. Alternating optimization is another popular method proposed by Ye et al. (2008) for dimensionality reduction with clustering (where alternating optimization of w and y is performed to solve the non-convex formulation (2)). The plots show that both K-means and alternating optimization fail when only a few dimensions of noise variables are present. The plots also show that with the introduction of a sparse regularizer (corresponding to the non-zero λ) the proposed



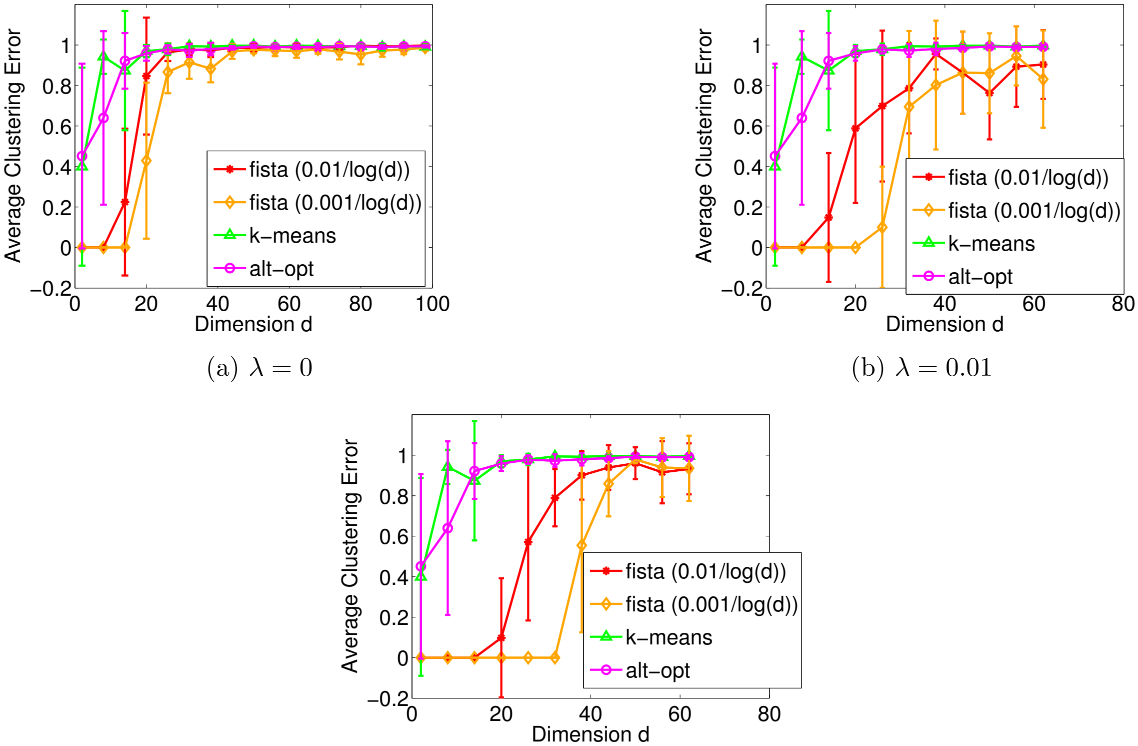
(a) cvx, max-cut comparison with $\lambda = 0$. Top: n varied with $d = 50$, $k = 6$. cvx crashed for $n \approx 1000$. Bottom: d varied with $n = 100$, $k = 2$.

(b) cvx comparison with $\lambda = 0.001$. Top: n varied with $d = 50$, $k = 6$. cvx crashed for $n \approx 1000$. Bottom: d varied with $n = 100$, $k = 2$.

Figure 3: Scalability experiments.

method becomes more robust to noisy dimensions. As observed earlier, the performance of FISTA is also sensitive to the choice of ε.

Finally we give a comparison of sparse discriminative clustering (cvx and FISTA) with max-margin clustering (Li et al., 2009) in Figure 5. We note that square loss is used in our framework whereas hinge loss is used in max-margin clustering. We have also included the behavior of K-means and alternating optimization methods in Figure 5 for completeness. From this plot, it is clear that the max-margin clustering is sensitive to noisy dimensions present in the data. Sparse discriminative clustering with square loss is able to maintain zero cluster error for a large number of noisy dimensions, while the performance of max-margin clustering starts deteriorating after adding a few noisy dimensions. However, we note from Figure 5 that for large dimensions, the hinge loss used in max-margin clustering is observed to provide a better solution than the square loss used in our framework.



(a) $\lambda = 0$

(b) $\lambda = 0.01$

(c) $\lambda = 0.001$

Figure 4: Comparison with k-means and alternating optimization, $n = 100$.

7.2 Experiments on real-world data

Experiments on two-class data. Experiments were conducted on real two-class classification datasets[2] to compare the performance of sparse discriminative clustering against non-sparse discriminative clustering, alternating optimization, K-means and max-margin

[2] The data sets were obtained from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

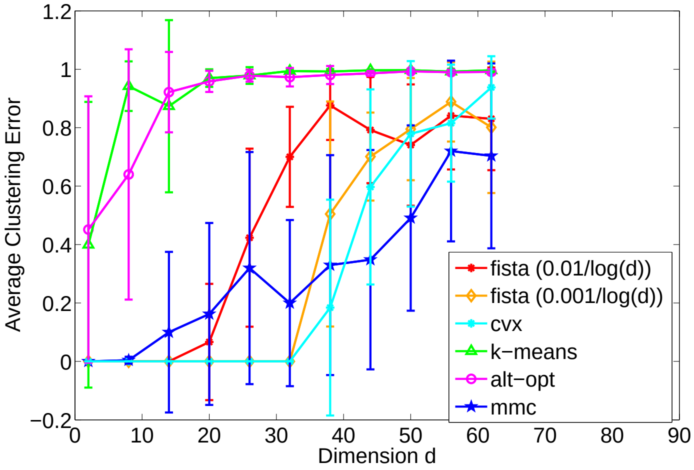


Figure 5: Comparison with k-means, alternating optimization and max-margin clustering (mmc), $n = 100$. The plots for FISTA, `cvx` and mmc correspond to the best choice of regularization parameters.

clustering algorithms. For sparse and non-sparse discriminative clustering, we consider the problem in Eq. (24) and the algorithm detailed in Section 6.3 (with the regularization $c = 0$ for the non-sparse case). The alternating optimization method is described in Proposition 1. For the two-class datasets, the clustering performance for a cluster $\bar{y} \in \{+1, -1\}^n$ obtained from an algorithm under comparison, was computed as $1 - (\bar{y}^\top y/n)^2$, where y is the original labeling. Here we explicitly compare the output of clustering with the original labels of the data points.

The dataset details and clustering performance results are summarized in Table 1. The experiments for discriminative clustering were conducted for different values of $a, c \in \{10^{-3}, 10^{-2}, 10^{-1}\}1_d$ associated with the ℓ_2-regularizer and ℓ_1-regularizer respectively. The range of cluster imbalance parameter was chosen to be $\nu \in \{0.01, 0.25, 0.5, 0.75, 1\}$. Note that for $\nu \neq 1$, the reformulation given in Eq. (14) was used, as explained after Eq. (17) in Section 3. The results given in Table 1 pertain to the best choices of these parameters. Similarly, the values of regularization parameter for max-margin clustering (Li et al., 2009) were chosen from the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$ and the cluster balance parameter was chosen from $\{0.1, 0.2, \dots, 0.9\}$. The results for alternating optimization and K-means show the average cluster error (and standard deviation) over 10 different runs. These results show that the cluster error is quite high for many datasets. This is primarily due to the absence of an ambient low-dimensional clustering of the two-class data, which can be identified by the simple linear model presented in this paper. Since K-means does not provide explicit dimensionality reduction, it might not be able to take advantage of the existence of an ambient low-dimensional clustering of the two-class data and its performance is poor. The results show that max-margin clustering achieves best clustering performance on most datasets. This improved performance of max-margin clustering may be due to the use of hinge loss, as opposed to square loss used in discriminative clustering in this paper. However for the heart dataset, we note that the sparse version with the square loss performs significantly better than the non-sparse version with the hinge loss (see additional experiments in Figure 5). The results also show that adding sparse regularizers to discriminative clustering helps in a better cluster identification when compared to the non-sparse case.

Table 1: Experiments on two-class datasets

Dataset	n	d	Cluster Error				
			Sparse Discriminative Clustering	Non-sparse Discriminative Clustering	Alternating Optimization	K-means	Max-margin Clustering
Heart	270	3	**0.52**	0.61	0.97 ± 0.03	0.91 ± 0.09	0.93
Diabetes	768	8	**0.88**	**0.88**	0.91 ± 0.05	0.93 ± 0.06	**0.88**
Breast-cancer	683	10	**0.15**	**0.15**	0.48 ± 0.17	0.68 ± 0.24	**0.15**
Australian	690	14	**0.5**	**0.5**	0.88 ± 0.17	0.87 ± 0.21	**0.5**
Liver-disorder	345	6	0.97	0.97	0.99 ± 0.01	0.99 ± 0.01	**0.73**
Sonar	208	60	**0.92**	0.95	0.98 ± 0.02	0.99 ± 0.01	**0.92**
DNA(1 vs 2,3)	1400	180	0.75	0.83	0.99 ± 0.01	0.98 ± 0.02	**0.71**
a1a	1605	113	0.74	0.75	0.98 ± 0.02	0.8 ± 0.08	**0.69**
w1a	2270	290	**0.11**	**0.11**	0.92 ± 0.08	0.16 ± 0.06	**0.11**

Experiments on real multi-label data. Experiments were also conducted on the Microsoft COCO dataset[3] to demonstrate the effectiveness of the proposed method in discovering multiple labels. We considered $n = 2000$ images from the dataset, each of which was labeled with a subset of $K = 80$ labels. The labels identified the objects in the images like person, car, chair, table, etc. and the corresponding features for each image were extracted from the last layer of a conventional convolutional neural network (CNN). The CNN was originally trained over the imagenet data (Krizhevsky et al., 2012).

For each image in the dataset, we obtained $d = 1000$ features. We then performed discriminative clustering on the 2000×1000 data matrix X and obtained the label matrix Y which was then subjected to the alternating optimization procedure (see Section 5.3).

It is clearly unlikely to recover perfect labels; therefore we now describe a way of measuring the amount of information which is recovered. In order to extract meaningful cluster information from the result so-obtained, we computed the correlation matrix $Y_k^\top \Pi_n Y_{true}$ where Y_{true} is the $n \times K$ label matrix containing actual labels and Π_n is the $n \times n$ centering matrix $I_n - \frac{1}{n} 1_n 1_n^\top$. The k predicted labels for the examples are present in the Y_k matrix of size $n \times k$. In order to choose an appropriate value of k, we plotted $\mathrm{Tr}(\Phi_{Y_{true}} \Phi_{Y_k})$ (shown in Figure 6 along with a K-means baseline), where $\Phi_{Y_k} = Y_k (Y_k{}^\top Y_k)^{-1} Y_k^\top$. From these plots, we chose $k = 30$ to be a suitable value for our interpretation purposes.

