Robust Subspace Clustering

Mahdi Soltanolkotabi*, Ehsan Elhamifar† and Emmanuel J. Candès‡

January 2013

Abstract

Subspace clustering refers to the task of finding a multi-subspace representation that best fits a collection of points taken from a high-dimensional space. This paper introduces an algorithm inspired by sparse subspace clustering (SSC) [17] to cluster noisy data, and develops some novel theory demonstrating its correctness. In particular, the theory uses ideas from geometric functional analysis to show that the algorithm can accurately recover the underlying subspaces under minimal requirements on their orientation, and on the number of samples per subspace. Synthetic as well as real data experiments complement our theoretical study, illustrating our approach and demonstrating its effectiveness.

Keywords. Subspace clustering, spectral clustering, LASSO, Dantzig selector, ℓ_1 minimization, multiple hypothesis testing, true and false discoveries, geometric functional analysis, nonasymptotic random matrix theory.

1 Introduction

1.1 Motivation

In many problems across science and engineering, a fundamental step is to find a lower dimensional subspace which best fits a collection of points taken from a high-dimensional space; this is classically achieved via Principal Component Analysis (PCA). Such a procedure makes perfect sense as long as the data points are distributed around a lower dimensional subspace, or expressed differently, as long as the data matrix with points as column vectors has approximately low rank. A more general model might sometimes be useful when the data come from a mixture model in which points do not lie around a single lower-dimensional subspace but rather around a union of subspaces. For instance, consider an experiment in which gene expression data are gathered on many cancer cell lines with unknown subsets belonging to different tumor types. One can imagine that the expressions from each cancer type may span a distinct lower dimensional subspace. If the cancer labels were known in advance, one would apply PCA separately to each group but we here consider the case where the observations are unlabeled. Finding the components of the mixture and assigning each point to a fitted subspace is called subspace clustering. Even when the mixture model holds, the full data matrix may not have low rank at all, a situation which is very different from that where PCA is applicable.

*Department of Electrical Engineering, Stanford University, Stanford CA
†Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley CA
‡Departments of Mathematics and of Statistics, Stanford University, Stanford CA

In recent years, numerous algorithms have been developed for subspace clustering and applied to various problems in computer vision/machine learning [43] and data mining [34]. At the time of this writing, subspace clustering techniques are certainly gaining momentum as they begin to be used in fields as diverse as identification and classification of diseases [29], network topology inference [19], security and privacy in recommender systems [47], system identification [4], hyperspectral imaging [16], identification of switched linear systems [27, 32], and music analysis [21] to name just a few. In spite of all these interesting works, tractable subspace clustering algorithms either lack a theoretical justification, or are guaranteed to work under restrictive conditions rarely met in practice. (We note that although novel and often efficient clustering techniques come about all the time, establishing igrorous theory for such techniques has proven to be quite difficult. In the context of subspace clustering, Section 5 offers a partial survey of the existing literature.) Furthermore, proposed algorithms are not always computationally tractable. Thus, one important issue is whether tractable algorithms that can (provably) work in less than ideal situations—that is, under severe noise conditions and relatively few samples per subspace—exist.

Elhamifar and Vidal [17] have introduced an approach to subspace clustering, which relies on ideas from the sparsity and compressed sensing literature, please see also the longer version [18] which was submitted while this manuscript was under preparation. *Sparse subspace clustering* (SSC) [17, 18] is computationally efficient since it amounts to solving a sequence of ℓ_1 minimization problems and is, therefore, tractable. Now the methodology in [17] is mainly geared towards noiseless situations where the points lie exactly on lower dimensional planes, and theoretical performance guarantees in such circumstances are given under restrictive assumptions. Continuing on this line of work, [37] showed that good theoretical performance could be achieved under broad circumstances. However, the model supporting the theory in [37] is still noise free.

This paper considers the subspace clustering problem in the presence of noise. We introduce a tractable clustering algorithm, which is a natural extension of SSC, and develop rigorous theory about its performance. In a nutshell, we propose a statistical mixture model to represent data lying near a union of subspaces, and prove that in this model, the algorithm is effective as long as there are sufficiently many samples from each subspace and that the subspaces are not too close to each other. In this theory, the performance of the algorithm is explained in terms of interpretable and intuitive parameters such as (1) the values of the principal angles between subspaces, (2) the number of points per subspace, (3) the noise level and so on. In terms of these parameters, our theoretical results indicate that the performance of the algorithm is in some sense near the limit of what can be achieved by any algorithm, regardless of tractability.

1.2 Problem formulation and model

We assume we are given data points lying near a union of unknown linear subspaces; there are L subspaces $S_1, S_2, \ldots, S_L$ of $\mathbb{R}^n$ of dimensions $d_1, d_2, \ldots, d_L$. These together with their number are completely unknown to us. We are given a point set $\mathcal{Y} \subset \mathbb{R}^n$ of cardinality N, which may be partitioned as $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2 \cup \ldots \cup \mathcal{Y}_L$; for each $\ell \in \{1, \ldots, L\}$, $\mathcal{Y}_\ell$ is a collection of N_ℓ vectors that are 'close' to subspace S_ℓ. The goal is to approximate the underlying subspaces using the point set $\mathcal{Y}$. One approach is first to assign each data point to a cluster, and then estimate the subspaces representing each of the groups with PCA.

Our statistical model assumes that each point $\boldsymbol{y} \in \mathcal{Y}$ is of the form

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}, \tag{1.1}$$

where $\boldsymbol{x}$ belongs to one of the subspaces and $\boldsymbol{z}$ is an independent stochastic noise term. We suppose that the inverse signal-to-noise ratio (SNR) defined as $\mathbb{E}\,\|\boldsymbol{z}\|_2^2/\|\boldsymbol{x}\|_{\ell_2}^2$ is bounded above. Each observation is thus the superposition of a noiseless sample taken from one of the subspaces and of a stochastic perturbation whose Euclidean norm is about σ times the signal strength so that $\mathbb{E}\,\|z\|_{\ell_2}^2 = \sigma^2\|x\|_{\ell_2}^2$. All the way through, we assume that

$$\sigma < \sigma^\star, \quad \text{and} \quad \max_\ell d_\ell < c_0\,\frac{n}{(\log N)^2}, \tag{1.2}$$

where $\sigma^\star < 1$ and c_0 are fixed numerical constants. The second assumption is here to avoid unnecessarily complicated expressions later on. While more substantial, the first is not too restrictive since it just says that the signal $\boldsymbol{x}$ and the noise $\boldsymbol{z}$ may have about the same magnitude. (With an arbitrary perturbation of Euclidean norm equal to two, one can move from any point $\boldsymbol{x}$ on the unit sphere to just about any other point.)

This is arguably the simplest model providing a good starting point for a theoretical investigation. For the noiseless samples $\boldsymbol{x}$, we consider the intuitive *semi-random model* introduced in [37], which assumes that the subspaces are fixed with points distributed uniformly at random on each subspace. One can think of this as a mixture model where each component in the mixture is a lower dimensional subspace. (One can extend the methods to affine subspace clustering as briefly explained in Section 2.)

1.3 What makes clustering hard?

Two important parameters fundamentally affect the performance of subspace clustering algorithms: (1) the distance between subspaces and (2) the number of samples we have on each subspace.

1.3.1 Distance/affinity between subspaces

Intuitively, any subspace clustering algorithm operating on noisy data will have difficulty segmenting observations when the subspaces are close to each other. We of course need to quantify closeness, and the following definition captures a notion of distance or similarity/affinity between subspaces.

Definition 1.1 *The principal angles $\theta^{(1)}, \ldots, \theta^{(d\wedge d')}$ between two subspaces S and S' of dimensions d and d', are recursively defined by*

$$\cos(\theta^{(i)}) = \max_{\boldsymbol{u}\in S}\max_{\boldsymbol{v}\in S'} \frac{\boldsymbol{u}^T\boldsymbol{v}}{\|\boldsymbol{u}\|_{\ell_2}\,\|\boldsymbol{v}\|_{\ell_2}} := \frac{\boldsymbol{u}_i^T\boldsymbol{v}_i}{\|\boldsymbol{u}_i\|_{\ell_2}\,\|\boldsymbol{v}_i\|_{\ell_2}}$$

with the orthogonality constraints $\boldsymbol{u}^T\boldsymbol{u}_j = 0$, $\boldsymbol{v}^T\boldsymbol{v}_j = 0$, $j = 1, \ldots, i-1$.

Alternatively, if the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthobases for S and S', then the cosine of the principal angles are the singular values of $\boldsymbol{U}^T\boldsymbol{V}$.

Definition 1.2 *The normalized affinity between two subspaces is defined by*

$$\mathrm{aff}(S, S') = \sqrt{\frac{\cos^2\theta^{(1)} + \ldots + \cos^2\theta^{(d\wedge d')}}{d\wedge d'}}.$$

The affinity is a measure of correlation between subspaces. It is low when the principal angles are nearly right angles (it vanishes when the two subspaces are orthogonal) and high when the principal angles are small (it takes on its maximum value equal to one when one subspace is contained in the other). Hence, when the affinity is high, clustering is hard whereas it becomes easier as the affinity decreases. Ideally, we would like our algorithm to be able to handle higher affinity values—as close as possible to the maximum possible value.

There are other ways of measuring the affinity between subspaces; for instance, by taking the cosine of the first principal angle. We prefer the definition above as it offers the flexibility of allowing for some principal angles to be small or zero. As an example, suppose we have a pair of subspaces with a nontrivial intersection. Then $|\cos\theta^{(1)}| = 1$ regardless of the dimension of the intersection whereas the value of the affinity would depend upon this dimension.

1.3.2 Sampling density

Another important factor affecting the performance of subspace clustering algorithms has to do with the distribution of points on each subspace. In the model we study here, this essentially reduces to the number of points we have on each subspace.[1]

Definition 1.3 *The sampling density ρ of a subspace is defined as the number of samples on that subspace per dimension. In our multi-subspace model the density of S_ℓ is, therefore, $\rho_\ell = N_\ell/d_\ell$.*[2]

With noisy data, one expects the clustering problem to become easier as the sampling density increases. Obviously, if the sampling density of a subspace S is smaller than one, then any algorithm will fail in identifying that subspace correctly as there are not sufficiently many points to identify all the directions spanned by S. Hence, we would like a clustering algorithm to be able to operate at values of the sampling density as low as possible, i.e. as close to one as possible.

2 Robust subspace clustering: methods and concepts

This section introduces our methodology through heuristic arguments confirmed by numerical experiments. Section 3 presents theoretical guarantees showing that the entire approach is mathematically valid. From now on, we arrange the N observed data points as columns of a matrix $\boldsymbol{Y} = [\boldsymbol{y_1}, \ldots, \boldsymbol{y_N}] \in \mathbb{R}^{n\times N}$. With obvious notation, $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}$.

2.1 The normalized model

In practice, one may want to normalize the columns of the data matrix so that for all i, $\|\boldsymbol{y}_i\|_{\ell_2} = 1$. Since with our SNR assumption, we have $\|\boldsymbol{y}\|_{\ell_2} \approx \|\boldsymbol{x}\|_{\ell_2}\sqrt{1+\sigma^2}$ *before* normalization, then *after* normalization:

$$\boldsymbol{y} \approx \frac{1}{\sqrt{1+\sigma^2}}(\boldsymbol{x} + \boldsymbol{z}),$$

where $\boldsymbol{x}$ is unit-normed, and $\boldsymbol{z}$ has i.i.d. random Gaussian entries with variance σ^2/n.

[1] In a general deterministic model where the points have arbitrary orientations on each subspace, we can imagine that the clustering problem becomes harder as the points align along an even lower dimensional structure.

[2] Throughout, we take $\rho_\ell \le e^{d_\ell/2}$. Our results hold for all other values by substituting ρ_ℓ with $\rho_\ell \wedge e^{d_\ell/2}$ in all the expressions.

For ease of presentation, we work—in this section and in the proofs—with a model $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$ in which $\|\boldsymbol{x}\|_{\ell_2} = 1$ instead of $\|\boldsymbol{y}\|_{\ell_2} = 1$ (the numerical Section 6 is the exception). The normalized model with $\|\boldsymbol{x}\|_{\ell_2} = 1$ and $\boldsymbol{z}$ i.i.d. $\mathcal{N}(0, \sigma^2/n)$ is nearly the same as before. In particular, all of our methods and theoretical results in Section 3 hold with both models in which either $\|\boldsymbol{x}\|_{\ell_2} = 1$ or $\|\boldsymbol{y}\|_{\ell_2} = 1$.

2.2 The SSC scheme

We describe the approach in [17], which follows a three-step procedure:

I Compute an affinity matrix encoding similarities between sample pairs as to construct a weighted graph $\boldsymbol{W}$.

II Construct clusters by applying spectral clustering techniques (e.g. [31]) to $\boldsymbol{W}$.

III Apply PCA to each of the clusters.

The novelty in [17] concerns step I, the construction of the affinity matrix. Interestingly, similar ideas were introduced earlier in the statistics literature for the purpose of graphical model selection [30]. Now the work [17] of interest here is mainly concerned with the noiseless situation in which $\boldsymbol{Y} = \boldsymbol{X}$ and the idea is then to express each column $\boldsymbol{x}_i$ of $\boldsymbol{X}$ as a sparse linear combination of all the other columns. The reason is that under any reasonable condition, one expects that the *sparsest* representation of $\boldsymbol{x_i}$ would only select vectors from the subspace in which $\boldsymbol{x_i}$ happens to lie in. Applying the ℓ_1 norm as the convex surrogate of sparsity leads to the following sequence of optimization problems

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^N} \ \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{\beta} \text{ and } \beta_i = 0. \tag{2.1}$$

Here, β_i denotes the ith element of $\boldsymbol{\beta}$ and the constraint $\beta_i = 0$ removes the trivial solution that decomposes a point as a linear combination of itself. Collecting the outcome of these N optimization problems as columns of a matrix $\boldsymbol{B}$, [17] sets the similarity matrix to be $\boldsymbol{W} = |\boldsymbol{B}| + |\boldsymbol{B}|^T$.[3] (This algorithm clusters linear subspaces but can also cluster affine subspaces by adding the constraint $\boldsymbol{\beta}^T \mathbf{1} = 1$ to (2.1).)