After choosing an arbitrary value of $k = 30$, we plotted the correlations between the actual and predicted labels. The heat map of the normalized absolute correlations is given in Figure 7, where the columns and rows corresponding to the 80 true labels and 30 predicted labels respectively, are ordered according to the sum of squared correlations (the top-scoring labels appear to the left-bottom). From this plot, we extract following highly correlated

[3] Dataset obtained from `http://mscoco.org/dataset`

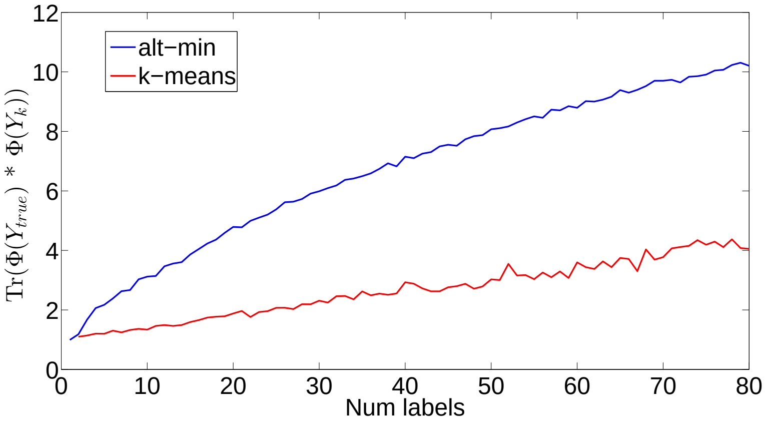


Figure 6: Plot of $\mathrm{Tr}(\Phi_{Y_{true}}\Phi_{Y_k})$.

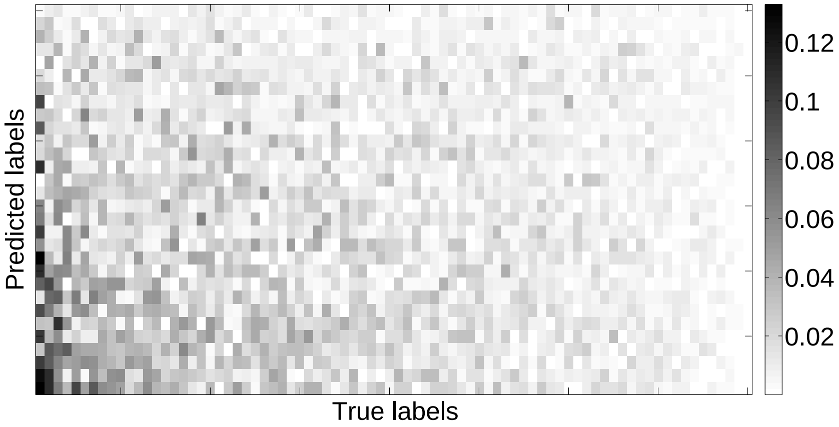


Figure 7: Heat map of correlations, $Y_k^\top \Pi_n Y_{true}$ with $k = 30$, with columns and rows ordered according to the sum of squared correlations.

labels: person, dining table, car, chair, cup, tennis racket, bowl, truck, fork, pizza, showing that these labels were partially recovered by our unsupervised technique (note that the CNN features are learned with supervision on the different dataset Imagenet, hence there is still some partial supervision).

8. Conclusion

In this paper, we provided a sparse extension of the discriminative clustering framework, and gave a first analysis of its theoretical performance in the totally unsupervised situation, highlighting provable scalings between ambient dimension d, number of observations and "clusterability" of irrelevant variables. We also proposed an efficient algorithm which is the first of its kind to be linear in the number of observations for discriminative clustering with the square loss. Our work could be extended in a number of ways, e.g., extending the sparse analysis to l-sparse case with higher l, extending the framework to nonlinear clustering using kernels, considering related weakly supervised learning extensions (Joulin and Bach, 2012), going beyond uniqueness of rank-one solutions, and improving the complexity of our algorithm to $O(nd)$, for example using stochastic gradient techniques.

Acknowledgments

The authors would like to thank Nicolas Boumal for interesting discussions and for his advice on max-cut implementation, Piotr Bojanowski for his help with multi-label datasets and the reviewers for their constructive and helpful comments which helped to significantly improve the paper.

References

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

F. Bach and Z. Harchaoui. DIFFRAC : a discriminative and flexible framework for clustering. In *Advances in Advances in Neural Information Processing Systems (NIPS)*, 2007.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

R. Bellman. A note on cluster analysis and dynamic programming. *Mathematical Biosciences*, 1973.

G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller. In search of non-Gaussian components of a high-dimensional distribution. *The Journal of Machine Learning Research*, 7:247–282, 2006.

Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Finding actors and actions in movies. In *Proc. IEEE International Conference on Computer Vision*, 2013.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*, volume 3 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer-Verlag, 2000. Theory and examples.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab Toolbox for Optimization on Manifolds. *Journal of Machine Learning Research*, 2014.

J. Bourgain, V. H. Vu, and P. M. Wood. On the singularity probability of discrete random matrices. *Journal of Functional Analysis*, 258(2):559–603, 2010.

S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, 2004.

T. De Bie and N. Cristianini. Convex methods for transduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 73–80, 2003.

F. De la Torre and T. Kanade. Discriminative cluster analysis. In *Proceedings of the conference on machine learning (ICML)*, 2006.

E. Diederichs, A. Juditsky, A. Nemirovski, and V. Spokoiny. Sparse non-Gaussian component analysis by semidefinite programming. *Machine learning*, 91(2):211–238, 2013.

C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and K-means clustering. In *Proceedings of the conference on machine learning (ICML)*, 2007.

D. Freedman. *Statistical models: theory and practice.* Cambridge University Press, 2009.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

A. Frieze and M. Jerrum. Improved approximation algorithms for MAX k-CUT and MAX BISECTION. In *Integer Programming and Combinatorial Optimization.* Springer, 1995.

M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theoret. Comput. Sci.*, 1(3):237–267, 1976.

M. X. Goemans and D. P. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *J. ACM*, 42(6):1115–1145, November 1995.

J. C. Gower and G. J. S. Ross. Minimum spanning trees and single Linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1), 1969.

M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.

M. Grant and S. Boyd. CVX: Matlab Software for Disciplined Convex Programming, version 2.1, March 2014.

Edouard Grave. A convex relaxation for weakly supervised relation extraction. In *EMNLP*, 2014.

G. Huang, J. Zhang, S. Song, and Z. Chen. Maximin separation probability clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

A. Joulin and F. Bach. A convex relaxation for weakly supervised classifiers. In *Proceedings of the conference on machine learning (ICML)*, 2012.

A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *Proc. CVPR*, 2010a.

A. Joulin, J. Ponce, and F. Bach. Efficient optimization for discriminative latent class models. In *Advances in Advances in Neural Information Processing Systems (NIPS)*, 2010b.

M. Journée, F. Bach, P-A Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 2010.

A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *International Symposium on Theory of Computing (STOC)*, pages 553–562, 2010.

R. M. Karp. Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Plenum, New York, 1972.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Advances in Neural Information Processing Systems (NIPS)*, 2012.

Rémi Lajugie, Piotr Bojanowski, Philippe Cuvillier, Sylvain Arlot, and Francis Bach. A weakly-supervised discriminative model for audio-to-score alignment. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016.

N. Le Roux and F. Bach. Local component analysis. In *Proceedings of the International Conference on Learning Representations*, 2013.

Y.-F. Li, I. W. Tsang, J. T.-Y. Kwok, and Z-H. Zhou. Tighter and convex maximum margin clustering. In *AISTATS*, pages 344–351, 2009.

Z. Q. Luo, W. K. Ma, A. C. So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *Signal Processing Magazine, IEEE*, 2010.

J. B. MacQueen. Some Methods for Classification and Analysis of MultiVariate Observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Annual Symposium on Foundations of Computer Science (FOCS)*, pages 93–102, 2010.

Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.*, 2007.

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 Advances in Neural Information Processing Systems (NIPS)*. 2002.

G. Niu, B. Dai, L. Shang, and M. Sugiyama. Maximum volume clustering: a new discriminative clustering approach. *Journal of Machine Learning Research*, 14(1):2641–2687, 2013.

R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *Annual Symposium on Foundations of Computer Science (FOCS)*, pages 165–176, 2006.

P. H Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1), 1966.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 1996.

J. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 2012.

J. von Neumann. Thermodynamik quantummechanischer Gesamheiten. *Gött. Nach*, (1): 273–291, 1927.

F. Wang, B. Zhao, and C. Zhang. Linear time maximum margin clustering. *IEEE Transactions on Neural Networks*, 2010.

H. Wang, F. Nie, and H. Huang. Multi-View Clustering and Feature Learning via Structured Sparsity. In *Proceedings of the conference on machine learning (ICML)*, volume 28, 2013.

Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefinite programming. In *Handbook on Semidefinite, Conic and Polynomial Optimization*. Springer, 2012.

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Advances in Advances in Neural Information Processing Systems (NIPS)*, 2004.

J. Ye, Z. Zhao, and M. Wu. Discriminative k-means for clustering. In *Advances in Advances in Neural Information Processing Systems (NIPS)*, 2008.

K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. *IEEE Transactions on Neural Networks*, 2009.

Appendix A. Joint clustering and dimension reduction

Given y, we need to optimize the Rayleigh quotient $\frac{w^\top X^\top y y^\top X w}{w^\top X^\top X w}$ with a rank-one matrix in the numerator, which leads to $w = (X^\top X)^{-1} X^\top y$. Given w, we will show that the averaged distortion measure of K-means once the means have been optimized is exactly equal to $(y^\top \Pi_n X w)^2 / \|\Pi_n y\|_2^2$. Given the data matrix $X \in \mathbb{R}^{n \times d}$, K-means to cluster the data into two components will tend to approximate the data points in X by the centroids $c_+ \in \mathbb{R}^d$ and $c_- \in \mathbb{R}^d$ such that

$$\begin{aligned} X &\approx \frac{(y+1_n)}{2} c_+^\top - \frac{(y-1_n)}{2} c_-^\top \text{ (since } y \in \{-1,1\}^n) \\ &= \frac{y}{2}(c_+^\top - c_-^\top) + \frac{1}{2} 1_n (c_+^\top + c_-^\top). \end{aligned}$$

The objective of K-means can now be written as problem $\mathcal{KM}$:

$$\begin{aligned} &\min_{y,c_+,c_-} \left\| X - \frac{y}{2}(c_+^\top - c_-^\top) - \frac{1}{2} 1_n (c_+^\top + c_-^\top) \right\|_F^2 \\ &= \min_{y,c_+,c_-} \left\| X - \frac{(y+1_n)}{2} c_+^\top - \frac{(1_n - y)}{2} c_-^\top \right\|_F^2 \\ &= \min_{y,c_+,c_-} \|X\|_F^2 + \|c_+^\top\|_F^2 \left\| \frac{(y+1_n)}{2} \right\|^2 + \|c_-^\top\|_F^2 \left\| \frac{(1-y_n)}{2} \right\|^2 + 2 c_-^\top c_+ \frac{(y+1_n)}{2}^\top \frac{(1_n - y)}{2} \\ &\qquad -2 \operatorname{tr} X^\top \left(\frac{(y+1_n)}{2} c_+^\top + \frac{(1_n - y)}{2} c_-^\top \right) \\ &= \min_{y,c_+,c_-} \|X\|_F^2 + \|c_+^\top\|_F^2 \frac{1}{2}(n + 1_n^\top y) + \|c_-^\top\|_F^2 \frac{1}{2}(n - 1_n^\top y) - 2 c_+^\top X^\top \left(\frac{y + 1_n}{2} \right) \\ &\qquad -2 c_-^\top X^\top \left(\frac{1_n - y}{2} \right). \end{aligned}$$

Fixing y and minimizing with respect to c_+ and c_-, we get closed-form expressions for c_+ and c_- as

$$c_+ = \frac{X^\top (y + 1_n)}{(n + 1_n^\top y)} \quad \text{and} \quad c_- = \frac{X^\top (1_n - y)}{(n - 1_n^\top y)}.$$