The issue here is that we only have access to the noisy data $\boldsymbol{Y}$: this makes the problem challenging, as unlike conventional sparse recovery problems where only the response vector $\boldsymbol{x}_i$ is corrupted, here both the covariates (columns of $\boldsymbol{X}$) and the response vector are corrupted. In particular, it may not be advisable to use (2.1) with $\boldsymbol{y}_i$ and $\boldsymbol{Y}$ in place of $\boldsymbol{x}_i$ and $\boldsymbol{X}$ as, strictly speaking, sparse representations no longer exist. Observe that the expression $\boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{\beta}$ can be rewritten as $\boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{\beta} + (\boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{\beta})$. Viewing $(\boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{\beta})$ as a perturbation, it is natural to use ideas from sparse regression to obtain an estimate $\hat{\boldsymbol{\beta}}$, which is then used to construct the similarity matrix. In this paper, we follow the same three-step procedure and shall focus on the first step in Algorithm 1; that is, on the construction of reliable similarity measures between pairs of points. Since we have noisy data, we shall not use (2.1) here. Also, we add denoising to Step III, check the output of Algorithm 1.

[3] We use the terminology similarity graph or matrix instead of affinity matrix as not to overload the word 'affinity'.

Algorithm 1 Robust SSC procedure

Input: A data set $\mathcal{Y}$ arranged as columns of $\boldsymbol{Y} \in \mathbb{R}^{n \times N}$.

1. For each $i \in \{1, \ldots, N\}$, produce a sparse coefficient sequence $\hat{\boldsymbol{\beta}}$ by regressing the ith vector $\boldsymbol{y}_i$ onto the other columns of $\boldsymbol{Y}$. Collect these as columns of a matrix $\boldsymbol{B}$.
2. Form the similarity graph G with nodes representing the N data points and edge weights given by $\boldsymbol{W} = |\boldsymbol{B}| + |\boldsymbol{B}|^T$.
3. Sort the eigenvalues $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_N$ of the normalized Laplacian of G in descending order, and set
$$\hat{L} = N - \underset{i=1,\ldots,N-1}{\arg\max} \ (\delta_i - \delta_{i+1}).$$
4. Apply a spectral clustering technique to the similarity graph using $\hat{L}$ as the estimated number of clusters to obtain the partition $\mathcal{Y}_1, \ldots, \mathcal{Y}_{\hat{L}}$.
5. Use PCA to find the best subspace fits $(\{S_\ell\}_1^L)$ to each of the partitions $(\{\mathcal{Y}_\ell\}_1^L)$ and denoise $\boldsymbol{Y}$ as to obtain clean data points $\hat{\boldsymbol{X}}$.

Output: Subspaces $\{S_\ell\}_1^L$ and cleaned data points $\hat{\boldsymbol{X}}$.

2.3 Performance metrics for similarity measures

Given the general structure of the method, we are interested in sparse regression techniques, which tend to select points in the same clusters (share the same underlying subspace) over those that do not share this property. Expressed differently, the hope is that whenever $B_{ij} \neq 0$, $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$ belong to the same subspace. We introduce metrics to quantify performance.

Definition 2.1 (False discoveries) *Fix i and $j \in \{1, \ldots, N\}$ and let $\boldsymbol{B}$ be the outcome of Step 1 in Algorithm 1. Then we say that (i, j) obeying $B_{ij} \neq 0$ is a false discovery if $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ do not belong to the same subspace.*

Definition 2.2 (True discoveries) *In the same situation, (i, j) obeying $B_{ij} \neq 0$ is a true discovery if $\boldsymbol{y}_j$ and $\boldsymbol{y}_i$ belong to the same cluster/subspace.*

When there are no false discoveries, we shall say that the *subspace detection property* holds. In this case, the matrix $\boldsymbol{B}$ is block diagonal after applying a permutation which makes sure that columns in the same subspace are contiguous. In some cases, the sparse regression method may select vectors from other subspaces and this property will not hold. However, it might still be possible to detect and construct reliable clusters by applying steps 2–5 in Algorithm 1.

2.4 LASSO with data-driven regularization

A natural sparse regression strategy is the LASSO:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^N} \ \frac{1}{2} \|\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \beta_i = 0. \tag{2.2}$$

Whether such a methodology should succeed is unclear as we are not under a traditional model for both the response $\boldsymbol{y}_i$ and the covariates $\boldsymbol{Y}$ are noisy, see [35] for a discussion of sparse regression under matrix uncertainty and what can go wrong. The main contribution of this paper is to show that if one selects λ in a data-driven fashion, then compelling practical and theoretical performance can be achieved.

2.4.1 About as many true discoveries as dimension?

The nature of the problem is such that we wish to make few false discoveries (and not link too many pairs belonging to different subspaces) and so we would like to choose λ large. At the same time, we wish to make many true discoveries, whence a natural trade off. The reason why we need many true discoveries is that spectral clustering needs to assign points to the same cluster when they indeed lie near the same subspace. If the matrix $\boldsymbol{B}$ is too sparse, this will not happen.

We now introduce a principle for selecting the regularization parameter. Suppose we have noiseless data so that $\boldsymbol{Y} = \boldsymbol{X}$, and thus solve (2.1) with equality constraints. Under our model, assuming there are no false discoveries the optimal solution is guaranteed to have exactly d—the dimension of the subspace the sample under study belongs to—nonzero coefficients with probability one. That is to say, when the point lies in a d-dimensional space, we find d 'neighbors'.

The selection rule we shall analyze in this paper is to take λ as large as possible (as to prevent false discoveries) while making sure that the number of true discoveries is also on the order of the dimension d, typically in the range $[0.5d, 0.8d]$. We can say this differently. Imagine that all the points lie in the same subspace of dimension d so that every discovery is true. Then we wish to select λ in such a way that the number of discoveries is a significant fraction of d, the number one would get with noiseless data. Which value of λ achieves this goal? We will see in Section 2.4.2 that the answer is around $1/\sqrt{d}$. To put this in context, this means that we wish to select a regularization parameter which depends upon the dimension d of the subspace our point belongs to. (We are aware that the dependence on d is unusual as in sparse regression the regularization parameter usually does not depend upon the sparsity of the solution.) In turn, this immediately raises another question: since d is unknown, how can we proceed? In Section 2.4.4 we will see that it is possible to guess the dimension and construct fairly reliable estimates.

2.4.2 Data-dependent regularization

We now discuss values of λ obeying the demands formulated in the previous section. Our arguments are informal and we refer the reader to Section 3 for rigorous statements and to Section 8 for proofs. First, it simplifies the discussion to assume that we have no noise (the noisy case assuming $\sigma \ll 1$ is similar). Following our earlier discussion, imagine we have a vector $\boldsymbol{x} \in \mathbb{R}^n$ lying in the d-dimensional span of the columns of an $n \times N$ matrix $\boldsymbol{X}$. We are interested in values of λ so that the minimizer $\hat{\boldsymbol{\beta}}$ of the LASSO functional

$$K(\boldsymbol{\beta}, \lambda) = \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1}$$

has a number of nonzero components in the range $[0.5d, 0.8d]$, say. Now let $\hat{\boldsymbol{\beta}}_{\mathrm{eq}}$ be the solution of the problem with equality constraints, or equivalently of the problem above with $\lambda \to 0^+$. Then

$$\frac{1}{2} \left\|\boldsymbol{x} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|_{\ell_2}^2 \le K(\hat{\boldsymbol{\beta}}, \lambda) \le K(\hat{\boldsymbol{\beta}}_{\mathrm{eq}}, \lambda) = \lambda \left\|\hat{\boldsymbol{\beta}}_{\mathrm{eq}}\right\|_{\ell_1}. \tag{2.3}$$

We make two observations: the first is that if $\hat{\boldsymbol{\beta}}$ has a number of nonzero components in the range $[0.5d, 0.8d]$, then $\|\boldsymbol{x} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_{\ell_2}^2$ has to be greater than or equal to a fixed numerical constant. The reason is that we cannot approximate to arbitrary accuracy a generic vector living in a d-dimensional subspace as a linear combination of about $d/2$ elements from that subspace. The second observation is that $\|\hat{\boldsymbol{\beta}}_{\mathrm{eq}}\|_{\ell_1}$ is on the order of $\sqrt{d}$, which is a fairly intuitive scaling (we have d coordinates, each of size about $1/\sqrt{d}$). This holds with the proviso that the algorithm operates

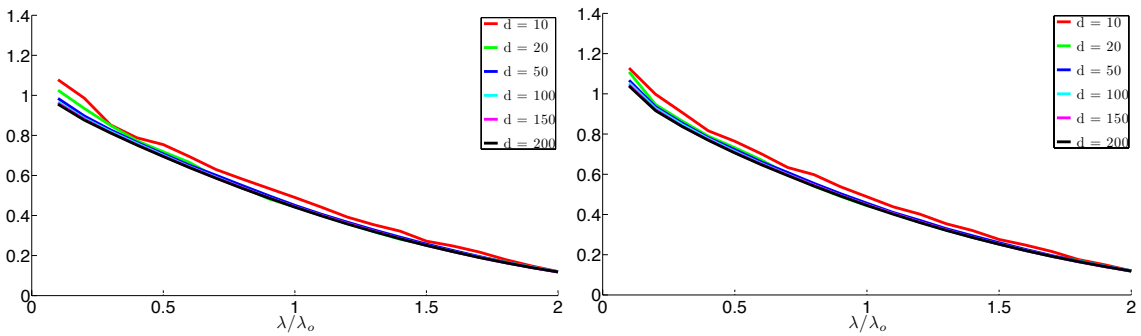


Figure 1: Average number of true discoveries normalized by subspace dimension for values of λ in an interval including the heuristic $\lambda_o = 1/\sqrt{d}$. (a) $\sigma = 0.25$. (b) $\sigma = 0.5$.

correctly in the noiseless setting and does not select columns from other subspaces. Then (2.3) implies that λ has to scale at least like $1/\sqrt{d}$. On the other hand, $\hat{\boldsymbol{\beta}} = \mathbf{0}$ if $\lambda \geq \|\boldsymbol{X}^T \boldsymbol{x}\|_{\ell_\infty}$. Now the informed reader knows that $\|\boldsymbol{X}^T \boldsymbol{x}\|_{\ell_\infty}$ scales at most like $\sqrt{(\log N)/d}$ so that choosing λ around this value yields no discovery (one can refine this argument to show that λ cannot be higher than a constant times $1/\sqrt{d}$ as we would otherwise have a solution that is too sparse). Hence, λ is around $1/\sqrt{d}$.

It might be possible to compute a precise relationship between λ and the expected number of true discoveries in an asymptotic regime in which the number of points and the dimension of the subspace both increase to infinity in a fixed ratio by adapting ideas from [6, 7]. We will not do so here as this is beyond the scope of this paper. Rather, we investigate this relationship by means of a numerical study.

Here, we fix a single subspace in $\mathbb{R}^n$ with $n = 2,000$. We use a sampling density equal to $\rho = 5$ and vary the dimension $d \in \{10, 20, 50, 100, 150, 200\}$ of the subspace as well as the noise level $\sigma \in \{0.25, 0.5\}$. For each data point, we solve (2.2) for different values of λ around the heuristic $\lambda_o = 1/\sqrt{d}$, namely, $\lambda \in [0.1\lambda_o, 2\lambda_o]$. In our experiments, we declare a discovery if an entry in the optimal solution exceeds 10^{-3}. Figures 1a and 1b show the number of discoveries per subspace dimension (the number of discoveries divided by d). One can clearly see that the curves corresponding to various subspace dimensions stack up on top of each other, thereby confirming that a value of λ on the order of $1/\sqrt{d}$ yields a fixed fraction of true discoveries. Further inspection also reveals that the fraction of true discoveries is around 50% near $\lambda = \lambda_o$, and around 75% near $\lambda = \lambda_o/2$.

2.4.3 The false-true discovery trade off

We now show empirically that in our model choosing λ around $1/\sqrt{d}$ typically yields very few false discoveries as well as many true discoveries; this holds with the proviso that the subspaces are of course not very close to each other.

In this simulation, 22 subspaces of varying dimensions in $\mathbb{R}^n$ with $n = 2,000$ have been independently selected uniformly at random; there are 5, 4, 3, 4, 4, and 2 subspaces of respective dimensions 200, 150, 100, 50, 20 and 10. This is a challenging regime since the sum of the subspace dimensions equals $2,200$ and exceeds the ambient dimension (the clean data matrix $\boldsymbol{X}$ has full

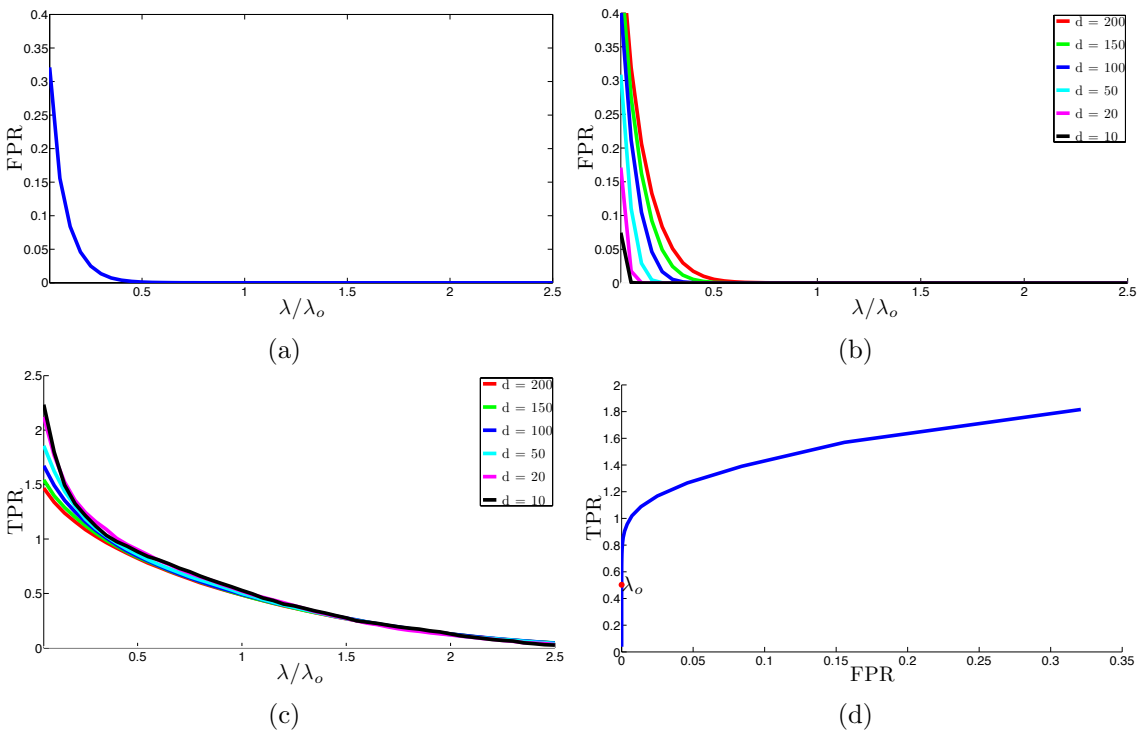


Figure 2: Performance of LASSO for values of λ in an interval including the heuristic $\lambda_o = 1/\sqrt{d}$. (a) Average number of false discoveries normalized by $(n-d)$ (FPR) on all m sampled data points). (b) FPR for different subspace dimensions. Each curve represents the average FPR over those samples originating from subspaces of the same dimension. (c) Average number of true discoveries per dimension for various dimensions (TPR). (d) TPR vs. FPR (ROC curve). The point corresponding to $\lambda = \lambda_o$ is marked as a red dot.

rank). We use a sampling density equal to $\rho = 5$ for each subspace and set the noise level to $\sigma = 0.3$. To evaluate the performance of the optimization problem (2.2), we proceed by selecting a subset of columns as follows: for each dimension, we take 100 cases at random belonging to subspaces of that dimension. Hence, the total number of test cases is $m = 600$ so that we only solve m optimization problems (2.2) out of the total N possible cases. Below, $\boldsymbol{\beta}^{(i)}$ is the solution to (2.2) and $\boldsymbol{\beta}_S^{(i)}$ its restriction to columns with indices in the same subspace. Hence, a nonvanishing entry in $\boldsymbol{\beta}_S^{(i)}$ is a true discovery and, likewise, a nonvanishing entry in $\boldsymbol{\beta}_{S^c}^{(i)}$ is false. For each data point we sweep the tuning parameter λ in (2.2) around the heuristic $\lambda_o = 1/\sqrt{d}$ and work with $\lambda \in [0.05\lambda_o, 2.5\lambda_o]$. In our experiments, a discovery is a value obeying $|B_{ij}| > 10^{-3}$.

In analogy with the signal detection literature we view the empirical averages of $\|\boldsymbol{\beta}_{S^c}^{(i)}\|_{\ell_0}/(n-d)$ and $\|\boldsymbol{\beta}_S^{(i)}\|_{\ell_0}/d$ as False Positive Rate (FPR) and True Positive Rate (TPR). On the one hand, Figures 2a and 2b show that for values around $\lambda = \lambda_o$, the FPR is zero (so there are no false discoveries). On the other hand, Figure 2c shows that the TPR curves corresponding to different dimensions are very close to each other and resemble those in Figure 1 in which all the points belong to the same cluster with no opportunity of making a false discovery. Hence, taking λ near

$1/\sqrt{d}$ gives a performance close to what can be achieved in a noiseless situation. That is to say, we have no false discovery and a number of true discoveries about $d/2$ if we choose $\lambda = \lambda_o$. Figure 2d plots TPR versus FPR (a.k.a. the Receiver Operating Characteristic (ROC) curve) and indicates that $\lambda = \lambda_o$ (marked by a red dot) is an attractive trade-off as it provides no false discoveries and sufficiently many true discoveries.

2.4.4 A two-step procedure

Returning to the selection of the regularization parameter, we would like to use λ on the order of $1/\sqrt{d}$. However, we do not know d and proceed by substituting an estimate. In the next section, we will see that we are able to quantify theoretically the performance of the following proposal: (1) run a hard constrained version of the LASSO and use an estimate $\hat{d}$ of dimension based on the ℓ_1 norm of the fitted coefficient sequence; (2) impute a value for λ constructed from $\hat{d}$. The two-step procedure is explained in Algorithm 2.

Algorithm 2 Two-step procedure with data-driven regularization

for $i = 1, \dots, N$ **do**

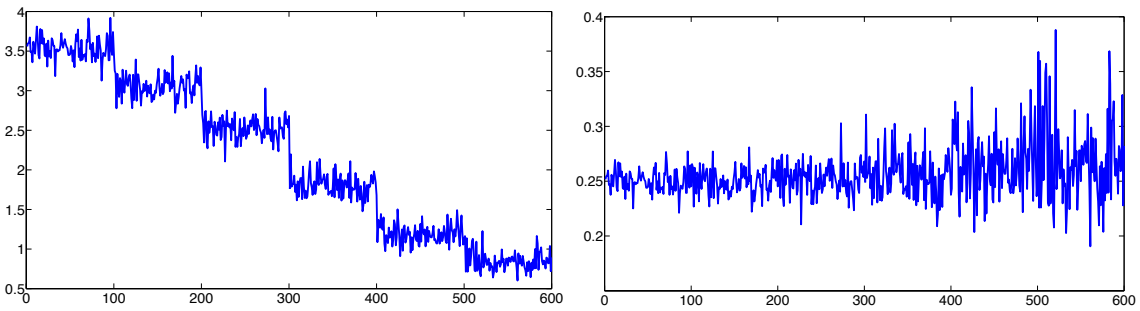
1. Solve
$$\boldsymbol{\beta}^\star = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^N} \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \|\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_{\ell_2} \le \tau \quad \text{and} \quad \boldsymbol{\beta}_i = 0. \tag{2.4}$$
2. Set $\lambda = f(\|\boldsymbol{\beta}^\star\|_{\ell_1})$.
3. Solve
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^N} \frac{1}{2}\|\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{\beta}_i = 0.$$
4. Set $\boldsymbol{B}_i = \hat{\boldsymbol{\beta}}$.

end for

To understand the rationale behind this, imagine we have noiseless data—i. e. $\boldsymbol{Y} = \boldsymbol{X}$—and are solving (2.1), which simply is our first step (2.4) with the proviso that $\tau = 0$. When there are no false discoveries, one can show that the ℓ_1 norm of $\boldsymbol{\beta}^\star$ is roughly of size $\sqrt{d}$ as shown in Lemma 8.2 from Section 8. This suggests using a multiple of $\|\boldsymbol{\beta}^\star\|_{\ell_1}$ as a proxy for $\sqrt{d}$. To drive this point home, take a look at Figure 3a which solves (2.4) with the same data as in the previous example and $\tau = 2\sigma$. The plot reveals that the values of $\|\boldsymbol{\beta}^\star\|_{\ell_1}$ fluctuate around $\sqrt{d}$. This is shown more clearly in Figure 3b, which shows that $\|\boldsymbol{\beta}^\star\|_{\ell_1}$ is concentrated around $\frac{1}{4}\sqrt{d}$ with, as expected, higher volatility at lower values of dimension.

Under suitable assumptions, we shall see in Section 3 that with noisy data, there are simple rules for selecting τ that guarantee, with high probability, that there are no false discoveries. To be concrete, one can take $\tau = 2\sigma$ and $f(t) \propto t^{-1}$. Returning to our running example, we have $\|\boldsymbol{\beta}^\star\|_{\ell_1} \approx \frac{1}{4}\sqrt{d}$. Plugging this into $\lambda = 1/\sqrt{d}$ suggests taking $f(t) \approx 0.25t^{-1}$. The plots in Figure 4 demonstrate that this is indeed effective. Experiments in Section 6 indicate that this is a good choice on real data as well.

The two-step procedure requires solving two LASSO problems for each data point and is useful when there are subspaces of large dimensions (in the hundreds, say) and some others of low-dimensions (three or four, say). In some applications such as motion segmentation in computer



(a) (b)

Figure 3: Optimal values of (2.4) for 600 samples using $\tau = 2\sigma$. The first 100 values correspond to points originating from subspaces of dimension $d = 200$, the next 100 from those of dimension $d = 150$, and so on through $d \in \{100, 50, 20, 10\}$. (a) Value of $\|\boldsymbol{\beta}^*\|_{\ell_1}$. (b) Value of $\|\boldsymbol{\beta}^*\|_{\ell_1} / \sqrt{d}$.

vision, the dimensions of the subspaces are all equal and known in advance. In this case, one can forgo the two-step procedure and simply set $\lambda = 1/\sqrt{d}$.

3 Theoretical Results

This section presents our main theoretical results concerning the performance of the two-step procedure (Algorithm 2). We make two assumptions:

- **Affinity condition.** We say that a subspace S_ℓ obeys the affinity condition if

$$\max_{k:k\neq\ell} \quad \text{aff}(S_\ell, S_k) \leq \kappa_0 / \log N, \tag{3.1}$$

  where κ_0 a fixed numerical constant.

- **Sampling condition.** We say that subspace S_ℓ obeys the *sampling condition* if

$$\rho_\ell \geq \rho^\star, \tag{3.2}$$

  where $\rho^\star$ is a fixed numerical constant.

The careful reader might argue that we should require lower affinity values as the noise level increases. The reason why σ does not appear in (3.1) is that we assumed a bounded noise level. For higher values of σ, the affinity condition would read as in (3.1) with a right-hand side equal to

$$\kappa = \frac{\kappa_0}{\log N} - \sigma \sqrt{\frac{d_\ell}{2n \log N}}.$$

3.1 Main results

From here on we use $d(i)$ to refer to the dimension of the subspace the vector $\boldsymbol{y}_i$ originates from. $N(i)$ and $\rho(i)$ are used in a similar fashion for the number and density of points on this subspace.

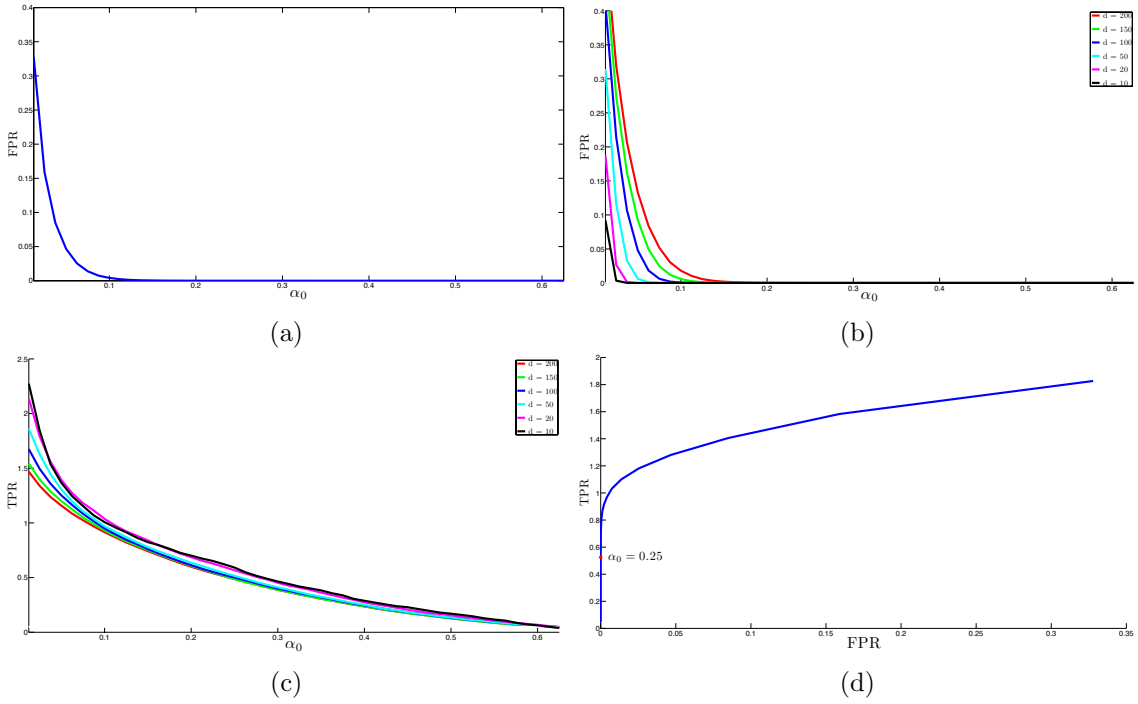


Figure 4: Performance of the two-step procedure using $\tau = 2\sigma$ and $f(t) = \alpha_0 t^{-1}$ for values of α_0 around the heuristic $\alpha_0 = 0.25$. (a) False positive rate (FPR). (b) FPR for various subspace dimensions. (c) True positive rate (TPR). (d) TPR vs. FPR.

Theorem 3.1 (No false discoveries) *Assume that the subspace attached to the ith column obeys the affinity and sampling conditions and that the noise level σ is bounded as in (1.2), where $\sigma^\star$ is a sufficiently small numerical constant. In Algorithm 2, take $\tau = 2\sigma$ and $f(t)$ obeying $f(t) \geq 0.707\sigma t^{-1}$. Then with high probability,[4] there is no false discovery in the ith column of $\boldsymbol{B}$.*

Theorem 3.2 (Many true discoveries) *Consider the same setup as in Theorem 3.1 with f also obeying $f(t) \leq \alpha_0 t^{-1}$ for some numerical constant α_0. Then with high probability,[5] there are at least*

$$c_0 \frac{d(i)}{\log \rho(i)} \tag{3.3}$$

true discoveries in the ith column (c_0 is a positive numerical constant).

The above results indicate that the algorithm works correctly in fairly broad conditions. To give an example, assume two subspaces of dimension d overlap in a smaller subspace of dimension s but are orthogonal to each other in the remaining directions (equivalently, the first s principal angles are 0 and the rest are $\pi/2$). In this case, the affinity between the two subspaces is equal to $\sqrt{s/d}$ and (3.1) allows s to grow almost linearly in the dimension of the subspaces. Hence, subspaces can have intersections of large dimensions. In contrast, previous work with perfectly noiseless data [17] would impose to have a first principal angle obeying $|\cos\theta^{(1)}| \leq 1/\sqrt{d}$ so that the subspaces are practically orthogonal to each other. Whereas our result shows that we can have an average of the cosines practically constant, the condition in [17] asks that the maximum cosine be very small.

In the noiseless case, [37] showed that when the sampling condition holds and

$$\max_{k:k\neq\ell} \quad \mathrm{aff}(S_\ell, S_k) \leq \kappa_0 \frac{\sqrt{\log \rho_\ell}}{\log N},$$

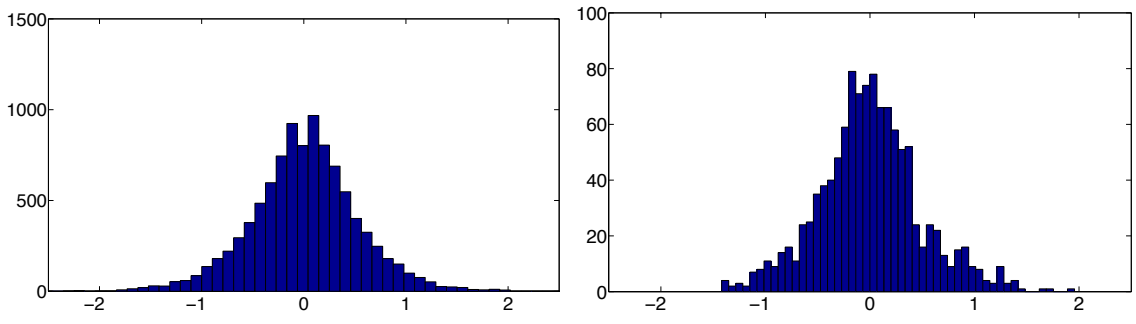
(albeit with slightly different values κ_0 and $\rho^\star$), then applying the noiseless version (2.1) of the algorithm also yields no false discoveries. Hence, with the proviso that the noise level is not too large, conditions under which the algorithm is provably correct are essentially the same.