Substituting these expressions in $\mathcal{KM}$, we have the following optimization problem in y:

$$\begin{aligned}
&\min_y \|X\|_F^2 - \frac{1}{2}\frac{\|X^\top(y+1_n)\|_F^2}{(n+1_n^\top y)} - \frac{1}{2}\frac{\|X^\top(1_n-y)\|_F^2}{(n-1_n^\top y)}\\
&= \min_y \|X\|_F^2 - \frac{1}{2}\frac{\operatorname{tr} XX^\top(y+1_n)(y+1_n)^\top}{(n+1_n^\top y)} - \frac{1}{2}\frac{\operatorname{tr} XX^\top(1_n-y)(1_n-y)^\top}{(n-1_n^\top y)}\\
&= \min_y \|X\|_F^2 - \frac{2}{(n+1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{y+1_n}{2}\right)\left(\frac{y+1_n}{2}\right)^\top\\
&\qquad -\frac{2}{(n-1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{1_n-y}{2}\right)\left(\frac{1_n-y}{2}\right)^\top\\
&= \min_y\ \operatorname{tr} XX^\top - \frac{2}{(n+1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{y+1_n}{2}\right)\left(\frac{y+1_n}{2}\right)^\top\\
&\qquad -\frac{2}{(n-1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{1_n-y}{2}\right)\left(\frac{1_n-y}{2}\right)^\top\\
&= \min_y\ \operatorname{tr} XX^\top\Bigg(I - \frac{1}{2(n+1_n^\top y)}(yy^\top + 1_n1_n^\top + y1_n^\top + 1_ny^\top)\\
&\qquad -\frac{1}{2(n-1_n^\top y)}(1_n1_n^\top + yy^\top - 1_ny^\top - y1_n^\top)\Bigg).
\end{aligned}$$

By the centering of X, we have $1_n^\top X = 0$ and hence $\operatorname{tr} XX^\top 1_n 1_n^\top = \operatorname{tr} XX^\top 1_n y^\top = \operatorname{tr} XX^\top y 1_n^\top = 0$. Therefore, we obtain

$$\begin{aligned}
&\min_y\ \operatorname{tr} XX^\top\left(I - \frac{1}{2(n+1_n^\top y)}(yy^\top) - \frac{1}{2(n-1_n^\top y)}(yy^\top)\right)\\
&= \min_y\ \operatorname{tr} XX^\top\left(I - (yy^\top)\left(\frac{1}{2(n+1_n^\top y)} + \frac{1}{2(n-1_n^\top y)}\right)\right)\\
&= \min_y\ \operatorname{tr} XX^\top\left(I - (yy^\top)\left(\frac{n}{n^2-(1_n^\top y)^2)}\right)\right)\\
&= \min_y\ \operatorname{tr} XX^\top\left(I - \frac{nyy^\top}{n^2-(1_n^\top y)^2}\right).
\end{aligned}$$

Thus we have the equivalent K-means problem as

$$\min_{y\in\{-1,1\}^n} \frac{1}{n}\operatorname{tr} Xww^\top X^\top\left(I - \frac{n}{n^2-(y^\top 1)^2}yy^\top\right) = 1 - \max_{y\in\{-1,1\}^n}\frac{(w^\top X^\top y)^2}{n^2-(y^\top 1)^2}.$$

Thus the averaged distortion measure of K-means with the optimized means is $\frac{(y^\top \Pi_n X w)^2}{\|\Pi_n y\|_2^2}$.

Appendix B. Full (unsuccessful) relaxation

It is tempting to find a direct relaxation of Eq. (2). It turns out to lead to a trivial relaxation, which we outline in this section. When optimizing Eq. (2) with respect to w, we obtain $\max\limits_{y\in\{-1,1\}^n} \frac{y^\top X(X^\top X)^{-1}X^\top y}{y^\top \Pi_n y}$, leading to a quasi-convex relaxation as $\max\limits_{\substack{Y\succcurlyeq 0,\\ \operatorname{diag}(Y)=1}} \frac{\operatorname{tr} YX(X^\top X)^{-1}X^\top}{\operatorname{tr} \Pi_n Y}$. Unfortunately, this relaxation always leads to trivial solutions as described below.

Consider the quasi-convex relaxation

$$\max_{Y\succcurlyeq 0, \mathrm{diag}(Y)=1} \frac{\operatorname{tr} YX(X^\top X)^{-1}X^\top}{\operatorname{tr} \Pi_n Y}. \tag{27}$$

By definition of Π_n this relaxation is equal to:

$$\max_{Y\succcurlyeq 0, \mathrm{diag}(Y)=1} \frac{1}{n}\frac{\operatorname{tr} YX(X^\top X)^{-1}X^\top}{1-\frac{1_n^\top Y 1_n}{n^2}}.$$

Let $\mathcal{A} = \{Y \succcurlyeq 0,\ \mathrm{diag}(Y) = 1\}$ the feasible set of this problem and define $\mathcal{B} = \{M \succcurlyeq 0,\ \mathrm{diag}(M) = 1 + \frac{1_n^\top M 1_n}{n^2}\}$. Let $Y \in \mathcal{A}$, then M defined by $M = \frac{Y}{1-\frac{1_n^\top Y 1_n}{n^2}}$ belongs to $\mathcal{B}$ since $1 + \frac{1_n^\top M 1_n}{n^2} = 1 + \frac{1_n^\top Y 1_n}{n^2 - 1_n^\top Y 1_n} = \frac{1}{1-\frac{1_n^\top Y 1_n}{n^2}} = \mathrm{diag}(M)$. Reciprocally for $M \in \mathcal{B}$, we can define $Y = \frac{M}{1+\frac{1_n^\top M 1_n}{n^2}}$, such that $\mathrm{diag}(Y) = 1$ and $Y \in \mathcal{A}$ and then verify that $M = \frac{Y}{1-\frac{1_n^\top Y 1_n}{n^2}}$. Thus the problem Eq. (27) is equivalent to the relaxation

$$\max_{M\succcurlyeq 0, \mathrm{diag}(M)=1+\frac{1_n^\top M 1_n}{n^2}} \frac{1}{n}\operatorname{tr} MX(X^\top X)^{-1}X^\top. \tag{28}$$

The Lagrangian function of this problem can be written as:

$$\begin{aligned} L(\mu) &= \operatorname{tr} MX(X^\top X)^{-1}X^\top - \frac{\mu^\top}{n}[\mathrm{diag}(M) - 1_n - \frac{1_n^\top M 1_n}{n^2}1_n] \\ &= \operatorname{tr} M[X(X^\top X)^{-1}X^\top - \mathrm{Diag}(\mu) + \frac{1_n^\top \mu}{n^2}1_n 1_n^\top] + \frac{1}{n}\mu^\top 1_n. \end{aligned}$$

Using $L(\mu)$ and the PSD constraint $M \succcurlyeq 0$, the dual problem is given by

$$\min_{\mu} \frac{\mu^\top 1_n}{n} \quad \text{s.t. } \mathrm{Diag}(\mu) - \frac{1_n^\top \mu}{n^2}1_n 1_n^\top \succcurlyeq X(X^\top X)^{-1}X^\top.$$

Since $X(X^\top X)^{-1}X^\top \succcurlyeq 0$, this implies for the dual variable μ:

$$\begin{aligned} \mathrm{Diag}(\mu) - \frac{1_n^\top \mu}{n^2}1_n 1_n^\top \succcurlyeq 0 \quad &\Leftrightarrow \quad 1_n^\top \mathrm{Diag}(\mu)^{-1} 1_n \leq \frac{n^2}{\mu^\top 1_n} \\ &\Leftrightarrow \quad \sum_{i=1}^n \frac{1}{\mu_i} \leq \frac{n^2}{\sum_{i=1}^n \mu_i} \\ &\Leftrightarrow \quad \frac{1}{n}\sum_{i=1}^n \frac{1}{\mu_i} \leq \frac{1}{\frac{1}{n}\sum_{i=1}^n \mu_i}. \end{aligned}$$

However for $\nu \in \mathbb{R}^n$, the harmonic mean $\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{\nu_i}\right]^{-1}$ is always smaller than the arithmetic mean $\frac{1}{n}\sum_{i=1}^n \nu_i$ with equality if and only if $\nu = c1_n$ for $c \in \mathbb{R}$.

Thus the dual variable μ is constant and the diagonal constraint in problem Eq. (28) simplifies itself as a trace constraint. Therefore the problem is equivalent to the trivial relaxation below for which each eigenvector of $X(X^\top X)^{-1}X^\top$ is a solution:

$$\max_{M\succcurlyeq 0,\ \mathrm{tr}(M)=n+\frac{1_n^\top M 1_n}{n}} \operatorname{tr} MX(X^\top X)^{-1}X^\top.$$

Appendix C. Equivalent relaxations

In this section, we give details about two equivalent relaxations.

C.1 First equivalent relaxation

We start from the penalized version of Eq. (5),

$$\min_{y\in\{-1,1\}^n,\ v\in\mathbb{R}^d} \frac{1}{n}\|\Pi_n y - Xv\|_2^2 + \nu\frac{(y^\top 1_n)^2}{n^2}, \tag{29}$$

which we expand as:

$$\min_{y\in\{-1,1\}^n,\ v\in\mathbb{R}^d} \frac{1}{n}\operatorname{tr}\Pi_n yy^\top - \frac{2}{n}\operatorname{tr}Xvy^\top + \frac{1}{n}\operatorname{tr}X^\top Xvv^\top + \nu\frac{(y^\top 1_n)^2}{n^2}, \tag{30}$$

and relax as, using $Y = yy^\top$, $P = yv^\top$ and $V = vv^\top$,

$$\min_{V,P,Y} \frac{1}{n}\operatorname{tr}\Pi_n Y - \frac{2}{n}\operatorname{tr}P^\top X + \frac{1}{n}\operatorname{tr}X^\top XV + \nu\frac{1_n^\top Y 1_n}{n^2} \text{ s.t. } \begin{pmatrix} Y & P \\ P^\top & V\end{pmatrix} \succcurlyeq 0,\ \operatorname{diag}(Y) = 1. \tag{31}$$

When optimizing Eq. (31) with respect to V and P, we get exactly Eq. (8). Indeed we solve this problem by fixing the matrix Y such that $Y = Y_0$ and $\operatorname{diag}(Y_0) = 1_n$. Then the Lagrangian function of the problem in Eq. (31) can be written as

$$\begin{aligned} L(A) &= \frac{1}{n}\operatorname{tr}\Pi_n Y - \frac{2}{n}\operatorname{tr}P^\top X + \frac{1}{n}\operatorname{tr}X^\top XV + \nu\frac{1_n^\top Y 1_n}{n^2} + \operatorname{tr}A(Y - Y_0) \\ &= \begin{pmatrix} Y & P \\ P^\top & V\end{pmatrix}\begin{pmatrix} \frac{1}{n}\Pi_n + \frac{\nu}{n^2}1_n1_n^\top + A & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X\end{pmatrix} - \operatorname{tr}AY_0. \end{aligned}$$

Using $L(A)$ and the psd constraint $\begin{pmatrix} Y & P \\ P^\top & V\end{pmatrix} \succcurlyeq 0$, we write the dual problem as

$$\min_A \operatorname{tr}AY_0 \text{ s.t. } \begin{pmatrix} \frac{1}{n}\Pi_n + \frac{\nu}{n^2}1_n1_n^\top + A & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X\end{pmatrix} \succcurlyeq 0.$$

From the Schur's complement condition of $\begin{pmatrix} \frac{1}{n}\Pi_n + \frac{\nu}{n^2}1_n1_n^\top + A & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X\end{pmatrix} \succcurlyeq 0$, we obtain $\frac{1}{n}\Pi_n + \frac{\nu}{n^2}1_n1_n^\top + A \succcurlyeq \frac{1}{n}X(X^\top X)^{-1}X^\top$. Substituting the bound for A we get the optimal objective function value

$$\mathcal{D}^* = \frac{1}{n}\operatorname{tr}X(X^\top X)^{-1}X^\top Y_0 - \frac{1}{n}\operatorname{tr}\Pi_n Y_0 - \frac{\nu}{n^2}1_n^\top Y_0 1_n.$$

Note that the optimal dual objective value $\mathcal{D}^*$ corresponds to a fixed Y_0. Hence by maximizing with respect to Y we obtain exactly Eq. (8) and therefore, the convex relaxation in Eq. (11) is equivalent to Eq. (8). Moreover the Karush-Kuhn-Tucker (KKT) conditions gives

$$P^\top - X + VX^\top X = 0 \text{ and } -YX + PX^\top X = 0$$

Thus the optimum is attained for $P = YX(X^\top X)^{-1}$ and $V = (X^\top X)^{-1}X^\top YX(X^\top X)^{-1}$.