Earlier, we argued that we would like to have, if possible, an algorithm provably working at (1) high values of the affinity parameters and (2) low values of the sampling density as these are the conditions under which the clustering problem is challenging. (Another property on the wish list is the ability to operate properly with high noise or low SNR and this is discussed next.) In this context, since the affinity is at most one, our results state that the affinity can be within a log factor from this maximum possible value. The number of samples needed per subspace is minimal as well. That is, as long as the density of points on each subspace is larger than a constant $\rho > \rho^\star$, the algorithm succeeds.

We would like to have a procedure capable of making no false discoveries and many true discoveries at the same time. Now in the noiseless case, whenever there are no false discoveries, the ith column contains exactly $d(i)$ true discoveries. Theorem 3.2 states that as long as the noise level σ is less than a fixed numerical constant, the number of true discoveries is roughly on the same order as in the noiseless case. In other words, a noise level of this magnitude does not fundamentally affect the performance of the algorithm. This holds even when there is great variation in the dimensions of the subspaces, and is possible because λ is appropriately tuned in an adaptive fashion.

[4] probability at least $1 - 2e^{-\gamma_1 n} - 6e^{-\gamma_2 d(i)} - e^{-\sqrt{N(i)d(i)}} - \frac{23}{N^2}$, for fixed numerical constants γ_1, γ_2.

[5] probability at least $1 - 2e^{-\gamma_1 n} - 6e^{-\gamma_2 d(i)} - e^{-\sqrt{N(i)d(i)}} - \frac{23}{N^2}$, for fixed numerical constants γ_1, γ_2.



(a) (b)

Figure 5: Histograms of the true discovery values from the two step procedure with $\alpha_0 = 0.25$ (multiplied by $\sqrt{d}$). (a) $d = 200$. (b) $d = 20$.

The number of true discoveries is shown to scale at least like dimension over the log of the density. This may suggest that the number of true discoveries decreases (albeit very slowly) as the sampling density increases. This behavior is to be expected: when the sampling density becomes exponentially large (in terms of the dimension of the subspace) the number of true discoveries become small since we need fewer columns to synthesize a point. In fact, the $d/\log\rho$ behavior seems to be the correct scaling. Indeed, when the density is low and ρ takes on a small value, (3.3) asserts that we make on the order of d discoveries, which is tight. Imagine now that we are in the high-density regime and ρ is exponential in d. Then as the points gets tightly packed, we expect to have only one discovery in accordance with (3.3).

Theorem 3.2 establishes that there are many true discoveries. This would not be useful for clustering purposes if there were only a handful of very large true discoveries and all the others of negligible magnitude. The reason is that the similarity matrix $\boldsymbol{W}$ would then be close to a sparse matrix and we would run the risk of splitting true clusters. This is not what happens and our proofs can show this although we do not do this in this paper for lack of space. Rather, we demonstrate this property empirically. On our running example, Figures 5a and 5b show that the histograms of appropriately normalized true discovery values resemble a bell-shaped curve.

Finally, we would like to comment on the fact that our main results hold when λ belongs to a fairly broad range of values. First, when all the subspaces have small dimensions, one can choose the same value of λ for all the data points since $1/\sqrt{d}$ is essentially constant. Hence, when we know a priori that we are in such a situation, there may be no need for the two step procedure. (We would still recommend the conservative two-step procedure because of its superior empirical performance on real data.) Second, the proofs also reveal that if we have knowledge of the dimension of the largest subspace $d_{\max}$, the first theorem holds with a fixed value of λ proportional to $\sigma/\sqrt{d_{\max}}$. Third, when the subspaces themselves are drawn at random, the first theorem holds with a fixed value of λ proportional to $\sigma(\log N)/\sqrt{n}$. (Both these statements follow by plugging these values of λ in the proofs of Section 8 and we omit the calclulations.) We merely mention these variants to give a sense of what our theorems can also give. As explained earlier, we recommend the more conservative two-step procedure with the proxy for $1/\sqrt{d}$. The reason is that using a higher value of λ allows for a larger value of κ_0, which says that the subspaces can be even closer. In other words, we can function in a more challenging regime. To drive this point home, consider the noiseless problem. When the subspaces are close, the equality constrained ℓ_1 problem may yield some false

discoveries. However, if we use the LASSO version—even though the data is noiseless—we may end up with no false discoveries while maintaining sufficiently many true discoveries.

4 The Bias-corrected Dantzig Selector

One can think of other ways of performing the first step in Algorithm 1 and this section discusses another approach based on a modification of the Dantzig selector, a popular sparse regression technique [13]. Unlike the two-step procedure, we do not claim any theoretical guarantees for this method and shall only explore its properties on real and simulated data.

Applied directly to our problem, the Dantzig selector takes the form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^N} \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \|\boldsymbol{Y}_{(-i)}^T(\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta})\|_{\ell_\infty} \le \lambda \quad \text{and} \quad \boldsymbol{\beta}_i = 0, \tag{4.1}$$

where $\boldsymbol{Y}_{(-i)}$ is $\boldsymbol{Y}$ with the ith column deleted. However, this is hardly suitable since the design matrix $\boldsymbol{Y}$ is corrupted. Interestingly, recent work [35, 36] has studied the problem of estimating a sparse vector from the standard linear model under uncertainty in the design matrix. The setup in these papers is close to our problem and we propose a modified Dantzig selection procedure inspired but not identical to the methods set forth in [35, 36].

4.1 The correction

If we had clean data, we would solve (2.1); this is (4.1) with $\boldsymbol{Y} = \boldsymbol{X}$ and $\lambda = 0$. Let $\boldsymbol{\beta}^I$ be the solution to this ideal noiseless problem. Applied to our problem, the main idea in [35, 36] would be to find a formulation that resembles (4.1) with the property that $\boldsymbol{\beta}^I$ is feasible. Since $\boldsymbol{x}_i = \boldsymbol{X}_{(-i)}\boldsymbol{\beta}^I{}_{(-i)}$, observe that we have the following decomposition:

$$\begin{aligned}\boldsymbol{Y}_{(-i)}^T(\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta}^I) &= (\boldsymbol{X}_{(-i)} + \boldsymbol{Z}_{(-i)})^T(\boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{\beta}^I) \\ &= \boldsymbol{X}_{(-i)}^T(\boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{\beta}^I) + \boldsymbol{Z}_{(-i)}^T\boldsymbol{z}_i - \boldsymbol{Z}_{(-i)}^T\boldsymbol{Z}\boldsymbol{\beta}^I.\end{aligned}$$

Then the conditional mean is given by

$$\mathbb{E}[\boldsymbol{Y}_{(-i)}^T(\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta}^I) \,|\, \boldsymbol{X}] = -\,\mathbb{E}\,\boldsymbol{Z}_{(-i)}^T\boldsymbol{Z}_{(-i)}\boldsymbol{\beta}^I{}_{(-i)} = -\sigma^2\boldsymbol{\beta}^I{}_{(-i)}.$$

In other words,

$$\sigma^2\boldsymbol{\beta}^I{}_{(-i)} + \boldsymbol{Y}_{(-i)}^T(\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta}^I) = \boldsymbol{\xi}$$

where $\boldsymbol{\xi}$ has mean zero. In Section 4.2, we compute the variance of the jth component ξ_j, given by

$$\mathbb{E}\,\xi_j^2 = \frac{\sigma^2}{n}(1 + \|\boldsymbol{\beta}^I\|_{\ell_2}^2) + \frac{\sigma^4}{n}\big(1 + (\beta_j^I)^2 + \|\boldsymbol{\beta}^I\|_{\ell_2}^2\big). \tag{4.2}$$

Owing to our Gaussian assumptions, $|\xi_j|$ shall be smaller than 3 or 4 times this standard deviation, say, with high probability.

Hence, we may want to consider a procedure of the form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^N} \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \|\boldsymbol{Y}_{(-i)}^T(\boldsymbol{y}_i - \boldsymbol{Y}\boldsymbol{\beta}) + \sigma^2\boldsymbol{\beta}_{(-i)}\|_{\ell_\infty} \le \lambda \quad \text{and} \quad \boldsymbol{\beta}_i = 0. \tag{4.3}$$

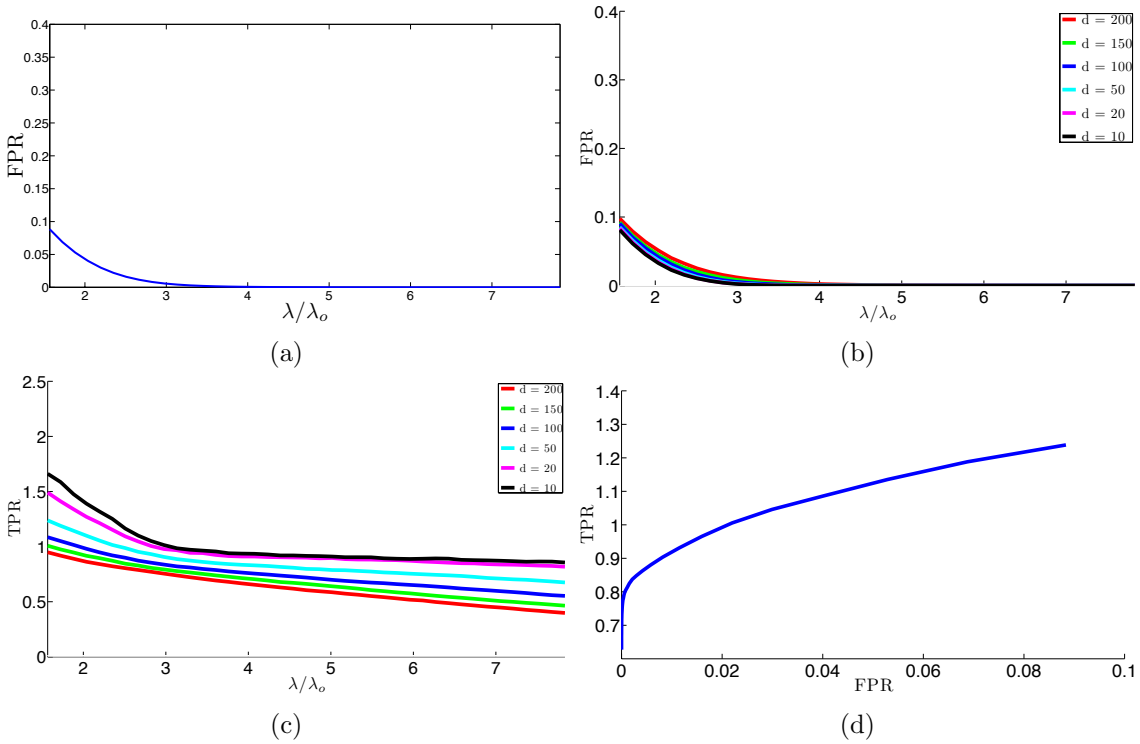


Figure 6: Performance of the bias-corrected Dantzig selector for values of λ that are multiples of the heuristic $\lambda_o = \sqrt{2/n}\,\sigma\,\sqrt{1+\sigma^2}$. (a) False positive rate (FPR). (b) FPR for different subspace dimensions. (c) True positive rate (TPR). (d) TPR vs. FPR.

It follows that if we take λ to be a reasonable multiple of (4.2), then $\boldsymbol{\beta}^I$ would obey the constraint in (4.3) with high probability. Hence, we would need to approximate the variance (4.2). Numerical simulations together with asymptotic calculations presented in Appendix C give that $\|\boldsymbol{\beta}^I\|_{\ell_2} \le 1$ with very high probability. Thus neglecting the term in $(\beta_j^I)^2$,

$$\mathbb{E}\,\xi_j^2 \approx \frac{\sigma^2}{n}(1+\sigma^2)(1+\|\boldsymbol{\beta}^I\|_{\ell_2}^2) \le 2\frac{\sigma^2}{n}(1+\sigma^2).$$

This suggests taking λ to be a multiple of $\sqrt{2/n}\,\sigma\,\sqrt{1+\sigma^2}$. This is interesting because the parameter λ does not depend on the dimension of the underlying subspace. We shall refer to (4.3) as the *bias-corrected Dantzig selector*, which resembles the proposal in [35, 36] for which the constraint is a bit more complicated and of the form $\|\boldsymbol{X}_{(-i)}^T(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{D}_{(-i)}\boldsymbol{\beta}\|_{\ell_\infty} \le \mu\|\boldsymbol{\beta}\|_{\ell_1} + \lambda$.

To get a sense about the validity of this proposal, we test it on our running example by varying $\lambda \in [\lambda_o, 8\lambda_o]$ around the heuristic $\lambda_o = \sqrt{2/n}\,\sigma\,\sqrt{1+\sigma^2}$. Figure 6 shows that good results are achieved around factors in the range $[4, 6]$.

In our synthetic simulations, both the two-step procedure and the corrected Dantzig selector seem to be working well in the sense that they yield many true discoveries while making very few false discoveries, if any. Comparing Figures 6b and 6c with those from Section 2 show that the corrected Dantzig selector has more true discoveries for subspaces of small dimensions (they are essentially the same for subspaces of large dimensions); that is, the two-step procedure is more conservative when it comes to subspaces of smaller dimensions. As explained earlier this is due to

our conservative choice of λ resulting in a TPR about half of what is obtained in a noiseless setting. Having said this, it is important to keep in mind that in these simulations the planes are drawn at random and as a result, they are sort of far from each other. This is why a less conservative procedure can still achieve a low FPR. When smaller subspaces are closer to each other or when the statistical model does not hold exactly as in real data scenarios, a conservative procedure may be more effective. In fact, experiments on real data in Section 6 confirm this and show that for the corrected Dantzig selector, one needs to choose values much larger than λ_o to yield good results.