C.2 Second equivalent relaxation

For $\nu = 1$, we solve the problem in Eq. (31) by fixing the matrix $V = V_0$. Then the Lagrangian function of this problem can be written as

$$\begin{aligned}\hat{L}(\mu, B) &= \frac{1}{n}\operatorname{tr}\Pi_n Y - \frac{2}{n}\operatorname{tr} P^\top X + \frac{1}{n}\operatorname{tr} X^\top X V + \nu\frac{1_n^\top Y 1_n}{n^2} + \mu^\top(\operatorname{diag}(Y) - 1_n) + \operatorname{tr} B(V - V_0) \\ &= \begin{pmatrix} Y & P \\ P^\top & V\end{pmatrix}\begin{pmatrix}\frac{1}{n}I_n + \operatorname{diag}(\mu) & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X + B\end{pmatrix} - \mu^\top 1_n - \operatorname{tr} BV_0.\end{aligned}$$

Using $\hat{L}(\mu, B)$ and the psd constraint $\begin{pmatrix} Y & P \\ P^\top & V\end{pmatrix} \succcurlyeq 0$, the dual problem is given by

$$\min_{\mu, B} \mu^\top 1_n + \operatorname{tr} BV_0 \text{ s.t. } \begin{pmatrix}\frac{1}{n}I_n + \operatorname{diag}(\mu) & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X + B\end{pmatrix} \succcurlyeq 0.$$

From the Schur's complement condition of $\begin{pmatrix}\frac{1}{n}I_n + \operatorname{diag}(\mu) & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X + B\end{pmatrix} \succcurlyeq 0$, we obtain $B \succcurlyeq \frac{1}{n^2}X^\top \operatorname{diag}(\mu + 1_n/n)^{-1}X - \frac{1}{n}X^\top X$. Substituting the bound for B we get the dual problem as

$$\begin{aligned} &\min_{\mu} \mu^\top 1_n + \frac{1}{n^2}\operatorname{tr} V_0 X^\top \operatorname{diag}(\mu + 1_n/n)^{-1}X - \frac{1}{n}\operatorname{tr} V_0 X^\top X \\ \Longrightarrow\quad &\min_{\mu} \sum_{i=1}^{n}\left(\mu_i + \frac{1}{n^2\mu_i + n}x_i^\top V_0 x_i\right) - \frac{1}{n}\operatorname{tr} V_0 X^\top X.\end{aligned}$$

Solving for μ_i, we get

$$\mu_i^* = \frac{1}{n}\sqrt{x_i^\top V_0 x_i} - \frac{1}{n}.$$

Substituting μ_i^* into the dual objective function, we get the optimal objective function value

$$\hat{D} = \frac{2}{n}\sum_{i=1}^{n}\sqrt{(XVX^\top)_{ii}} - 1 - \frac{1}{n}\operatorname{tr} V_0 X^\top X.$$

Furthermore the KKT conditions give

$$Y \operatorname{diag}(\nu + 1_n/n) - \frac{1}{n}PX^\top = 0 \text{ and } P^\top \operatorname{diag}(\nu + 1_n/n) - \frac{1}{n}VX^\top = 0.$$

Thus we obtain the following closed form expressions:

$$\begin{aligned} P &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2}XV \\ Y &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2}XVX^\top \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2}.\end{aligned}$$

The optimal dual objective value $\hat{D}$ corresponds to a fixed V_0. Therefore, maximizing with respect to V leads to the problem:

$$\min_{V \succcurlyeq 0} \quad 1 - \frac{2}{n}\sum_{i=1}^{n}\sqrt{(XVX^\top)_{ii}} + \frac{1}{n}\operatorname{tr}(VX^\top X). \tag{32}$$

Appendix D. Auxiliary results for Section 5.1

Here we provide auxiliary results for Section 5.1.

D.1 Auxiliary lemma

The matrix $X(X^\top X)^{-1}X^\top$ has the following properties (see e.g., (Freedman, 2009)).

Lemma 10 *The matrix $H = X(X^\top X)^{-1}X^\top$ is the orthogonal projection onto the column space of the design matrix X since:*

- *H is symmetric.*
- *H is idempotent $(H^2) = H$.*
- *X is invariant under H, that is $HX = X$.*

D.2 Rank-one solution of the relaxation Eq. (8)

We denote by $(x_i)_{i=1\dots n}$ the lines (or rows) of X.

Lemma 11 *The rank-one solution $Y_* = yy^\top$ is always a solution of the relaxation Eq. (8).*

Proof We give an elementary proof of this result without using convex optimization tools. Using Proposition 1 and Lemma 10 we have $Hy = y$, thus

$$\operatorname{tr} HY_* = \operatorname{tr} Hyy^\top = \operatorname{tr} yy^\top = n.$$

Moreover all $M \succcurlyeq 0$ can always be decomposed as $\sum_{i=1}^n \lambda_i u_i u_i^\top$ with $\lambda_i \geq 0$ and $(u_i)_{i=1,\dots,n}$ an orthonormal family. Since H is an orthogonal projection $(u_i)^\top H u_i = (Hu_i)^\top H u_i = \|Hu_i\|^2 \leq \|u_i\|^2 \leq 1$. Thus $\operatorname{tr} HM = \sum_{i=1}^n \lambda_i \operatorname{tr} Hu_i(u_i)^\top = \sum_{i=1}^n \lambda_i (u_i)^\top H u_i \leq \sum_{i=1}^n \lambda_i = \operatorname{tr} M$.

Then for all matrix M feasible we have $\operatorname{tr} HM \leq n$ since $\operatorname{diag}(M) = 1_n$ and $\operatorname{tr} HY_* = n$ which conclude the lemma. ■

D.3 Rank-one solution of the relaxation Eq. (12)

Lemma 12 *The rank-one solution $V_* = vv^\top$ is always a solution of the relaxation Eq. (12).*

Proof The Karush-Kuhn-Tucker (KKT) optimality conditions for the problem are for the dual variable $A \preccurlyeq 0$:

$$\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^\top}{\sqrt{x_i^\top V x_i}} - \frac{1}{n} XX^\top = A \text{ and } AV = 0 \;\; \text{(Complementary Slackness)}.$$

Since $x_i^\top w = y_i$, $\sqrt{x_i^\top V_* x_i} = |y_i| = 1$, V_* and the dual variable $A = 0$ satisfy the KKT conditions and then V_* is solution of this problem. ■

D.4 Proof of Proposition 2

In the following lemma, we use a Taylor expansion to lower-bound f around its minimum.

Lemma 13 *For $d \geq 3$ and $\delta \in [0,1)$.*

If $\beta \geq 3$ and $m^2 \leq \frac{\beta-3}{2(d+\beta-4)}$, then with probability at least $1 - d\exp\big(-\frac{\delta^2 n m^2}{2R^4 d^2}\big)$, for any symmetric matrix Δ:

$$f(V_*) - f(V_* + \Delta) > 2(1-\delta)m^2\|\Delta\|_F^2 + o(\|\Delta\|^2) \geq 0.$$

Otherwise with probability at least $1 - d\exp\big(-\frac{\delta^2 n\mu_1}{4R^4 d^2}\big)$, for any symmetric matrix Δ:

$$f(V_*) - f(V_* + \Delta) > (1-\delta)\mu_1\|\Delta\|_F^2 + o(\|\Delta\|^2) \geq 0,$$

with $\mu_1 \geq \frac{m^2(\beta-1)}{1+(d+\beta-2)m^2}$. Moreover we also have with probability at least $1 - d\exp\big(-\frac{\delta^2 n\mu_2}{4R^4 d^2}\big)$, for any symmetric matrix $\Delta \in \Delta_{\min}^{\perp}$:

$$f(V_*) - f(V_* + \Delta) > (1-\delta)\mu_2\|\Delta\|_F^2 + o(\|\Delta\|^2) \geq 0,$$

where $\mu_2 = \min\{2m^2, m^2(\beta-1), 2m\}$ and $\Delta_{\min} = \begin{pmatrix} 1 & 0 \\ 0 & c_{\min} I_{d-1}\end{pmatrix}$ is defined in the proof and satisfies

$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2 - 1|}.$$

This lemma directly implies Proposition 2.

Proof

For $\Delta \in \mathcal{S}(d)$ and $\delta \in \mathbb{R}$ we compute for $f(V) = \frac{1}{n}\sum_{i=1}^n \sqrt{x_i^\top V x_i}$,

$$\frac{d^2}{d\delta^2} f(V + \delta\Delta) = -\frac{1}{4n}\sum_{i=1}^n \frac{(x_i^\top \Delta x_i)^2}{\sqrt{x_i^\top (V+\delta\Delta) x_i}^3}.$$

Thus the second directional derivative in $V = V_*$ along Δ is

$$\nabla_\Delta^2 f(V_*) = \lim_{\delta\to 0}\frac{d^2}{d\delta^2} f(V+\delta\Delta) = -\frac{1}{4n}\sum_{i=1}^n (x_i^\top \Delta x_i)^2.$$

Let $\mathcal{T}_x$ be the semidefinite positive quadratic form of $\mathcal{S}(d)$ defined for $\Delta \in \mathcal{S}(d)$, by

$$\mathcal{T}_x : \Delta \mapsto (x^\top \Delta x)^2. \tag{33}$$

Then there exists a positive linear operator T_x from $\mathcal{S}(d)$ to $\mathcal{S}(d)$ such that $\mathcal{T}_x(\Delta) = \langle \Delta, T_x \Delta\rangle$.

Therefore the function f will be strictly concave if for all directions $\Delta \in \mathcal{S}(d)$

$$\frac{1}{n}\sum_{i=1}^n \mathcal{T}_{x_i}(\Delta) > 0. \tag{34}$$

We will bound the empirical expectation in Eq. (34) by first showing that its expectation remains away from 0. Then we will use a concentration inequality for matrices to control the distance between the sum in Eq. (34) and its expectation.

We first derive conditions so that the result is true in expectation, i.e., for the operator $\mathcal{T}$ defined by $\mathcal{T} = \mathbb{E}\mathcal{T}_x$ for x following the same law as $(y, z^\top)^\top$. We denote by $m = \mathbb{E}z^2$ and by $\beta = \mathbb{E}z^4/m^2$ its kurtosis.

We let $\Delta = \begin{pmatrix} a & b^\top \\ b & C \end{pmatrix}$ and then have $x^\top \Delta x = a + 2yb^\top z + z^\top C z$. Thus

$$\mathcal{T}_x(\Delta) = a^2 + 4ayb^\top z + 2az^\top Cz + 4b^\top(zz^\top)b + (z^\top Cz)^2 + 4yb^\top z(z^\top Cz).$$

Therefore we can express the value of the operator $\mathcal{T}$ only in function of the elements of Δ:

$$\mathcal{T}(\Delta) = (a + m\operatorname{tr} C)^2 + 4m\|b\|_2^2 + 2m^2\|C - \operatorname{Diag}(\operatorname{diag}(C))\|_F^2 + m^2(\beta - 1)\|\operatorname{diag}(C)\|^2,$$

where we have used

$$\begin{aligned}
\mathbb{E}(z^\top Cz)^2 &= \mathbb{E}\sum_{i,j,k,l} z_i z_j z_k z_l c_{i,j} c_{k,l} \\
&= \mathbb{E}\sum_i (z_i)^4 c_{i,i}^2 + \mathbb{E}\sum_{i,k\neq i} z_i^2 z_k^2 c_{i,i} c_{k,k} + 2\mathbb{E}\sum_{i,j\neq i} z_i^2 z_j^2 c_{i,j}^2 \\
&= \beta m^2 \sum_i c_{i,i}^2 + m^2 \sum_{i,k\neq i} c_{i,i} c_{k,k} + 2m^2 \sum_{i,j\neq i} c_{i,j}^2 \\
&= m^2(\beta - 3)\sum_i c_{i,i}^2 + m^2 \sum_{i,k} c_{i,i} c_{k,k} + 2m^2 \sum_{i,j} c_{i,j}^2 \\
&= m^2(\beta - 3)\|\operatorname{diag}(C)\|^2 + m^2\big(2\|C\|_F^2 + \operatorname{tr}(C)^2\big) \\
&= m^2(\beta - 3)\|\operatorname{diag}(C)\|^2 + m^2\big(2\|C - \operatorname{Diag}(\operatorname{diag}(C))\|_F^2 + \operatorname{tr}(C)^2\big).
\end{aligned}$$

Since $\beta \geq 1$, we get

$$\mathcal{T}(\Delta) \geq (a + m\operatorname{tr} C)^2 + 4m\|b\|_2^2 + 2m^2(\|C\|_F^2 - \|\operatorname{diag}(C)\|^2).$$

Thus $\mathcal{T}(\Delta) = 0$ if and only if $\beta = 1$ with $b = 0_{d-1}$ and $C = \operatorname{diag}(c)$ with $c^\top 1_d = -\frac{a}{m_2}$. With the condition $\beta = 1$ meaning that $\operatorname{var}(z^2) = 0$ and thus z^2 is constant a.s., i.e., z follows a Rademacher law.

However we would like to bound $\mathcal{T}(\Delta)$ away from zero by some constant and for that we are looking for the smallest eigenvalue of the operator $\mathbb{E}T_x$. Unfortunately we are not able to solve the optimization problem

$$\min_{\Delta \in \mathcal{S}(d), \|\Delta\|_F^2 = 1} \mathcal{T}(\Delta),$$

and we have to compute all the spectrum of this operator to be able to find the smallest using $\mathbb{E}T_x\Delta = 1/2\nabla\mathcal{T}(\Delta)$.

We have

$$1/2\nabla\mathcal{T}(\Delta) = \begin{pmatrix} a + m\operatorname{tr}(C) & 2mb^\top \\ 2mb & (a + m\operatorname{tr}(C))m_2 I_{d-1} + 2m^2 C \\ & +m^2(\beta - 3)\operatorname{Diag}(\operatorname{diag}(C)) \end{pmatrix}.$$

- For all $b \in \mathbb{R}^{d-1}$ we have for $\Delta = \begin{pmatrix} 0 & b^\top \\ b & 0 \end{pmatrix}$, $1/2\nabla\mathcal{T}(\Delta) = 2m\Delta$. Thus $2m$ is an eigenvalue of multiplicity $d-1$.
- For all $C \in \mathbb{R}^{(d-1)\times(d-1)}$ with $\operatorname{diag}(C) = 0_{d-1}$ we have for $\Delta = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$, $1/2\nabla\mathcal{T}(\Delta) = 2m^2\Delta$. Thus $2m^2$ is an eigenvalue of multiplicity $\frac{(d-1)(d-2)}{2}$.
- For all $c \in \mathbb{R}^{d-1}$ with $c^\top 1_{d-1} = 0$ we have for $\Delta = \begin{pmatrix} 0 & 0 \\ 0 & \operatorname{diag}(C) \end{pmatrix}$, $1/2\nabla\mathcal{T}(\Delta) = m^2(\beta-1)\Delta$. Thus $m^2(\beta-1)$ is an eigenvalue of multiplicity $d-2$.
- For all $a, c \in \mathbb{R}^2$ we have for $\Delta = \begin{pmatrix} a & 0 \\ 0 & cI_{d-1} \end{pmatrix}$,
$$\begin{aligned} 1/2\nabla\mathcal{T}(\Delta) &= \begin{pmatrix} a + m(d-1)c & 0 \\ 0 & [ma + m^2(d+\beta-2)c]I_{d-1} \end{pmatrix} \\ &= \operatorname{Diag}\Big[\begin{pmatrix} 1 & m1_{d-1}^\top \\ m1_{d-1} & (d+\beta-2)m^2 I_{d-1} \end{pmatrix}\begin{pmatrix} a \\ c1_{d-1} \end{pmatrix}\Big]. \end{aligned}$$
Thus an eigenvalue of $\begin{pmatrix} 1 & (d-1)m \\ m & (d+\beta-2)m^2 \end{pmatrix}$ with an eigenvector $[a, c]^\top$ would be an eigenvalue of the operator $\mathbb{E}T_x$ with a corresponding eigenvector $\begin{pmatrix} a & 0 \\ 0 & cI_{d-1} \end{pmatrix}$. This matrix has two simple eigenvalues
$$\mu_\pm = \frac{1 + (d+\beta-2)m^2 \pm \sqrt{(1 + (d+\beta-2)m^2)^2 - 4m^2(\beta-1)}}{2}. \tag{35}$$

Moreover when we add all the multiplicity of the found eigenvalues we get $d-1+\frac{(d-1)(d-2)}{2}+d-2+2 = \frac{d(d+1)}{2}$ which is the dimension of $S(d)$, therefore we have found all the eigenvalues of the linear operator $\mathbb{E}T_x$.

We will prove now than the smallest eigenvalue is μ_- when the dimension d is large enough with regards to m^2 and $2m^2$ otherwise.

Lemma 14 *Let μ_1 and μ_2 be the two smallest eigenvalues of the operator $\mathbb{E}T_x$. Let us assume that $d \geq 3$ (the case $d = 2$ will also be done in the proof).*

If $\beta \geq 3$ and $m^2 \leq \frac{\beta-3}{2(d+\beta-4)}$ then
$$\mu_1 = 2m^2.$$

Otherwise
$$\mu_1 = \mu_- \geq \frac{m^2(\beta-1)}{1 + (d+\beta-2)m^2} \text{ and } \mu_2 = \min\{2m^2, m^2(\beta-1), 2m\}.$$

Moreover we denote by $\Delta_{\min} = \begin{pmatrix} 1 & 0 \\ 0 & c_{\min}I_{d-1} \end{pmatrix}$ the eigenvector associated to μ_- for which we have set without loss of generality the first component $a = 1$. Then
$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2 - 1|}.$$

Unfortunately μ_- can become small when the dimension increases as explained by the tight bound $\mu_- \geq \frac{m^2(\beta-1)}{1+(d+\beta-2)m^2}$. However the corresponding eigenvector has a particular structure we will be able to exploit.

Proof First we note that $\mu_- \leq m^2(\beta-1)$ and compute

$$\begin{aligned}
\mu_- \geq 2m^2 \quad &\Leftrightarrow \quad 1+(d+\beta-2)m^2 - \sqrt{(1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1)} - 4m^2 \geq 0 \\
&\Leftrightarrow \quad 1+(d+\beta-2)m^2 - 4m^2 \geq \sqrt{(1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1)} \\
&\Leftrightarrow \quad (1+(d+\beta-2)m^2 - 4m^2)^2 \geq (1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1) \\
&\qquad\;\; \text{and } 1+(d+\beta-6)m^2 \geq 0 \\
&\Leftrightarrow \quad 16m^4 - 8m^2(1+(d+\beta-2)m^2) \geq -4m^2(\beta-1) \\
&\qquad\;\; \text{and } 1+(d+\beta-6)m^2 \geq 0 \\
&\Leftrightarrow \quad \underbrace{2(d+\beta-4)m^2 \leq \beta-3}_{\text{R1}} \text{ and } \underbrace{1+(d+\beta-6)m^2 \geq 0}_{\text{R2}}.
\end{aligned}$$

- If $d=2$,
  - If $\beta \leq 3$ we have necessary that $\beta \leq 2$ and the first term R1 implies $m^2 \geq \frac{3-\beta}{2(2-\beta)}$ and the second term R2 implies $m^2 \leq 1/(4-\beta)$. Thus we should have $(4-\beta)(3-\beta) \leq 2(2-\beta)$ which is not possible since the polynomial $\beta^2-5\beta+8 \geq 0$.
  - If $\beta \geq 3$, the first term R1 implies $m^2 \leq \frac{\beta-3}{2(\beta-2)} \leq 1$ and the second term R2 implies $m^2 \leq 1/(4-\beta) \leq \frac{\beta-3}{2(\beta-2)} \leq 1$ for $\beta \leq 4$ and is always satisfied otherwise.
- If $d \geq 3$, the first term R1 implies that $\beta \geq 3$ for which the second term R2 is always satisfied. It also implies that $m^2 \leq \frac{\beta-3}{2(d+\beta-4)} \leq 1$.

We denote by $\Delta_{\min} = \begin{pmatrix} 1 & 0 \\ 0 & c_{\min} I_{d-1} \end{pmatrix}$ the eigenvector for which we have set without loss of generality $a=1$ and

$$c_{\min} = \frac{-1}{2(d-1)m}\Big[\sqrt{((d+\beta-2)m^2-1)^2 + 4(d-1)m^2} - (d+\beta-2)m^2 + 1\Big].$$

Consequently $c_{\min} \leq 0$ and by convexity of the square root we have

$$\sqrt{((d+\beta-2)m^2-1)^2 + 4(d-1)m^2} \leq ((d+\beta-2)m^2-1) + \frac{2(d-1)m^2}{|(d+\beta-2)m^2-1|}.$$

Therefore

$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2-1|}.$$

■

We will control now the behavior of the empirical expectation by its expectation thanks to concentration theory. By definition T_x is a symmetric positive linear operator as its projection $T_x^\perp$ onto the orthogonal space of $\Delta_{\min}$. We can thus apply the Matrix Chernoff inequality from Tropp (2012, Theorem 5.1.1) to these two operators using $\|T_x\|_{op} \leq$

$\|xx^\top\|^2 \leq \mathrm{tr}(xx^\top)^2 \leq \|x\|_2^4 \leq R^4 d^2$. Then:

$$\mathbb{P}\bigg(\lambda_{\min}\Big(\sum_{k=1} T_{x_k}\Big) \leq n\delta\mu_1\bigg) \leq d\Big[\frac{e^{-(1-\delta)}}{\delta^\delta}\Big]^{n\mu_1/(2R^4d^2)} \leq de^{-(1-\delta)^2 n\mu_1/(4R^4d^2)},$$

$$\mathbb{P}\bigg(\lambda_{\min}\Big(\sum_{k=1} T^\perp_{x_k}\Big) \leq n\delta\mu_2\bigg) \leq d\Big[\frac{e^{-(1-\delta)}}{\delta^\delta}\Big]^{n\mu_2/(2R^4d^2)} \leq de^{-(1-\delta)^2 n\mu_2/(4R^4d^2)},$$

For $m=1$ and $d \geq 3$ we have $\mu_1 = \mu_- \geq \frac{\beta-1}{\beta+d} \geq \min\{\frac{\beta-1}{2\beta}, \frac{\beta-1}{2d}\} \geq \min\{1/3, \frac{\beta-1}{2d}\}$. ■

D.5 Noise robustness for the one dimensional balanced problem

We want a condition on ε such that the solution of the relaxation Eq. (8) recovers the right y. We recall the dual problem of the relaxation Eq. (8)

$$\min \mu^\top 1_n \text{ s.t. } \mathrm{Diag}(\mu) \succcurlyeq X(X^\top X)^{-1}X^\top.$$

The KKT conditions are:

- Dual feasibility: $\mathrm{Diag}(\mu) \succcurlyeq X(X^\top X)^{-1}X^\top$.