4.2 Variance calculation

By definition,

$$\xi_j = \langle \boldsymbol{x}_j, \boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{\beta}^I \rangle + \langle \boldsymbol{z}_j, \boldsymbol{z}_i \rangle - (\boldsymbol{z}_j^T \boldsymbol{z}_j - \sigma^2)\beta_j^I - \sum_{k:k\neq i,j} \boldsymbol{z}_j^T \boldsymbol{z}_k \beta_k^I := I_1 + I_2 + I_3 + I_4.$$

A simple calculation shows that for $\ell_1 \neq \ell_2$, $\mathrm{Cov}(I_{\ell_1}, I_{\ell_2}) = 0$ so that

$$\mathbb{E}\,\xi_j^2 = \sum_{\ell=1}^{4} \mathrm{Var}(I_\ell).$$

We compute

$$\mathrm{Var}(I_1) = \frac{\sigma^2}{n}(1 + \|\boldsymbol{\beta}^I\|_{\ell_2}^2), \qquad \mathrm{Var}(I_3) = \frac{\sigma^4}{n} 2(\beta_j^I)^2,$$
$$\mathrm{Var}(I_2) = \frac{\sigma^4}{n}, \qquad \mathrm{Var}(I_4) = \frac{\sigma^4}{n}[\|\boldsymbol{\beta}^I\|_{\ell_2}^2 - (\beta_j^I)^2]$$

and (4.2) follows.

5 Comparisons With Other Works

We now briefly comment on other approaches to subspace clustering. Since this paper is theoretical in nature, we shall focus on comparing theoretical properties and refer to [18], [43] for a detailed comparison about empirical performance. Three themes will help in organizing our discussion.

- *Tractability.* Is the proposed method or algorithm computationally tractable?
- *Robustness.* Is the algorithm provably robust to noise and other imperfections?
- *Efficiency.* Is the algorithm correctly operating near the limits we have identified above? In our model, how many points do we need per subspace? How large can the affinity between subspaces be?

One can broadly classify existing subspace clustering techniques into three categories, namely, algebraic, iterative and statistical methods.

Methods inspired from algebraic geometry have been introduced for clustering purposes. In this area, a mathematically intriguing approach is the *generalized principal component analysis* (GPCA) presented in [44]. Unfortunately, this algorithm is not tractable in the dimension of the subspaces, meaning that a polynomial-time algorithm does not exist. Another feature is that GPCA is not robust to noise although some heuristics have been developed to address this issue, see e.g. [28].

As far as the dependence upon key parameters is concerned, GPCA is essentially optimal. An interesting approach to make GPCA robust is based on semidefinite programming [33]. However, this novel formulation is still intractable in the dimension of the subspaces and it is not clear how the performance of the algorithm depends upon the parameters of interest.

A representative example of an iterative method—the term is taken from the tutorial [43]—is the K-subspace algorithm [39], a procedure which can be viewed as a generalization of K-means. Here, the subspace clustering problem is formulated as a non-convex optimization problem over the choice of bases for each subspace as well as a set of variables indicating the correct segmentation. A cost function is then iteratively optimized over the basis and the segmentation variables. Each iteration is computationally tractable. However, due to the non-convex nature of the problem, the convergence of the sequence of iterates is only guaranteed to a local minimum. As a consequence, the dependence upon the key parameters is not well understood. Furthermore, the algorithm can be sensitive to noise and outliers. Other examples of iterative methods may be found in [1, 10, 25, 48].

Statistical methods typically model the subspace clustering problem as a mixture of degenerate Gaussian observations. Two such approaches are *mixtures of probabilistic PCA* (MPPCA) [38] and *agglomerative lossy compression* (ALC) [26]. MPPCA seeks to compute a maximum-likelihood estimate of the parameters of the mixture model by using an expected-maximization (EM) style algorithm. ALC searches for a segmentation of the data by minimizing the code length necessary (with a code based on Gaussian mixtures) to fit the points up to a given distortion. Once more, due to the non-convex nature of these formulations, the dependence upon the key parameters and the noise level is not understood.

Many other methods apply spectral clustering to a specially constructed graph [3,9,14,15,20,46, 49]. They share the same difficulties as stated above and [43] discusses advantages and drawbacks. An approach of this kind is termed *Sparse Curvature Clustering* (SCC) [14, 15], please see also see [2, 3]. This approach is not tractable in the dimension of the subspaces as it requires building a tensor with $N^{(d+2)}$ entries and involves computations with this tensor. Some theoretical guarantees for this algorithm are given in [14] although its limits of performance and its robustness to noise have not been studied. An approach similar to SSC is called *low-rank representation* (LRR) [23]. The LRR algorithm is tractable but its robustness to noise and its dependence upon key parameters is not understood. The work in [22] formulates the robust subspace clustering problem as a non-convex geometric minimization problem over the Grassmanian. Because of the non-convexity, this formulation may not be tractable. On the positive side, this algorithm is provably robust and can accommodate noise levels up to $\mathcal{O}(1/(Ld^{3/2}))$. However, the density ρ required for favorable properties to hold is an unknown function of the dimensions of the subspaces (e.g. ρ could depend on d in a super polynomial fashion). Also, the bound on the noise level seems to decrease as the dimension d and number of subspaces L increases. In contrast, our theory requires $\rho \geq \rho^\star$ where $\rho^\star$ is a fixed numerical constant. While this manuscript was under preparation we learned of [19] which establishes robustness to sparse outliers but with a dependence on the key parameters that is super-polynomial in the dimension of the subspaces demanding $\rho \geq C_0\, d^{\log n}$. (Numerical simulations in [19] seem to indicate that ρ cannot be a constant.)

Finally, the papers [24,35,36] also address regression under corrupted covariates. However, there are three key differences between these studies and our work. First, our results show that LASSO without any change is robust to corrupted covariates whereas these works require modifications to either LASSO or the Dantzig selector. Second, the modeling assumptions for the uncorrupted covariates are significantly different. These papers assume that $\boldsymbol{X}$ has i.i.d. rows and obeys the

restricted eigenvalue condition (REC) whereas we have columns sampled from a mixture model so that the design matrices do not have much in common. Last, for clustering and classification purposes, we need to verify that the support of the solution is correct whereas these works establish closeness to an oracle solution in an ℓ_2 sense. In short, our work is far closer to multiple hypothesis testing.

6 Numerical Experiments

In this section, we perform numerical experiments corroborating our main results and suggesting their applications to temporal segmentation of motion capture data. In this application we are given sensor measurements at multiple joints of the human body captured at different time instants. The goal is to segment the sensory data so that each cluster corresponds to the same activity. Here, each data point corresponds to a vector whose elements are the sensor measurements of different joints at a fixed time instant.

We use the Carnegie Mellon Motion Capture dataset (available at http://mocap.cs.cmu.edu), which contains 149 subjects performing several activities (data are provided in [50]). The motion capture system uses 42 markers per subject. We consider the data from subject 86 in the dataset, consisting of 15 different trials, where each trial comprises multiple activities. We use trials 2 and 5, which feature more activities (8 activities for trial 2 and 7 activities for trial 5) and are, therefore, harder examples relative to the other trials. Figure 7 shows a few snapshots of each activity (walking, squatting, punching, standing, running, jumping, arms-up, and drinking) from trial 2. The right plot in Figure 7 shows the singular values of three of the activities in this trial. Notice that all the curves have a low-dimensional knee, showing that the data from each activity lie in a low-dimensional subspace of the ambient space ($n = 42$ for all the motion capture data).

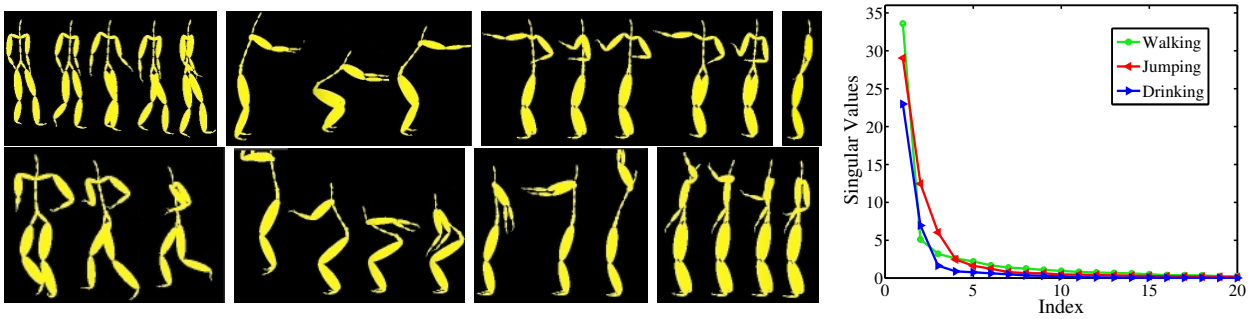
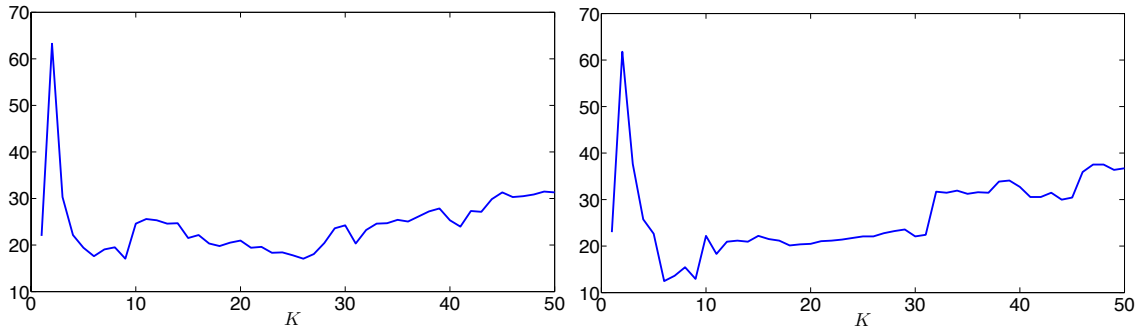


Figure 7: Left: eight activities performed by subject 86 in the CMU motion capture dataset: walking, squatting, punching, standing, running, jumping, arms-up, and drinking. Right: singular values of the data from three activities (walking, jumping, drinking) show that the data from each activity lie approximately in a low-dimensional subspace.

We compare three different algorithms: a baseline algorithm, the two-step procedure and the bias-corrected Dantzig selector. We evaluate these algorithms based on the *clustering error*. That is, we assume knowledge of the number of subspaces and apply spectral clustering to the similarity matrix built by the algorithm. After the spectral clustering step, the clustering error is simply the ratio of misclassified points to the total number of points. We report our results on half of



(a) Trial 2. (b) Trial 5.

Figure 8: Minimum clustering error (%) for each K in the baseline algorithm.

the examples—downsampling the video by a factor 2 keeping every other frame—as to make the problem more challenging. (As a side note, it is always desirable to have methods that work well on a smaller number of examples as one can use split-sample strategies for tuning purposes).[6]

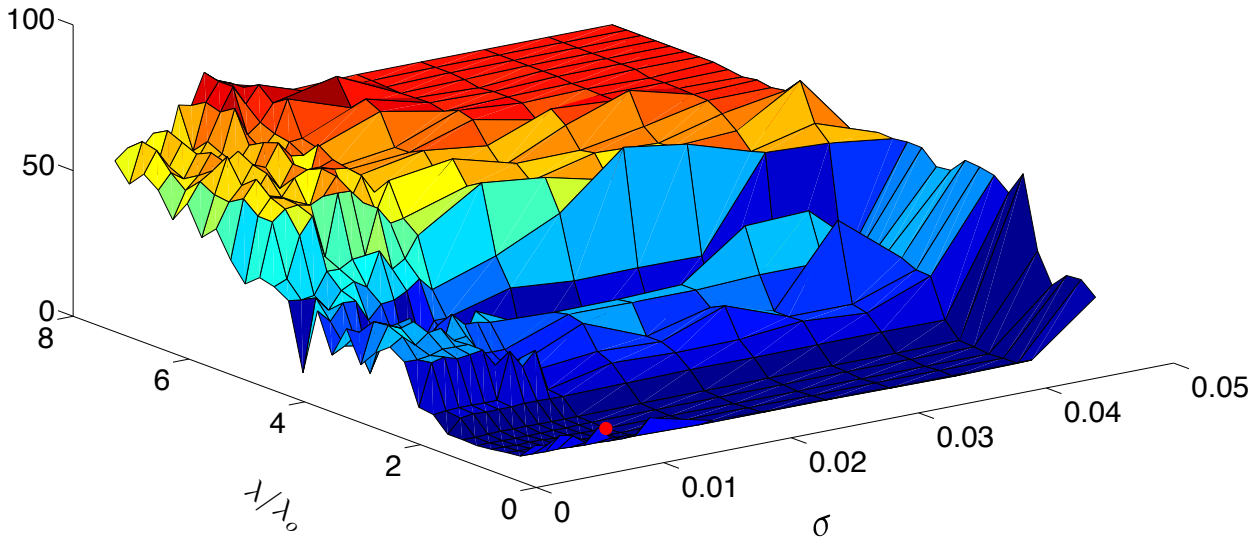
As a baseline for comparison, we apply spectral clustering to a standard similarity graph built by connecting each data point to its K-nearest neighbors. For pairs of data points, $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, that are connected in the K-nearest neighbor graph, we define the similarities between them by $W_{ij} = \exp(-\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2/t)$, where $t > 0$ is a tuning parameter (a.k.a. temperature). For pairs of data points, $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, that are not connected in the K-nearest neighbor graph, we set $W_{ij} = 0$. Thus, pairs of neighboring data points that have small Euclidean distances from each other are considered to be more similar, since they have high similarity W_{ij}. We then apply spectral clustering to the similarity graph and measure the clustering error. For each value of K, we record the minimum clustering error over different choices of the temperature parameter $t > 0$ as shown in Figures 8a and 8b. The minimum clustering error for trials 2 and 5 are 17.06% and 12.47%.

For solving the LASSO problems in the two-step procedure, we developed a computational routine made publicly available [40] based on TFOCS [8] solving the optimization problems in parallel. For the corrected Dantzig selector we use a homotopy solver in the spirit of [41].