- Primal feasibility: $\mathrm{Diag}(Y) = 1_n$ and $Y \succcurlyeq 0$.

- Complimentary slackness : $Y[\mathrm{Diag}(\mu) - X(X^\top X)^{-1}X^\top] = 0$

For $Y = yy^\top$ a rank one matrix, the last condition implies $\mathrm{Diag}(\mu)y = Hy$ and

$$\mu_i = \frac{(X(X^\top X)^{-1}X^\top y)_i}{y_i}.$$

For $X = y+\varepsilon$, we denote by $\tilde{y} = y+\varepsilon$, then $X(X^\top X)^{-1}X^\top = \frac{\tilde{y}\tilde{y}^\top}{\|\tilde{y}\|^2}$ and $X(X^\top X)^{-1}X^\top y = \frac{\tilde{y}^\top y}{\|\tilde{y}\|^2}\tilde{y}$. Thus

$$\mu_i = \frac{\tilde{y}^\top y}{\|\tilde{y}\|^2}\frac{\tilde{y}_i}{y_i}.$$

Assume that all $\tilde{y}_i y_i$ have the same sign, without loss of generality we assume $\tilde{y}_i y_i > 0$. By definition of μ, $\mu \geq 0$. To show the dual feasibility we have to show that $\mathrm{Diag}(\mu) \succcurlyeq H$ which is equivalent to $\mathrm{Diag}(\frac{\tilde{y}_i}{y_i}) \succcurlyeq \frac{\tilde{y}\tilde{y}^\top}{\tilde{y}^\top y}$, to $I_n - \mathrm{Diag}(\sqrt{\frac{y_i}{\tilde{y}_i}})\frac{\tilde{y}\tilde{y}^\top}{\tilde{y}^\top y}\mathrm{Diag}(\sqrt{\frac{y_i}{\tilde{y}_i}}) \succcurlyeq 0$ and to $\sum y_i\tilde{y}_i \leq \tilde{y}^\top y$ which is obviously true. Reciprocally if μ is dual feasible then $\mathrm{Diag}(\mu) \succcurlyeq 0$ and all the $\tilde{y}_i y_i$ have the same sign.

Therefore we have shown that y is solution of the relaxation Eq. (8) if and only if all the $\tilde{y}_i y_i$ have the same sign. If ε and y are independent this is equivalent to $\|\varepsilon\|_\infty \leq 1$ a.s.

D.6 The rank-one candidates are not solutions of the relaxation Eq. (12)

We assume now that $1_n^\top y \neq 0$ thus $y \neq \Pi_n y$, which means we do not have the same proportion in the two clusters. Let us assume that $\Pi_n y$ takes two values $\{\pi y_-, \pi y_+\}$ that is by definition of Π_n: $\pi y_+ = 1 - \frac{1_n^\top y}{n}$ and $\pi y_- = -1 - \frac{1_n^\top y}{n}$. For V_* defined as before, we get $x_i^\top V_* x_i = (\pi y_i)^2$ and with $I_\pm$ the set of indices such that $\Pi_n y_i = \pi y_\pm$ respectively, the KKT conditions for $V = V_*$ can be written as

$$\frac{1}{n}\Big[\sum_{i \in I_+}\Big(\frac{1}{\pi y_+} - 1\Big)x_i x_i^\top + \sum_{i \in I_-}\Big(\frac{1}{-\pi y_-} - 1\Big)x_i x_i^\top\Big] = A_n \preccurlyeq 0 \text{ and } A_n V_* = 0.$$

We check that with $n_\pm = \#\{I_\pm\}$:

$$\begin{aligned} w^\top A_n w = 0 &= \sum_{i \in I_+}\Big(\frac{1}{\pi y_+} - 1\Big)(\pi y_+)^2 + \sum_{i \in I_-}\Big(\frac{1}{-\pi y_-} - 1\Big)(\pi y_-)^2 \\ &= n_+\Big(\frac{1}{\pi y_+} - 1\Big)(\pi y_+)^2 + n_-\Big(\frac{1}{-\pi y_-} - 1\Big)(\pi y_-)^2 \\ &= n_+ \pi y_+ - n_- \pi y_- - \big(n_+(\pi y_+)^2 + n_-(\pi y_-)^2\big) \\ &= y^\top \Pi_n y - (\Pi_n y)^\top \Pi_n y = y^\top \Pi_n y - y^\top \Pi_n y = 0. \end{aligned}$$

And $A_n = \frac{1}{2n}\big[\sum_{i \in I_+} \alpha_+ x_i x_i^\top + \sum_{i \in I_-} \alpha_- x_i x_i^\top\big]$ with $\alpha_+ = \big(\frac{1}{\pi y_+} - 1\big)$ and $\alpha_- = \big(\frac{1}{-\pi y_-} - 1\big)$. Unfortunately $\alpha_+ \alpha_- \leq 0$, and A_n is not necessary negative. Even worse we will show that $\mathbb{E}A$ is not semi-definite negative which will conclude the proof since by the law of large number $\lim_{n \to \infty} \frac{1}{n} A_n = \mathbb{E}A$. Assume that the proportions of the two clusters stay constant with $n_\pm = \rho_\pm n$, then

$$\mathbb{E}A = \rho_+ \alpha_+ \begin{pmatrix} (\pi y_+)^2 & 0 \\ 0 & I \end{pmatrix} + \rho_- \alpha_- \begin{pmatrix} (\pi y_-)^2 & 0 \\ 0 & I \end{pmatrix}.$$

And $\rho_+ \alpha_+ (\pi y_+)^2 + \rho_- \alpha_- (\pi y_-)^2 = 0$ since $w^\top A_n w = 0$. Then

$$\begin{aligned} \rho_+ \alpha_+ + \rho_- \alpha_- &= \frac{\rho_+ \pi y_- - \rho_- \pi y_+ - \pi y_+ \pi y_-}{\pi y_+ \pi y_-} \\ &= \frac{-(\rho_+ + \rho_-) - \frac{1_n^\top y}{n}(\rho_+ - \rho_-) + (1 - (1_n^\top y)^2)}{-(1 - (\frac{1_n^\top y}{n})^2)} \\ &= \frac{\frac{1_n^\top y}{n}(\rho_+ - \rho_-) + (\frac{1_n^\top y}{n})^2)}{(1 - (\frac{1_n^\top y}{n})^2)} = \frac{2(\frac{1_n^\top y}{n})^2}{(1 - (\frac{1_n^\top y}{n})^2)} \geq 0. \end{aligned}$$

Thus $A = \frac{2(1_n^\top y)^2}{(n^2 - (1_n^\top y)^2)} \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$ is not semi-definite negative and V_* is not solution of the relaxation Eq. (12).

Appendix E. Auxiliary results for sparse extension

In this section, we provide some results for sparse extension.

E.1 There is a rank-one solution of the relaxation Eq. (18)

Lemma 15 *The rank-one solution $V_* = v^* {v^*}^\top$ is solution of the relaxation Eq. (18) if the design matrix X is such that $\frac{1}{n}X^\top X$ has all its diagonal entries less than one.*

Proof The KKT conditions for problem Eq. (18) are

$$\frac{1}{n}\sum_{i=1}^{n}\frac{x_i x_i^\top}{\sqrt{x_i^\top V x_i}} - \lambda U - \frac{1}{n}X^\top X = A \preccurlyeq 0 \text{ and } AV = 0,$$

with U such that $U_{ij} = \operatorname{sign}(V_{ij})$ if $V_{ij} \neq 0$ and $U_{ij} \in [-1, 1]$ otherwise. For $V_* = v^* {v^*}^\top$ this gives

$$A = \frac{(1+\lambda)}{n}X^\top X - \lambda U - \frac{1}{n}X^\top X = \lambda\Big[\frac{X^\top X}{n} - U\Big] \text{ with } U_{1,1} = 1 \text{ and } U_{i,j} \in [-1,1] \text{ otherwise.}$$

We check that $AV_* = 0$. If the design matrix X is such that $\frac{1}{n}X^\top X$ has all its diagonal entries less than one, we can choose a sub-gradient U such that the dual variable $A = 0$ and thus V_* is solution. Otherwise by property of semi-definite matrices, there is a diagonal entry of $\frac{1}{n}X^\top X$ which is bigger than 1 which prevents A to be semi-definite negative since the corresponding diagonal entry of $\frac{X^\top X}{n} - U$ will be positive. This shows that V_* does not solve the problem. ■

E.2 Proof of proposition 6

Lemma 16 *For $\delta \in [0, 1)$, with probability $1 - 5d^2 \exp\big(-\frac{\delta^2 n(\beta-1)}{2dR^4(1/m^2+\beta+d)}\big)$, for any direction Δ such that $V_* + \Delta \succcurlyeq 0$, we have:*

$$g(V_*) - g(V_* + \Delta) > (1-\delta)\Big[\lambda\|\Delta - \operatorname{Diag}(\Delta)\|_1 + \frac{\beta-1}{\beta+d+1/m^2}\frac{(1+\lambda)^3}{4}\|\operatorname{Diag}(\Delta)\|_2^2\Big] + o(\|\Delta\|^2) \geq 0.$$

Moreover we also have with probability at least $1 - 5d^2 \exp\big(-\frac{\delta^2 n m^2(\beta-1)}{2dR^4}\big)$, for any symmetric matrix Δ such that $V_* + \Delta \succcurlyeq 0$ and $\operatorname{Diag}(\Delta) \in (e_{\min})^\perp$:

$$g(V_*) - g(V_* + \Delta) > (1-\delta)\Big[\lambda\|\Delta - \operatorname{Diag}(\Delta)\|_1 + m^2(\beta-1)\frac{(1+\lambda)^3}{4}\|\operatorname{Diag}(\Delta)\|_2^2\Big] + o(\|\Delta\|^2) \geq 0,$$

where $e_{\min} = [1, c_{\min}1_{d-1}]$ is defined in the proof and $c_{\min}$ satisfies

$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2 - 1|}.$$

E.2.1 Proof outline

We will investigate under which conditions on X the solution is unique, first for a deterministic design matrix. We make the following deterministic assumptions on X for $\delta, \zeta \geq 0$ and $\mathcal{S} \subset \mathbb{R}^d$:

(A1) $\|\frac{X^\top X}{n}\|_\infty \leq 1$ **(A3)** $\|\frac{Z^\top Z}{n} - \operatorname{Diag}(\operatorname{diag}(\frac{1}{n}Z^\top Z))\|_\infty \leq \delta$
(A2) $\|\frac{Z^\top y}{n}\|_\infty \leq \delta$ **(A4)** $\lambda_{\min}^{\mathcal{S}}(\frac{X^{\odot 2}(X^{\odot 2})^\top}{n}) \geq \zeta > 0$,

where we denoted by $\odot$ the Hadamard (i.e., pointwise) product between matrices and $\lambda_{\min}^{\mathcal{S}}$ the minimum eigenvalue of a linear operator restricted to a subspace $\mathcal{S}$. Then with $g(V) = \frac{2}{n}\sum_{i=1}^n \sqrt{x_i^\top V x_i} - \lambda\|V\|_1 - \frac{1}{n}\operatorname{tr} X^\top X V$, we can certify that g will decrease around the solution V_*.