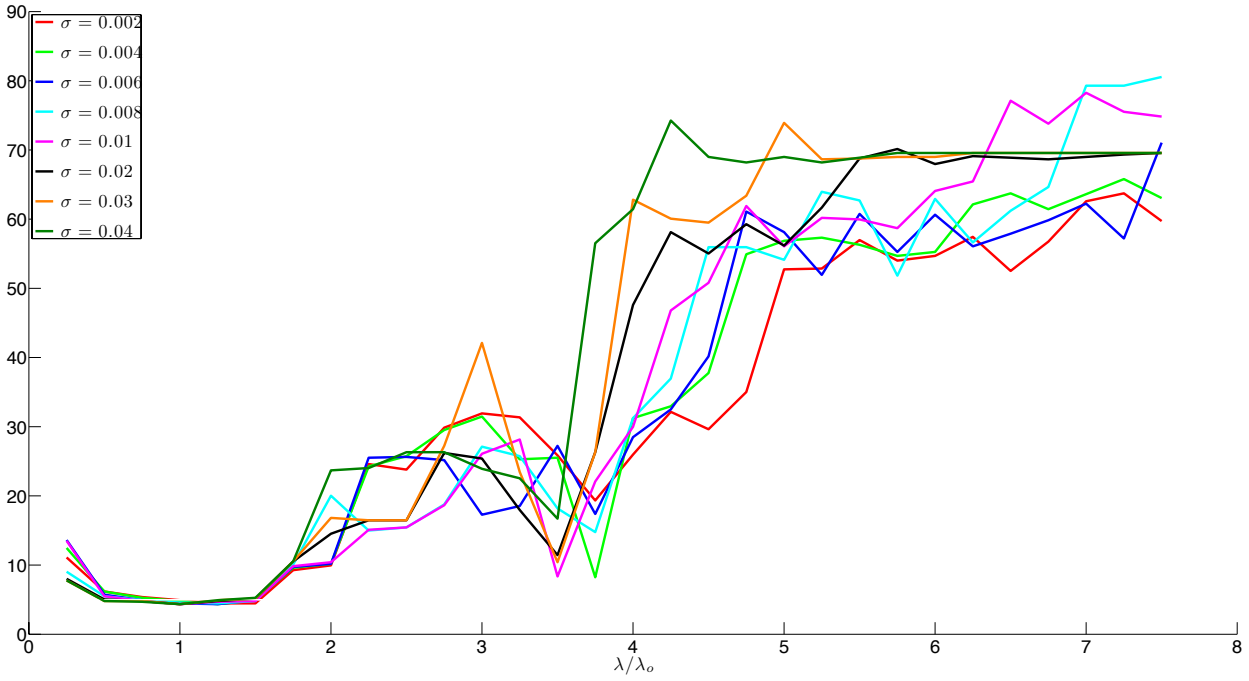
For both the two-step procedure and the bias-corrected Dantzig selector we normalize the data points as a preprocessing step. We work with a noise σ in the interval $[0.001, 0.045]$, and for each value of σ, we vary λ around $1/\lambda_o = 4\,\|\boldsymbol{\beta}^\star\|_{\ell_1}$ and $\lambda_o = \sqrt{2/n}\,\sigma\,\sqrt{1+\sigma^2}$. After building the similarity graph from the sparse regression output, we apply spectral clustering as explained earlier. Figures 9a, 9b and 10 show the clustering error (on trial 5) and the red point indicates the location where the minimum clustering error is reached. Figures 9a and 9b show that for the two-step procedure the value of the clustering error is not overly sensitive to the choice of σ—especially around $\lambda = \lambda_o$. Notice that the clustering error for the robust versions of SSC are significantly lower than the baseline algorithm for a wide range of parameter values. The reason the baseline algorithm performs poorly in this case is that there are many points that are in small Euclidean distances from each other, but belong to different subspaces.

Finally a summary of the clustering errors of these algorithms on the two trials are reported in Table 1. Robust versions of SSC outperform the baseline algorithm. This shows that the multiple subspace model is better for clustering purposes. The two-step procedure seems to work slightly

[6] We have adopted this subsampling strategy to make our experiments reproducible. For tuning purposes, a random strategy may be preferrable.



(a)



(b)

Figure 9: Clustering error (%) for different values of λ and σ on trial 5 using the two step procedure (a) 3D plot (minimum clustering error appears in red). (b) 2D cross sections.

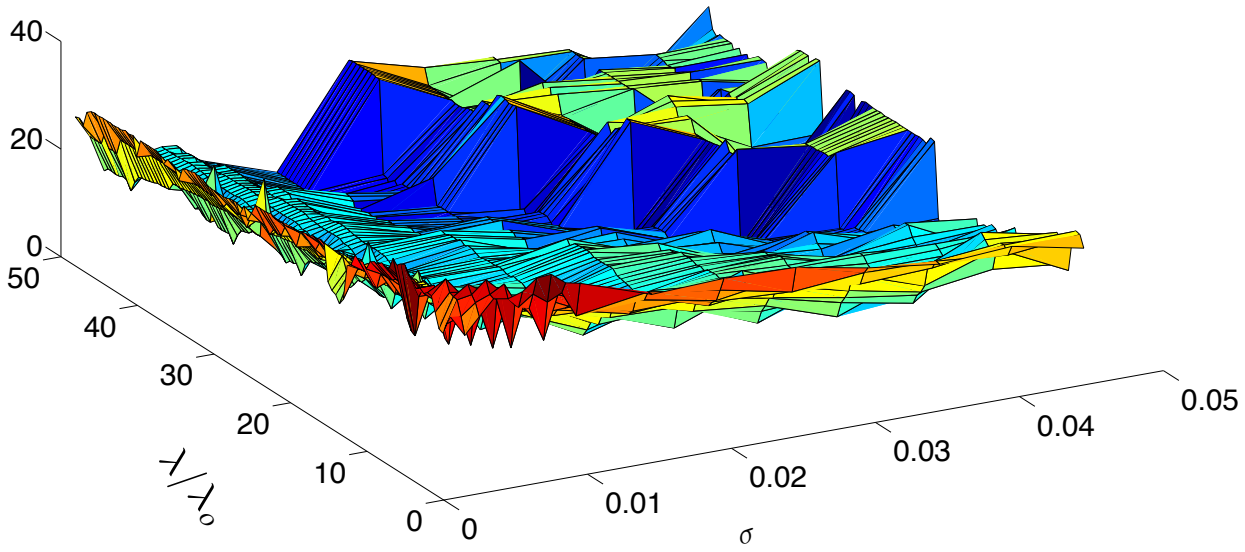


Figure 10: Clustering error (%) for different values of λ and σ on trial 5 using the corrected Dantzig Selector.

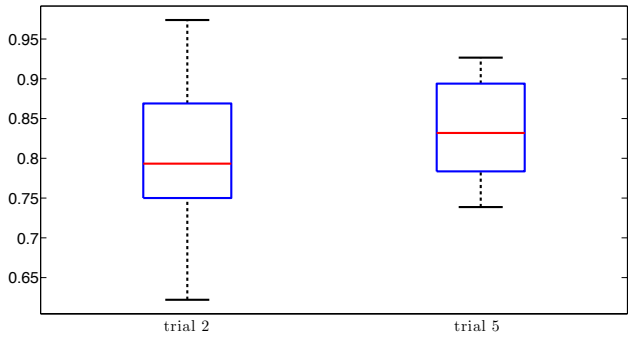


Figure 11: Box plot of the affinities between subspaces for trials 2 and 5.

better than the corrected Dantzig selector for these two examples. Table 2 reports the optimal parameters that achieve the minimum clustering error for each algorithm. The table indicates that on real data, choosing λ close to λ_o also works very well. Also, one can see that in comparison with the synthetic simulations of Section 4, a more conservative choice of the regularization parameter λ is needed for the corrected Dantzig selector as λ needs to be chosen much higher than λ_o to achieve the best results. This may be attributed to the fact that the subspaces in this example are very close to each other and are not drawn at random as was the case with our synthetic data. To get a sense of the affinity values, we fit a subspace of dimension d_ℓ to the N_ℓ data points from the ℓth group, where d_ℓ is chosen as the smallest nonnegative integer such that the partial sum of the d_ℓ top singular values is at least 90% of the total sum. Figure 11 shows that the affinities are higher than 0.75 for both trials.

	Baseline Algorithm	Two-step procedure	Corrected Dantzig selector
Trial 2	17.06%	**3.54%**	9.53%
Trial 5	12.47%	**4.35%**	4.92%

Table 1: Minimum clustering error.

	Baseline algorithm	Two-step procedure	Corrected Dantzig selector
Trial 2	K=9, t=0.0769	$\sigma = 0.03$, $\lambda = 1.25\lambda_o$	$\sigma = 0.004$, $\lambda = 41.5\lambda_o$
Trial 5	K=6, t=0.0455	$\sigma = 0.01$, $\lambda = \lambda_o$	$\sigma = 0.03$, $\lambda = 45.5\lambda_o$

Table 2: Optimal parameters.

7 Discussion and Open Problems

In this paper, we have developed a tractable algorithm that can provably cluster data points in a fairly challenging regime in which subspaces can overlap along many dimensions and in which the number of points per subspace is rather limited. Our results about the performance of the robust SSC algorithm are expressed in terms of interpretable parameters. This is not a trivial achievement: one of the challenges of the theory for subspace clustering is precisely that performance depends on many different aspects of the problem such as the dimension of the ambient space, the number of subspaces, their dimensions, their relative orientations, the distribution of points around each subspace, the noise level, and so on. Nevertheless, these results only offer a starting point as our work leaves open lots of questions, and at the same time, suggests topics for future research. Before presenting the proofs, we would like to close by listing a few questions colleagues may find of interest.

- We have shown that while having the affinities and sampling densities near what is information theoretically possible, robust versions of SSC that can accommodate noise levels σ of order one exist. It would be interesting to establish fundamental limits relating the key parameters to the maximum allowable noise level. What is the maximum allowable noise level for any algorithm regardless of tractability?

- It would be interesting to extend the results of this paper to a deterministic model where both the orientation of the subspaces and the noiseless samples are non-random. We leave this to a future publication.

- Our work in this paper concerns the construction of the similarity matrix and the correctness of sparse regression techniques. The full algorithm then applies clustering techniques to clean up errors introduced in the first step. It would be interesting to develop theoretical guarantees for this step as well. A potential approach is the interesting formulation developed in [5].

- A natural direction is the development of clustering techniques that can provably operate with missing and/or sparsely corrupted entries (the work [37] only deals with grossly corrupted columns). The work in [19] provides one possible approach but requires a very high sampling density as we already mentioned. The paper [18] develops another heuristic approach without any theoretical justification.

- One of the advantages of the suggested scheme is that it is highly parallelizable. When the algorithm is run sequentially, it would be interesting to see whether one can reuse computations to solve all the ℓ_1-minimization problems more effectively.

8 Proofs

We prove all of our results in this section. Before we begin, we introduce some notation. If $\boldsymbol{A}$ is a matrix with N columns and T a subset of $\{1,\dots,N\}$, $\boldsymbol{A}_T$ is the submatrix with columns in T. Similarly, $\boldsymbol{x}_T$ is the restriction of the vector $\boldsymbol{x}$ to indices in T. Throughout we use L_m to denote $\log m$ up to a fixed numerical constant. The value of this constant may change from line to line. Likewise, C is a generic numerical constant whose value may change at each occurrence.

Next, we work with $\boldsymbol{y} := \boldsymbol{y}_1$ for convenience, assumed to originate from S_1, which is no loss of generality. It is also convenient to partition $\boldsymbol{Y}$ as $\{\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}, \dots, \boldsymbol{Y}^{(L)}\}$, where for each ℓ, $\boldsymbol{Y}^{(\ell)}$ are those noisy columns from subspace S_ℓ; when $\ell = 1$, we exclude the response $\boldsymbol{y}_1$ from $\boldsymbol{Y}^{(1)}$. With this notation, the problem (2.2) takes the form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{N-1}} \quad \frac{1}{2}\left\|\boldsymbol{y} - \left(\boldsymbol{Y}^{(1)}\boldsymbol{\beta}^{(1)} + \ldots + \boldsymbol{Y}^{(L)}\boldsymbol{\beta}^{(L)}\right)\right\|_{\ell_2}^2 + \lambda\left\|\boldsymbol{\beta}^{(1)}\right\|_{\ell_1} + \ldots + \lambda\left\|\boldsymbol{\beta}^{(L)}\right\|_{\ell_1}. \tag{8.1}$$

Throughout, $\boldsymbol{y}_\parallel/\boldsymbol{Y}_\parallel^{(1)}$ denotes the projection of the vector/matrix $\boldsymbol{y}/\boldsymbol{Y}^{(1)}$ onto S_1. Similarly, we use $\boldsymbol{y}_\perp$ to denote projection onto the orthogonal complement $S_1^\perp$; hence, $\boldsymbol{y} = \boldsymbol{y}_\parallel + \boldsymbol{y}_\perp$ and $\boldsymbol{Y}^{(1)} = \boldsymbol{Y}_\parallel^{(1)} + \boldsymbol{Y}_\perp^{(1)}$. Moreover, $\boldsymbol{U}_1 \in \mathbb{R}^{n\times d}$ and $\boldsymbol{U}_1^\perp \in \mathbb{R}^{n\times(n-d)}$ are orthonormal bases for S_1 and $S_1^\perp$.

Since $\boldsymbol{y}_\parallel = \boldsymbol{x} + \boldsymbol{z}_\parallel$ with $\|\boldsymbol{x}\|_{\ell_2} = 1$ and $\mathbb{E}\,\|\boldsymbol{z}_\parallel\|_{\ell_2}^2 = \sigma^2 d/n$, it is obvious that under the stated assumptions, $\|\boldsymbol{y}_\parallel\|_{\ell_2} \in [3/4, 5/4]$ with very high probability as shown in Lemma A.4. The same applies to all the columns of $\boldsymbol{Y}_\parallel^{(1)}$. From now on, we will operate under these two assumptions, which hold simultaneously over an event of probability at least $1 - 1/N^2$.

8.1 Intermediate results

In this section, we record a few important results that shall be used to establish the no-false and many true discoveries theorems. Now the reader interested in our proofs may first want to pass over this section rather quickly, and return to it once it is clear how our arguments reduce to the technical lemmas below.

8.1.1 Preliminaries

Our first lemma rephrases Lemma 7.5 in [37] and bounds the size of the dot product between random vectors. We omit the proof.

Lemma 8.1 *Let $\boldsymbol{A} \in \mathbb{R}^{d_1\times N_1}$ be a matrix with columns sampled uniformly at random from the unit sphere of $\mathbb{R}^{d_1}$, $\boldsymbol{w} \in \mathbb{R}^{d_2}$ be a vector sampled uniformly at random from the unit sphere of $\mathbb{R}^{d_2}$ and independent of $\boldsymbol{A}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d_1\times d_2}$ be a deterministic matrix. We have*

$$\left\|\boldsymbol{A}^T\boldsymbol{\Sigma}\boldsymbol{w}\right\|_{\ell_\infty} \le \sqrt{\log a \log b}\,\frac{\|\boldsymbol{\Sigma}\|_F}{\sqrt{d_1}\sqrt{d_2}}, \tag{8.2}$$

with probability at least $1 - \frac{2}{\sqrt{a}} - \frac{2N_1}{\sqrt{b}}$.

We are interested in this because (8.2) relates the size of the dot products with the affinity between subspaces as follows: suppose the unit-norm vector $\boldsymbol{x}_i$ is drawn uniformally at random from S_i, then

$$\boldsymbol{X}^{(j)^T}\boldsymbol{x}_i = \boldsymbol{A}^T\boldsymbol{\Sigma}\boldsymbol{w};$$

$\boldsymbol{A}$ and $\boldsymbol{w}$ are as in the lemma and $\boldsymbol{\Sigma} = \boldsymbol{U}^{(j)^T}\boldsymbol{U}^{(i)}$, where $\boldsymbol{U}^{(j)}$ (resp. $\boldsymbol{U}^{(i)}$) is an orthobasis for S_j (resp. S_i). By definition, $\|\boldsymbol{\Sigma}\|_F = \sqrt{d_j \wedge d_i}\,\mathrm{aff}(S_j, S_i)$.

8.1.2 The first step of Algorithm 2

As claimed in Section 2, the first step of Algorithm 2 returns an optimal value that is a reasonable proxy for the unknown dimension.

Lemma 8.2 *Let* Val(Step 1)*be the optimal value of* (2.4) *with* $\tau = 2\sigma$*. Assume* $\rho_1 > \rho^\star$ *as earlier. Then*

$$\frac{1}{10}\sqrt{\frac{d_1}{\log \rho_1}} \le \mathrm{Val}(\text{Step 1}) \le 2\sqrt{d_1}. \tag{8.3}$$

The upper bound holds with probability at least $1 - e^{\gamma_1 n} - e^{-\gamma_2 d_1}$*. The lower bound holds with probability at least* $1 - e^{-\gamma_3 d_1} - \frac{10}{N^2}$*.*

Proof We begin with the upper bound. Let $\boldsymbol{\beta}_0 = \boldsymbol{X}^{(1)^T}\left(\boldsymbol{X}^{(1)}\boldsymbol{X}^{(1)^T}\right)^{-1}\boldsymbol{x}$ be the minimum ℓ_2-norm solution to the noiseless problem $\boldsymbol{X}\boldsymbol{\beta}_0 = \boldsymbol{x}$. We show that $\boldsymbol{\beta}_0$ is feasible for (2.4). We have

$$\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}_0 = \boldsymbol{x} - \boldsymbol{X}\boldsymbol{\beta}_0 + (\boldsymbol{z} - \boldsymbol{Z}\boldsymbol{\beta}_0) = \boldsymbol{z} - \boldsymbol{Z}\boldsymbol{\beta}_0,$$

which gives

$$\mathcal{L}(\boldsymbol{z} - \boldsymbol{Z}\boldsymbol{\beta}_0 \,|\, \boldsymbol{X}, \boldsymbol{x}) = \mathcal{N}(0, V\boldsymbol{I}_n), \quad V = (1 + \|\boldsymbol{\beta}_0\|_{\ell_2}^2)\sigma^2/n$$

(the notation $\mathcal{L}(Y|X)$ is the conditional law of Y given X). Hence, the conditional distribution of $\|\boldsymbol{z} - \boldsymbol{Z}\boldsymbol{\beta}_0\|_{\ell_2}^2$ is that of a chi square and (A.1) gives

$$\|\boldsymbol{z} - \boldsymbol{Z}\boldsymbol{\beta}_0\|_{\ell_2} \le \sqrt{2(1 + \|\boldsymbol{\beta}_0\|_{\ell_2}^2)}\sigma$$

with probability at least $1 - e^{-\gamma_1 n}$. On the other hand,

$$\|\boldsymbol{\beta}_0\|_{\ell_2} \le \frac{\|\boldsymbol{x}\|_{\ell_2}}{\sigma_{\min}(\boldsymbol{X}^{(1)})}$$

and applying Lemma A.2 gives

$$\|\boldsymbol{\beta}_0\|_{\ell_2} \le \frac{1}{\sqrt{\frac{N_1}{d_1}} - 2},$$

which holds with probability at least $1 - e^{-\gamma_2 d_1}$. If $N_1 > 9d_1$, then $\|\boldsymbol{\beta}_0\|_{\ell_2} \le 1$ and thus $\boldsymbol{\beta}_0$ is feasible. Therefore,

$$\|\boldsymbol{\beta}^\star\|_{\ell_1} \le \|\boldsymbol{\beta}_0\|_{\ell_1} \le \sqrt{N_1}\,\|\boldsymbol{\beta}_0\|_{\ell_2} \le \frac{\sqrt{d_1}}{1 - 2\sqrt{\frac{d_1}{N_1}}} \le 2\sqrt{d_1},$$

where the last inequality holds provided that $N_1 \ge 16d_1$.

We now turn to the lower bound and let $\boldsymbol{\beta}^\star$ be an optimal solution. Notice that $\|\boldsymbol{y}_\| - \boldsymbol{Y}_\|\boldsymbol{\beta}^\star\|_{\ell_2} \le \|\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{\beta}^\star\|_{\ell_2} \le 2\sigma$ so that $\boldsymbol{\beta}^\star$ is feasible for

$$\min \ \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \left\|\boldsymbol{y}_\| - \boldsymbol{Y}_\|\boldsymbol{\beta}\right\|_{\ell_2} \le 2\sigma. \tag{8.4}$$

We bound the optimal value of this program from below. The dual of (8.4) is

$$\max \ \langle \boldsymbol{y}_\|, \boldsymbol{\nu}\rangle - 2\sigma \|\boldsymbol{\nu}\|_{\ell_2} \quad \text{subject to} \quad \left\|\boldsymbol{Y}_\|^T\boldsymbol{\nu}\right\|_{\ell_\infty} \le 1. \tag{8.5}$$

Slater's condition holds and the primal and dual optimal values are equal. To simplify notation set $\boldsymbol{A} = \boldsymbol{Y}_\|^{(1)}$. Define

$$\boldsymbol{\nu}^\star \in \arg\max \ \langle \boldsymbol{y}_\|, \boldsymbol{\nu}\rangle \quad \text{subject to} \quad \left\|\boldsymbol{A}^T\boldsymbol{\nu}\right\|_{\ell_\infty} \le 1.$$

Notice that $\boldsymbol{\nu}^\star$ has random direction on subspace S_1. Therefore, combining Lemmas 8.1, A.1, and B.4 together with the affinity condition implies that for $\ell \neq 1$, $\|\boldsymbol{Y}_\|^{(\ell)^T}\boldsymbol{\nu}^*\|_{\ell_\infty} \le 1$ with high probability. In short, $\boldsymbol{\nu}^\star$ is feasible for (8.5).

Since $\boldsymbol{y}_\|$ has random direction, the arguments (with $t = 1/6$) in Step 2 of the proof of Theorem 2.9 in [37] give

$$\langle \boldsymbol{y}_\|, \boldsymbol{\nu}^\star\rangle \ge \frac{1}{\sqrt{2\pi e}}\sqrt{\frac{d_1}{\log \rho_1}}.$$

Also, by Lemma B.4,

$$\|\boldsymbol{\nu}^\star\|_{\ell_2} \le \frac{16}{3}\sqrt{\frac{d_1}{\log \rho_1}}.$$

Since $\boldsymbol{\nu}^\star$ is feasible for (8.5), the optimal value of (8.5) is greater or equal than

$$\langle \boldsymbol{y}_\|, \boldsymbol{\nu}^\star\rangle - 2\sigma \|\boldsymbol{\nu}^\star\|_{\ell_2} \ge \frac{1}{10}\sqrt{\frac{d_1}{\log \rho_1}},$$

where the inequality follows from the bound on the noise level. This concludes the proof. ∎

8.1.3 The reduced and projected problems

When there are no false discoveries, the solution to (8.1) coincides with that of the *reduced problem*

$$\hat{\boldsymbol{\beta}}^{(1)} \in \arg\min \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\boldsymbol{\beta}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\boldsymbol{\beta}^{(1)}\right\|_{\ell_1}. \tag{8.6}$$

Not surprisingly, we need to analyze the properties of the solution to this problem. In particular, we would like to understand something about the orientation and the size of the residual vector $\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}$.

A problem close to (8.6) is the *projected problem*

$$\tilde{\boldsymbol{\beta}}^{(1)} \in \arg\min \frac{1}{2}\left\|\boldsymbol{y}_\| - \boldsymbol{Y}_\|^{(1)}\boldsymbol{\beta}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\boldsymbol{\beta}^{(1)}\right\|_{\ell_1}. \tag{8.7}$$

The difference with the reduced problem is that the goodness of fit only involves the residual sum of squares of the projected residuals. Intuitively, the solutions to the two problems (8.6) and (8.7) should be close. Our strategy is to gain some insights about the solution to the reduced problem by studying the properties of the projected problem.

8.1.4 Properties of the projected problem

The sole purpose of this subsection is to state this:

Lemma 8.3 *Let $\tilde{\boldsymbol{\beta}}^{(1)}$ be any solution to the projected problem and assume that $N_1/d_1 \geq \rho^\star$ as before. Then there exists an absolute constant C such that for all $\lambda > 0$,*

$$\left\|\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \leq C \tag{8.8}$$

holds with probability at least $1 - 5e^{-\gamma_1 d_1} - e^{-\sqrt{N_1 d_1}}$.

This estimate shall play a crucial role in our arguments. It is a consequence of sharp estimates obtained by Wojtaszczyk [45] in the area of compressed sensing. As not to interrupt the flow, we postpone its proof to Section 8.3.

In the asymptotic regime ($\rho_1 = N_1/d_1$ fixed and $d_1 \to \infty$), one can sharpen the upper bound (8.8) by taking $C = 1$. This leverages the asymptotic theory developed in [6] and [7] as explained in Appendix C.

8.1.5 Properties of the reduced problem

We now collect two important facts about the residuals to the reduced problem. The first concerns their orientation.

Lemma 8.4 (Isotropy of the residuals) *The projection of the residual vector $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}$ onto either S_1 or $S_1^\perp$ has uniform orientation.*

Proof Consider any unitary transformation $\boldsymbol{U}^\parallel$ (resp. $\boldsymbol{U}^\perp$) leaving S_1 (resp. $S_1^\perp$) invariant. Since

$$\begin{aligned}\frac{1}{2}\left\|\boldsymbol{U}^\parallel(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\boldsymbol{\beta}^{(1)})\right\|_{\ell_2}^2 + \frac{1}{2}\left\|\boldsymbol{U}^\perp(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\boldsymbol{\beta}^{(1)})\right\|_{\ell_2}^2 + \lambda\left\|\boldsymbol{\beta}^{(1)}\right\|_{\ell_1} \\ = \frac{1}{2}\left\|\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\boldsymbol{\beta}^{(1)}\right\|_{\ell_2}^2 + \frac{1}{2}\left\|\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\boldsymbol{\beta}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\boldsymbol{\beta}^{(1)}\right\|_{\ell_1},\end{aligned}$$

the LASSO functional is invariant and this gives

$$\hat{\boldsymbol{\beta}}^{(1)}(\boldsymbol{U}^\parallel\boldsymbol{y}_\parallel, \boldsymbol{U}^\perp\boldsymbol{y}_\perp, \boldsymbol{U}^\parallel\boldsymbol{Y}_\parallel^{(1)}, \boldsymbol{U}^\perp\boldsymbol{Y}_\perp^{(1)}) = \hat{\boldsymbol{\beta}}^{(1)}(\boldsymbol{y}_\parallel, \boldsymbol{y}_\perp, \boldsymbol{Y}_\parallel^{(1)}, \boldsymbol{Y}_\perp^{(1)}).$$

Let $\boldsymbol{r}^\parallel(\boldsymbol{y}_\parallel, \boldsymbol{Y}_\parallel^{(1)}) = \boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}$ and $\boldsymbol{r}^\perp(\boldsymbol{y}_\perp, \boldsymbol{Y}_\perp^{(1)}) = \boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)}$ be the projections of the residuals. Since $\boldsymbol{y}_\parallel$ and $\boldsymbol{Y}_\parallel^{(1)}$ are invariant under rotations leaving S_1 invariant, we have

$$\begin{aligned}\boldsymbol{r}^\parallel(\boldsymbol{U}^\parallel\boldsymbol{y}_\parallel, \boldsymbol{U}^\parallel\boldsymbol{Y}_\parallel^{(1)}) &= \boldsymbol{U}^\parallel\boldsymbol{y}_\parallel - \boldsymbol{U}^\parallel\boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}(\boldsymbol{U}^\parallel\boldsymbol{y}_\parallel, \boldsymbol{U}^\perp\boldsymbol{y}_\perp, \boldsymbol{U}^\parallel\boldsymbol{Y}_\parallel^{(1)}, \boldsymbol{U}^\perp\boldsymbol{Y}_\perp^{(1)}) \\ &= \boldsymbol{U}^\parallel\boldsymbol{y}_\parallel - \boldsymbol{U}^\parallel\boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}(\boldsymbol{y}_\parallel, \boldsymbol{y}_\perp, \boldsymbol{Y}_\parallel^{(1)}, \boldsymbol{Y}_\perp^{(1)}) \\ &\sim \boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}(\boldsymbol{y}_\parallel, \boldsymbol{y}_\perp, \boldsymbol{Y}_\parallel^{(1)}, \boldsymbol{Y}_\perp^{(1)}) \\ &= \boldsymbol{r}^\parallel(\boldsymbol{y}_\parallel, \boldsymbol{Y}_\parallel^{(1)}),\end{aligned}$$

where $X \sim Y$ means that the random variables X and Y have the same distribution. Therefore, the distribution of $\boldsymbol{r}^\parallel(\boldsymbol{y}_\parallel, \boldsymbol{Y}_\parallel^{(1)}) = \boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}$ is invariant under rotations leaving S_1 invariant.

In other words, the projection $\boldsymbol{r}^{\|}$ has uniform orientation. In a similar manner we conclude that $\boldsymbol{r}^{\perp}$ has uniform orientation has well. ∎

The next result controls the size of the residuals.