Lemma 17 *Let us assume that the noise matrix verifies assumptions (A1,A2,A3,A4), then for any direction* Δ *such that* $V_* + \Delta \succcurlyeq 0$ *and* $\operatorname{diag}(\Delta) \in \mathcal{S}$ *we have:*

$$g(V_*) - g(V_* + \Delta) \geq \lambda(1-\delta)\|\Delta - \operatorname{Diag}(\operatorname{diag}(\Delta))\|_1 + \zeta\frac{(1+\lambda)^3}{4}\|\operatorname{Diag}(\Delta)\|_2^2 + o(\|\Delta\|^2) > 0.$$

Let us assume now that $(z^i)_{i=1,.,d}$ are i.i.d of law z symmetric with $\mathbb{E}z = \mathbb{E}z^3 = 0$, $\mathbb{E}z^2 = m = 1$, $\mathbb{E}z^4/(\mathbb{E}z^2)^2 = \beta$ and such that $\|z\|_\infty$ is a.s. bounded by $0 \leq R \leq 1$. Then the matrix X satisfies a.s. assumption (A1). Using multiple Hoeffding's inequalities we will prove lemma 17.

Lemma 18 *If* z *does not follow a Rademacher law, the design matrix* X *satisfies assumptions (A1,A2,A3,A4) with probability greater than* $1 - 8d^2\exp\big(-\frac{\delta^2 n(\beta-1)}{2d(\beta+d)R^4}\big)$ *for* $\mathcal{S} = \mathbb{R}^d$, *and with probability greater than* $1 - 8d^2\exp\big(-\frac{\delta^2 n\min\{\beta-1,2\}}{2dR^4}\big)$ *for* $\mathcal{S} = [1, c_{\min}1_{d-1}]^\perp$ *where* c_{min} *is defined in the proof and satisfies*

$$|e_{\min}| \leq \frac{1}{d+\beta-3}.$$

This lemma concludes the proof of proposition 6. We will now prove these two lemmas.

E.2.2 Proof of lemma 17

Proof Since the dual variable A for the PSD constraint is 0 (see the proof of lemma 15), this constraint $V \succcurlyeq 0$ is not active and we will show that the function decreases in a set of directions Δ which include the one for which $V_* + \Delta \succcurlyeq 0$.

Therefore we consider a direction $\Delta = \begin{pmatrix} a & b^\top \\ b & C \end{pmatrix}$, with $C \succcurlyeq 0$, which is slightly more general than $V_* + \Delta \succcurlyeq 0$. We denote by $f(W) = \frac{2}{n}\sum_{i=1}^n \sqrt{x_i^\top W x_i} - \frac{1}{n}\operatorname{tr} X^\top X W$ the smooth part of g. By Taylor-Young, we have for all W:

$$f(W) - f(W+\Delta) = -\langle f'(W), \Delta\rangle - \frac{1}{2}\langle \Delta, f''(W)\Delta\rangle + o(\|\Delta\|^2).$$

Thus:

$$g(W) - g(W+\Delta) = -\langle f'(W), \Delta\rangle - \frac{1}{2}\langle \Delta, f''(W)\Delta\rangle + \lambda(\|W+\Delta\|_1 - \|W\|_1) + o(\|\Delta\|^2).$$

Letting $W = V_*$ this gives with $X^\top X = \begin{pmatrix} n & y^\top Z \\ Z^\top y & Z^\top Z \end{pmatrix}$,

$$\begin{aligned} g(W) - g(W+\Delta) &= -\lambda \langle \frac{X^\top X}{n}, \Delta \rangle - \frac{1}{2} \langle \Delta, f''(V_*)\Delta \rangle + \lambda(a + 2\|b\|_1 + \|C\|_1) + o(\|\Delta\|^2) \\ &= \lambda \big[2(\|b\|_1 - \frac{1}{n} b^\top Z^\top y) + \|C\|_1 - \frac{1}{n} \operatorname{tr}(Z^\top Z C) \big] - \frac{1}{2} \langle \Delta, f''(V_*)\Delta \rangle + o(\|\Delta\|^2). \end{aligned}$$

And with Hölder's inequality and assumption (A2)

$$\|b\|_1 - \frac{1}{n} b^\top Z^\top y \geq \|b\|_1 (1 - \|\frac{1}{n} Z^\top y\|_\infty) \geq (1-\delta)\|b\|_1.$$

Nevertheless we will show in lemma 19 that $\|C\|_1 - \frac{1}{n}\operatorname{tr}(Z^\top Z C) \geq (1-\delta)\|C - \operatorname{diag}(C)\|_1$, thus

$$g(W) - g(W+\Delta) \geq \lambda(1-\delta)(2\|b\|_1 + \|C - \operatorname{diag}(C)\|_1) + o(\|\Delta\|^2). \tag{36}$$

However in Eq. (36), $g(W) - g(W+\Delta) = 0$ for $b = 0$ and C diagonal, therefore we have to investigate second order conditions, i.e., to show for $\Delta = \operatorname{diag}(e)$ with $e \in \mathbb{R}^d$ that $-\langle \Delta, f''(V_*)\Delta \rangle > 0$.

And with assumption (A4)

$$\begin{aligned} -\frac{4}{(1+\lambda)^3} \langle \operatorname{diag}(e), f''(V_*) \operatorname{diag}(e) \rangle &= \frac{1}{n} \sum_{i=1}^n (x_i^\top \operatorname{diag}(e) x_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\sum_{j=1}^d e_j (x_i^j)^2)^2 \\ &= \frac{1}{n} \sum_{i=1}^n e^\top [x_i^{\odot 2} (x_i^{\odot 2})^\top] e \\ &\geq \lambda_{\min} (\frac{X^{\odot 2} (X^{\odot 2})^\top}{n}) \|e\|^2 \geq \zeta \|e\|_2^2. \end{aligned}$$

Thus we can conclude:

$$g(W) - g(W+\Delta) \geq \lambda(1-\delta)(2\|b\|_1 + \|C - \operatorname{diag}(C)\|_1) + \zeta \frac{(1+\lambda)^3}{4} \|e\|_2^2 + o(\|\Delta\|^2).$$

■

E.2.3 Auxiliary lemma

Lemma 19 *For all matrix C symmetric semi-definite positive we have under assumptions (A1) and (A3):*

$$\operatorname{tr}\Big(S - \frac{Z^\top Z}{n}\Big) C \geq (1-\delta)\|C - \operatorname{diag}(C)\|_1 > 0.$$

Proof We denote by $\Sigma^n = \frac{Z^\top Z}{n}$. We always have $\|C\|_1 - \mathrm{tr}(\Sigma^n C) = \mathrm{tr}(S - \Sigma^n)C$ where $S_{i,j} = \mathrm{sign}(C_{i,j})$, thus if $\mathrm{diag}(C) > 0$ then $\mathrm{diag}(S) = 1$ and $\mathrm{diag}(S - \Sigma^n) \geq 0$ from assumption (A1). Moreover since ${\Sigma^n}_{i,j} \in [-1, 1]$ then $\mathrm{sign}(S - \Sigma^n) = \mathrm{sign}(S)$.

Thus $\mathrm{tr}(S-\Sigma^n)C = \sum_i C_{i,i}(S-\Sigma^n)_{i,i} + \sum_{i\neq j} C_{i,j}(S-\Sigma^n)_{i,j} \geq \sum_{i\neq j} C_{i,j}(S-\Sigma^n)_{i,j} \geq 0$. Furthermore from assumption (A3), $|(\Sigma^n)_{i,j}| \leq \delta$ for $i \neq j$. Therefore

$$\mathrm{tr}(S - \Sigma^n)C \geq \sum_{i\neq j} C_{i,j}(S - \Sigma^n)_{i,j} \geq \sum_{i\neq j} |C_{i,j}|(1-\delta) \geq (1-\delta)\|C - \mathrm{diag}(C)\|_1 > 0.$$

If there is a diagonal element of C which is 0, then the corresponding row and column in C will also be 0 and we can look at the same problem as before by erasing off from C and Σ^n the corresponding column and row. ■

E.2.4 Proof of lemma 18

Proof We will first show that the noise matrix Z satisfies assumptions (A2,A3). By Hoeffding's inequality we have with probability $1 - 2\exp(-\delta^2 n/(2R^2))$

$$\frac{1}{n}|\sum_{i=1}^n z_i^j| \leq \delta.$$

Then, since the law of z is symmetric $y_i z_i$ will have the same law as z_i and with probability $1 - 2\exp(-\delta^2 n/(2R^2))$, the design matrix Z satisfies assumption (A2):

$$\|\frac{Z^\top y}{n}\|_\infty \leq \delta.$$

Likewise we have with probability $1 - 2\exp(-\delta^2 n/(2R^4))$ that for $j \neq j'$

$$|\frac{1}{n}\sum_{i=1}^n z_i^j z_i^{j'}| \leq \delta.$$

Thus we also have with probability $1 - 2d^2\exp(-\delta^2 n/(2R^4))$ that Z satisfies assumption (A3):

$$\|\frac{1}{n}Z^\top Z - \mathrm{diag}(\frac{1}{n}Z^\top Z)\|_\infty \leq \delta.$$

Thus with probability $1 - 4d^2\exp(-\delta^2 n/(2R^4))$, the noise matrix Z satisfies assumptions (A1, A2, A3).

We proceed as in the proof of proposition 2 to show that X satisfies assumption (A4). We first derive a condition to have the result in expectation, then we use an inequality concentration on matrix to bound the empirical expectation. This will be very similar, but we will get a better scaling since Δ is diagonal.

Using the same arguments as in the proof of proposition 2 we have for the diagonal matrix $\Delta = \mathrm{diag}(e)$ with $e = (a, c) \in \mathbb{R}^d$:

$$e^\top \mathbb{E}(x^{\odot 2}(x^{\odot 2})^\top)e = \mathbb{E}(x^\top \Delta x)^2 = (a + mc^\top 1_{n-1})^2 + m^2(\beta - 1)\|c\|_2^2 > 0 \ \text{ if } \ \beta > 1.$$

We can show that $m^2(\beta-1)$ is an eigenvalue of multiplicity $d-2$ and $\mu_\pm$ are eigenvalues of multiplicity one of the operator $\Delta \mapsto \mathbb{E}(x^\top \Delta x)^2$ with eigenvectors $e_\pm$. Thus we have

$$\begin{aligned} \lambda_{\min}(\mathbb{E}x^{\odot 2}(x^{\odot 2})^\top) &= \frac{1+(d+\beta-2)m^2 - \sqrt{(1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1)}}{2} \qquad (37)\\ &\geq \frac{m^2(\beta-1)}{1+(d+\beta-2)m^2}, \end{aligned}$$

and

$$\lambda_{\min}^{e_-^\perp}(\mathbb{E}x^{\odot 2}(x^{\odot 2})^\top) = m^2(\beta-2).$$

Moreover

$$\lambda_{\max}\Big(x^{\odot 2}(x^{\odot 2})^\top\Big) = (x^{\odot 2})^\top x^{\odot 2} = \sum_{j=1}^d (x_i)^4 \leq dR^4.$$

Thus we can apply the Matrix Chernoff inequality from (Tropp, 2012) for $\mu_{\mathcal{S}} = \lambda_{\min}^{\mathcal{S}}(\mathbb{E}x^{\odot 2}(x^{\odot 2})^\top)$:

$$\mathbb{P}\Bigg(\lambda_{\min}^{\mathcal{S}}\Big(\frac{X^{\odot 2}(X^{\odot 2})^\top}{n}\Big) \leq (1-\delta)\mu_{\mathcal{S}}\Bigg) \leq de^{-\delta^2 n\mu_{\mathcal{S}}/(2dR^4)}.$$

Thus with probability $1-5d^2\exp(-\delta^2 n\mu_-/(2dR^4))$ the design matrix X satisfies assumptions (A1,A2,A3,A4) with $\zeta=(1-\delta)\mu_-$ and $\mathcal{S}=\mathbb{R}^d$. And with probability $1-5d^2\exp(-\delta^2 n\min\{\beta-1,2\}/(2dR^4))$ the design matrix X satisfies assumptions (A1,A2,A3,A4) with $\zeta=(1-\delta)\min\{\beta-1,2\}$ and $\mathcal{S}=e_-^\perp$. ■

Appendix F. Proof of multi-label results

We first prove lemma 7:

Proof Let $A\in\mathbb{R}^{k\times k}$ symmetric semi-definite positive such that $\operatorname{diag}(\tilde{y}A\tilde{y}^\top)=1_n$, then

$$\operatorname{diag}(\tilde{y}A\tilde{y}^\top) = \sum_{i=0}^k a_{i,i}1_n + 2\sum_{i=1}^k a_{0,i}y_i + 2\sum_{1\leq i<j\leq k} a_{i,j}y_i\odot y_j$$

thus

$$2\sum_{i=1}^k a_{0,i}y_i + 2\sum_{1\leq i<j\leq k} a_{i,j}y_i\odot y_j = (1-\sum_{i=0}^k a_{i,i})1_n$$

And this system admits as unique solution 0_n if and only if the family $\{1_n,(y_i)_{1\leq i\leq k},(y_iy_j)_{1\leq i<j\leq k}\}$ is *linearly independent.* ■

Then we prove the lemma 8:

Proof Since $a_0+\sum_{i=1}^k a_i^2\alpha_i \geq \alpha_{\min}\sum_{i=0}^k a_i^2 = \alpha_{\min}$ we should have $\alpha\geq\alpha_{\min}$. We have already seen that such Y satisfies the constraint. The KKT conditions are: $B=$

$\mathrm{diag}(\mu) - H - \nu 11^\top \succcurlyeq 0$ and $BY = 0$. Since $y_i = \Pi_n y_i + \frac{(y_i^\top 1_n)}{n} 1_n$.

$$\begin{aligned} Hy_i &= H\Pi_n y_i + (y_i^\top 1_n) H 1_n \\ &= \Pi_n y \\ &= (y_i - \frac{1_n^\top y_i}{n} 1_n). \end{aligned}$$

Thus

$$\begin{aligned} HY &= \sum_{i=1}^{k} a_i^2 H y_i y_i^\top \\ &= \sum_{i=1}^{k} a_i^2 (y_i - \frac{1_n^\top y_i}{n} 1_n) y_i^\top \\ &= \sum_{i=1}^{k} a_i^2 (y_i y_i^\top - \frac{1_n^\top y_i}{n} 1_n y_i^\top) \end{aligned}$$

and $\mathrm{tr}(HY) = \sum_{i=1}^k a_i^2 (n - n\alpha_i) = n(1 - a_0^2 + a_0^2 - \alpha) = n(1-\alpha)$.

Furthermore since $1_n^\top \mathrm{diag}(Y) = n$ and $1_n^\top M 1_n = n^2 \alpha$, for $\mu = 1_n$ and $\nu = 1/n$, $B.Y = n - n(1-\alpha) - n\alpha = 0$. And since $B = I_n - \frac{1}{n} 1_n 1_n^\top - H$, $B^2 = B$ and $B^\top = B$, thus B is a symmetric projection and consequently symmetric semi-definite positive.

Hence the primal variable Y and the dual variables $\mu = 1_n$ and $\nu = 1/n$ satisfy the KKT conditions, thus Y is solution of this problem. ■

Appendix G. Efficient optimization problem

We now give the details of an efficient optimization algorithm.

G.1 Dual computation

We consider the following strongly-convex approximation of Eq. (24), augmented with the von-Neumann entropy:

$$\max_{V \succcurlyeq 0} \frac{1}{n} \sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \| \mathrm{Diag}(c) V \mathrm{Diag}(c) \|_1 - \varepsilon \, \mathrm{tr}[(A^{\frac{1}{2}} V A^{\frac{1}{2}}) \log(A^{\frac{1}{2}} V A^{\frac{1}{2}})] \text{ s.t. } \mathrm{tr}(A^{\frac{1}{2}} V A^{\frac{1}{2}}) = 1.$$

Introducing dual variables, we have

$$\begin{aligned} \min_{u \in \mathbb{R}_+^n, C : |C_{ij}| \leqslant c_i c_j} \quad \max_{V \succcurlyeq 0} \quad & \frac{1}{2n} \sum_{i=1}^n \Big(u_i ((XVX^\top)_{ii}) + \frac{1}{u_i} \Big) - \mathrm{tr}\, CV - \varepsilon \, \mathrm{tr}[(A^{\frac{1}{2}} V A^{\frac{1}{2}}) \log(A^{\frac{1}{2}} V A^{\frac{1}{2}})] \\ \text{s.t.} \quad & \mathrm{tr}(A^{\frac{1}{2}} V A^{\frac{1}{2}}) = 1. \end{aligned}$$

By fixing u and C, and letting $Q = A^{\frac{1}{2}} V A^{\frac{1}{2}}$, we can write the max problem as

$$\begin{aligned} \max_{Q \succcurlyeq 0} \quad & \mathrm{tr}\, A^{-\frac{1}{2}} (\frac{1}{2n} X^\top \mathrm{Diag}(u) X - C) A^{-\frac{1}{2}} Q - \varepsilon \, \mathrm{tr}[Q \log(Q)] \\ \text{s.t.} \quad & \mathrm{tr}\, Q = 1. \end{aligned}$$

This problem is of the form

$$\max_{Q \succcurlyeq 0} \operatorname{tr} DQ - \varepsilon \sum_{i=1}^{n} \sigma_i(Q) \log \sigma_i(Q)$$
$$\text{s.t. } \operatorname{tr} Q = 1$$

where $D = A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}}$ and $\sigma_i(Q)$ denotes the i-th largest eigen value of the matrix Q. If we consider the matrix D to be of the form $D = U \operatorname{Diag}(\theta)U^\top$ with θ denoting the vector of ordered eigen values of D, then it turns out that at optimality Q has the form $Q = U \operatorname{Diag}(\sigma)U^\top$, with σ denoting the ordered vector of eigen values of Q.

Therefore the above optimization problem can be cast in terms of σ as:

$$\max_{\sigma \in \mathbb{R}^n} \theta^\top \sigma - \varepsilon \sum_{i=1}^{n} \sigma_i \log \sigma_i$$
$$\text{s.t. } \sum_{i=1}^{n} \sigma_i = 1.$$

The solution of this problem is $\sigma_i = \frac{e^{\theta_i/\varepsilon}}{\sum_{j=1}^n e^{\theta_j/\varepsilon}}$, which leads to

$$\min_{\theta \in \mathbb{R}^n} \phi^\varepsilon(\theta) = \varepsilon \log \sum_{i=1}^{n} \left(e^{\frac{\theta_i}{\varepsilon}}\right).$$

In terms of the original matrix variables, we have

$$\min \phi^\varepsilon(D) = \varepsilon \log \operatorname{tr} e^{\frac{D}{\varepsilon}}.$$

Using the appropriate expansion of D, we have the overall optimization problem as

$$\min_{u \in \mathbb{R}^n_+, C : |C_{ij}| \leqslant c_i c_j} \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{u_i} + \phi^\varepsilon(A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}}). \tag{38}$$

At optimality, we have

$$A^{\frac{1}{2}}VA^{\frac{1}{2}} = \left(e^{\frac{(A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}})}{\varepsilon}}\right) / \operatorname{tr}\left(e^{\frac{(A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}})}{\varepsilon}}\right).$$

The error of approximation is at most $\varepsilon \log d$ and the Lipschitz constant associated with the function $\phi^\varepsilon(\cdot)$ is $\frac{1}{\varepsilon}$.

G.2 Algorithm details

We write the optimization problem Eq. (38) as:

$$\min_{u \in \mathbb{R}^n_+} F(u, C) + H(u, C)$$

where

$$H(u, C) = \phi^{\varepsilon}(A^{-\frac{1}{2}}(\frac{1}{2n}X^{\top}\operatorname{Diag}(u)X - C)A^{-\frac{1}{2}})$$

is the smooth part and

$$F(u, C) = \mathbb{I}_{C:|C_{ij}| \leqslant c_i c_j} + \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{u_i}$$

is the non-smooth part.

The gradient ∇_u of $H(u, C)$ with respect to u is

$$\nabla_u = \operatorname{diag}(B^{\top}U\operatorname{Diag}(\sigma)U^{\top}B).$$

where $B = \frac{1}{\sqrt{2n}}A^{-\frac{1}{2}}X^{\top}$ and the gradient of $H(u, C)$ with respect to C is

$$\nabla_C = (A^{-\frac{1}{2}}U\operatorname{Diag}(\sigma)U^{\top}A^{-\frac{1}{2}}).$$

The Lipschitz constant L associated with the gradient $\nabla H(u, C)$ is

$$L = \frac{2}{\varepsilon}\max\Big(\lambda_{max}(B^{\top}B \odot B^{\top}B), \lambda_{max}^{2}(A^{-1})\Big), \tag{39}$$

where $\lambda_{max}(M)$ denotes the maximum eigen value of matrix M. Computing L takes $O(\max(n, d)^3)$ time and L needs to be computed once at the beginning of the algorithm.

The resultant FISTA procedure is described in Algorithm 1. Note that the FISTA procedure first computes intermediate iterates $(\bar{u}^{k-\frac{1}{2}}, \bar{C}^{k-\frac{1}{2}})$ (Step 7, Algorithm 1) by taking descent steps along the respective gradient directions. Then two distinct problems in u and C (respectively Steps 8 and 9 in Algorithm 1) are solved. The sub-problem in u (Step 8) can be efficiently solved using a Newton procedure followed by a thresholding step, as illustrated in Algorithm 2. The sub-problem in C (Step 9) can also be solved using a simple thresholding step.

Algorithm 1 *FISTA Algorithm to solve Eq. (38)*

```
Input $X$.
Compute Lipschitz constant $L$.
Let $(u^0, C^0)$ be an arbitrary starting point.
Let $(\bar{u}^0, \bar{C}^0) = (u^0, C^0)$, $t_0 = 1$.
Set the maximum iterations to be $K$.
for $k = 1, 2, \dots, K$ do        ▷ The loop can also be terminated based on duality gap.
    $(\bar{u}^{k-\frac{1}{2}}, \bar{C}^{k-\frac{1}{2}}) = \Big(\bar{u}^k - \frac{1}{L}\nabla_{\bar{u}^k}, \bar{C}^k - \frac{1}{L}\nabla_{\bar{C}^k}\Big)$.
    Obtain $u^k = \mathrm{argmin}_{u \in \mathbb{R}^n_+} \Big\{ \frac{L}{2}\|u - \bar{u}^{k-\frac{1}{2}}\|^2 + \frac{1}{2n}\sum_{i=1}^n \frac{1}{u_i}\}$ by Algorithm 2.
    Obtain $C^k = \mathrm{argmin}_C \Big\{ \mathbb{I}_{C:|C_{ij}| \leqslant c_i c_j} + \frac{L}{2}\|C - \bar{C}^{k-\frac{1}{2}}\|_F^2 \Big\}$ by thresholding.
    $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$.
    $(\bar{u}^k, \bar{C}^k) = (u^k, C^k) + \frac{(t_{k-1}-1)}{t_k}\Big((u^k, C^k) - (u^{k-1}, C^{k-1})\Big)$.
end for
Output $(u^K, C^K)$.
```

Algorithm 2 *Newton method to solve u sub-problem*

```
Input $u^{k-\frac{1}{2}}$, $n$, $L$.
$u_i^0 = \max(u_i^{k-\frac{1}{2}}, \frac{1}{(2nL)^{\frac{1}{3}}}),\ i = 1, 2, \dots, n$.
Set $\mathcal{M}$ to be the max number of Newton steps.
for $t = 1, 2, \dots, \mathcal{M}$ do
    for $i = 1, 2, \dots, n$ do
        $u_i^t = \frac{2nL(u_i^{t-1})^3 u_i^{k-\frac{1}{2}} + 3u_i^t}{2(nL(u_i^{t-1})^3 + 1)}$.
    end for
end for
Output $\max(u^{\mathcal{M}}, 0)$.
```