Lemma 8.5 (Size of residuals) *If $N_1/d_1 \geq \rho^\star$, then for all $\lambda > 0$,*

$$\left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \leq C\,\sigma. \tag{8.9}$$

Also,

$$\left\|\boldsymbol{y}_{\|} - \boldsymbol{Y}_{\|}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \leq \frac{32}{3}\lambda\sqrt{\frac{d_1}{\log(N_1/d_1)}} + C\sigma. \tag{8.10}$$

Both these inequalities hold with probability at least $1 - e^{-\gamma_1(n-d_1)} - 5e^{-\gamma_2 d_1} - e^{-\sqrt{N_1 d_1}}$, where γ_1 and γ_2 are fixed numerical constants. Thus if $\lambda > \sigma/\sqrt{8d_1}$, then

$$\left\|\boldsymbol{y}_{\|} - \boldsymbol{Y}_{\|}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \leq C\lambda\sqrt{d_1}. \tag{8.11}$$

Proof We begin with (8.9). Since $\hat{\boldsymbol{\beta}}^{(1)}$ is optimal for the reduced problem,

$$\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1} \leq \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1}.$$

Conversely, since $\tilde{\boldsymbol{\beta}}^{(1)}$ is optimal for the projected problem,

$$\frac{1}{2}\left\|\boldsymbol{y}_{\|} - \boldsymbol{Y}_{\|}^{(1)}\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1} \leq \frac{1}{2}\left\|\boldsymbol{y}_{\|} - \boldsymbol{Y}_{\|}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1}.$$

Now Parseval equality

$$\left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 = \left\|\boldsymbol{y}_{\|} - \boldsymbol{Y}_{\|}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2$$

(and similarly for $\tilde{\boldsymbol{\beta}}^{(1)}$) together with the last two inequalities give

$$\left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \leq \left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}.$$

Now observe that $\boldsymbol{y}_{\|}$, $\boldsymbol{y}_{\perp}$, $\boldsymbol{Y}_{\|}^{(1)}$ and $\boldsymbol{Y}_{\perp}^{(1)}$ are all independent from each other. Since $\tilde{\boldsymbol{\beta}}^{(1)}$ is a function of $\boldsymbol{y}_{\|}$ and $\boldsymbol{Y}_{\|}^{(1)}$, it is independent from $\boldsymbol{y}_{\perp}$ and $\boldsymbol{Y}_{\perp}^{(1)}$. Hence,

$$\mathcal{L}(\boldsymbol{U}_1^{\perp T}(\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\tilde{\boldsymbol{\beta}}^{(1)}) \,|\, \boldsymbol{y}_{\|}, \boldsymbol{Y}_{\|}^{(1)}) = \mathcal{N}(0, V\,\boldsymbol{I}_{n-d_1}), \quad V = \frac{\sigma^2}{n}\Big(1 + \left\|\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2\Big).$$

Conditionally then, it follows from the chi-square tail bound (A.1) with $\epsilon = 1$ that

$$\left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 \leq 2\sigma^2\Big(1 - \frac{d_1}{n}\Big)\Big(1 + \left\|\tilde{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2\Big),$$

which holds with probability at least $1 - e^{-\gamma_1(n-d_1)}$, where $\gamma_1 = \frac{(1-\log 2)}{2}$. Unconditionally, our first claim follows from Lemma 8.3.

We now turn our attention to (8.10). Our argument uses the solution to the *noiseless* projected problem

$$\bar{\boldsymbol{\beta}}^{(1)} \in \arg\min \left\|\boldsymbol{\beta}^{(1)}\right\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{y}_{\|} = \boldsymbol{Y}_{\|}^{(1)}\boldsymbol{\beta}^{(1)}; \tag{8.12}$$

this is the solution to the projected problem as $\lambda \to 0^+$. With this, we proceed until (8.13) as in [11, 12]. Since $\hat{\boldsymbol{\beta}}^{(1)}$ is optimal for the reduced problem,

$$\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1} \le \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \lambda\left\|\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1}.$$

Put $\boldsymbol{h} = \hat{\boldsymbol{\beta}}^{(1)} - \bar{\boldsymbol{\beta}}^{(1)}$. Standard simplifications give

$$\frac{1}{2}\left\|\boldsymbol{Y}^{(1)}\boldsymbol{h}\right\|_{\ell_2}^2 + \lambda\left\|\bar{\boldsymbol{\beta}}^{(1)} + \boldsymbol{h}\right\|_{\ell_1} \le \langle \boldsymbol{y} - \boldsymbol{Y}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}, \boldsymbol{Y}^{(1)}\boldsymbol{h}\rangle + \lambda\left\|\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1}.$$

Letting S be the support of $\bar{\boldsymbol{\beta}}^{(1)}$, we have

$$\left\|\bar{\boldsymbol{\beta}}^{(1)} + \boldsymbol{h}\right\|_{\ell_1} = \left\|\bar{\boldsymbol{\beta}}_S + \boldsymbol{h}_S\right\|_{\ell_1} + \|\boldsymbol{h}_{S^c}\|_{\ell_1} \ge \left\|\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1} + \langle \operatorname{sgn}(\bar{\boldsymbol{\beta}}_S), \boldsymbol{h}_S\rangle + \|\boldsymbol{h}_{S^c}\|_{\ell_1}.$$

This yields

$$\frac{1}{2}\left\|\boldsymbol{Y}^{(1)}\boldsymbol{h}\right\|_{\ell_2}^2 + \lambda\|\boldsymbol{h}_{S^c}\|_{\ell_1} \le \langle \boldsymbol{y} - \boldsymbol{Y}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}, \boldsymbol{Y}^{(1)}\boldsymbol{h}\rangle - \lambda\langle \operatorname{sgn}(\bar{\boldsymbol{\beta}}_S^{(1)}), \boldsymbol{h}_S\rangle.$$

By definition, $\boldsymbol{y}_{\|} - \boldsymbol{Y}_{\|}^{(1)}\bar{\boldsymbol{\beta}}^{(1)} = 0$, and thus

$$\frac{1}{2}\left\|\boldsymbol{Y}^{(1)}\boldsymbol{h}\right\|_{\ell_2}^2 + \lambda\|\boldsymbol{h}_{S^c}\|_{\ell_1} \le \langle \boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}, \boldsymbol{Y}_{\perp}^{(1)}\boldsymbol{h}\rangle - \lambda\langle \operatorname{sgn}(\bar{\boldsymbol{\beta}}_S^{(1)}), \boldsymbol{h}_S\rangle. \tag{8.13}$$

Continue with

$$\langle \boldsymbol{Y}_{\perp}^{(1)}\boldsymbol{h}, \boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}\rangle \le \left\|\boldsymbol{Y}_{\perp}^{(1)}\boldsymbol{h}\right\|_{\ell_2}\left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \le \frac{1}{2}\left\|\boldsymbol{Y}_{\perp}^{(1)}\boldsymbol{h}\right\|_{\ell_2}^2 + \frac{1}{2}\left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2$$

so that

$$\frac{1}{2}\left\|\boldsymbol{Y}_{\|}^{(1)}\boldsymbol{h}\right\|_{\ell_2}^2 + \lambda\|\boldsymbol{h}_{S^c}\|_{\ell_1} \le \frac{1}{2}\left\|\boldsymbol{y}_{\perp} - \boldsymbol{Y}_{\perp}^{(1)}\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 - \lambda\langle \operatorname{sgn}(\bar{\boldsymbol{\beta}}_S^{(1)}), \boldsymbol{h}_S\rangle. \tag{8.14}$$

Now set $\boldsymbol{A} = \boldsymbol{Y}_{\|}^{(1)}$ for notational convenience. Since $\bar{\boldsymbol{\beta}}^{(1)}$ is optimal, there exists $\boldsymbol{\nu}$ such that

$$\boldsymbol{v} = \boldsymbol{A}^T\boldsymbol{\nu}, \quad \boldsymbol{v}_S = \operatorname{sgn}(\bar{\boldsymbol{\beta}}_S^{(1)}) \text{ and } \|\boldsymbol{v}_{S^c}\|_{\ell_\infty} \le 1.$$

Also, Corollary B.4 gives

$$\|\boldsymbol{\nu}\|_{\ell_2}^2 \le \frac{256}{9}\frac{d_1}{\log(N_1/d_1)}. \tag{8.15}$$

With this

$$\langle \operatorname{sgn}(\bar{\boldsymbol{\beta}}_S^{(1)}), \boldsymbol{h}_S\rangle = \langle \boldsymbol{v}_S, \boldsymbol{h}_S\rangle = \langle \boldsymbol{\nu}, \boldsymbol{A}\boldsymbol{h}\rangle - \langle \boldsymbol{v}_{S^c}, \boldsymbol{h}_{S^c}\rangle.$$

We have

$$|\langle \boldsymbol{\nu}, \boldsymbol{A}\boldsymbol{h}\rangle| \le \|\boldsymbol{A}\boldsymbol{h}\|_{\ell_2}\|\boldsymbol{\nu}\|_{\ell_2} \le \frac{1}{4\lambda}\|\boldsymbol{A}\boldsymbol{h}\|_{\ell_2}^2 + \lambda\|\boldsymbol{\nu}\|_{\ell_2}^2$$

and, therefore,

$$\lambda|\langle \mathrm{sgn}(\bar{\boldsymbol{\beta}}_S^{(1)}), \boldsymbol{h}_S\rangle| \le \frac{1}{4}\|\boldsymbol{A}\boldsymbol{h}\|_{\ell_2}^2 + \lambda^2\|\boldsymbol{\nu}\|_{\ell_2}^2 + \lambda\|\boldsymbol{h}_{S^c}\|_{\ell_1}.$$

Plugging this into (8.14), we obtain

$$\frac{1}{4}\|\boldsymbol{A}\boldsymbol{h}\|_{\ell_2}^2 \le \frac{1}{2}\left\|\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\bar{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}^2 + \lambda^2\|\boldsymbol{\nu}\|_{\ell_2}^2. \tag{8.16}$$

This concludes the proof since by definition $\boldsymbol{A}\boldsymbol{h} = \boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}$ and since we already know that $\|\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\bar{\boldsymbol{\beta}}^{(1)}\|_{\ell_2}^2 \le C^2\sigma^2$. ∎

8.2 Proof of Theorem 3.1

First, Lemma 8.2 asserts that by using τ and $f(t)$ as stated, our choice of λ obeys

$$\lambda > \frac{\sigma}{\sqrt{8d_1}}. \tag{8.17}$$

All we need is to demonstrate that when λ is as above, there are no false discoveries. To do this, it is sufficient to establish that the solution $\hat{\boldsymbol{\beta}}^{(1)}$ to the *reduced problem* obeys

$$\left\|\boldsymbol{Y}^{(\ell)^T}(\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)})\right\|_{\ell_\infty} < \lambda, \quad \text{for all } \ell \ne 1. \tag{8.18}$$

This is a consequence of this:

Lemma 8.6 *Fix $\boldsymbol{A} \in \mathbb{R}^{d\times N}$ and $T \subset \{1,2,\dots,N\}$. Suppose that there is a solution $\boldsymbol{x}^\star$ to*

$$\min \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 + \lambda\|\boldsymbol{x}\|_{\ell_1} \quad \textit{subject to} \quad \boldsymbol{x}_{T^c} = \boldsymbol{0}$$

obeying $\|\boldsymbol{A}_{T^c}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star)\|_{\ell_\infty} < \lambda$. Then any optimal solution $\hat{\boldsymbol{x}}$ to

$$\min \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 + \lambda\|\boldsymbol{x}\|_{\ell_1}$$

must also satisfy $\hat{\boldsymbol{x}}_{T^c} = \boldsymbol{0}$.

Proof Consider a perturbation $\boldsymbol{x}^\star + t\boldsymbol{h}$. For $t > 0$ sufficiently small, the value of the LASSO functional at this point is equal to

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x}^\star + t\boldsymbol{h})\|_{\ell_2}^2 + \lambda\|\boldsymbol{x} + t\boldsymbol{h}\|_{\ell_1} = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star\|_{\ell_2}^2 - t\langle \boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star), \boldsymbol{h}\rangle + \frac{t^2}{2}\|\boldsymbol{A}\boldsymbol{h}\|_{\ell_2}^2 + \lambda\|\boldsymbol{x}\|_{\ell_1} + \lambda t\langle \mathrm{sgn}(\boldsymbol{x}_T), \boldsymbol{h}_T\rangle + \lambda t\|\boldsymbol{h}_{T^c}\|_{\ell_1}.$$

Now since the optimality conditions give that $\boldsymbol{A}_T^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star) = \lambda\mathrm{sgn}(\boldsymbol{x}_T)$ and that by assumption, $\boldsymbol{A}_{T^c}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star) = \lambda\epsilon_{T^c}$ with $\|\epsilon_{T^c}\|_{\ell_\infty} < 1$, the value of the LASSO functional is equal to

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^\star\|_{\ell_2}^2 + \lambda\|\boldsymbol{x}\|_{\ell_1} + \frac{t^2}{2}\|\boldsymbol{A}\boldsymbol{h}\|_{\ell_2}^2 + \lambda t\big(\|\boldsymbol{h}_{T^c}\|_{\ell_1} - \langle \epsilon_{T^c}, \boldsymbol{h}_{T^c}\rangle\big).$$

Clearly, when $\boldsymbol{h}_{T^c} \neq 0$, the value at $\boldsymbol{x}^\star + t\boldsymbol{h}$ is strictly greater than that at $\boldsymbol{x}^\star$, which proves the claim. ∎

We return to (8.18) and write

$$\boldsymbol{Y}^{(\ell)^T}(\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}) = \boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}) + \boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)}) + \boldsymbol{Z}^{(\ell)^T}(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}) + \boldsymbol{Z}^{(\ell)^T}(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)}).$$

To establish (8.18), we shall control the ℓ_∞ norm of each term by using Lemma 8.1 and the estimates concerning the size of the residuals. For ease of presentation we assume $d_1 \geq d_\ell$ and $d_\ell \leq n - d_1$—the proof when $d_\ell > d_1$ is similar.

The term $\boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)})$. Using Lemma 8.1 with $a = 2\sqrt{\log N}$, $b = 2\sqrt{2\log N}$, we have

$$\left\|\boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)})\right\|_{\ell_\infty} \leq \sqrt{32}\log N \frac{\mathrm{aff}(S_1, \mathcal{S}_\ell)}{\sqrt{d_1}} \left\|\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2};$$

this holds uniformly over $\ell \neq 1$ with probability at least $1 - \frac{4}{N^2}$. Now applying Lemma 8.5 we conclude that

$$\left\|\boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)})\right\|_{\ell_\infty} \leq \lambda L_N \mathrm{aff}(S_1, S_\ell) := \lambda I_1.$$

The term $\boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)})$. As before,

$$\left\|\boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)})\right\|_{\ell_\infty} \leq \sqrt{32}\log N \frac{1}{\sqrt{n - d_1}} \left\|\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2},$$

which holds uniformly over $\ell \neq 1$ with probability at least $1 - \frac{4}{N^2}$ (we used the fact that the affinity is at most one.) Applying Lemma 8.5 gives

$$\left\|\boldsymbol{X}^{(\ell)^T}(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)})\right\|_{\ell_\infty} \leq L_N \frac{\sigma}{\sqrt{n}} := I_2.$$

The terms $\boldsymbol{Z}^{(\ell)^T}(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)})$ and $\boldsymbol{Z}^{(\ell)^T}(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)})$. Since $\boldsymbol{Z}^{(\ell)}$ is a Gaussian matrix with entries $\mathcal{N}(0, \sigma^2/n)$, applying Lemma A.1 gives

$$\begin{aligned}\left\|\boldsymbol{Z}^{(\ell)^T}(\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)})\right\|_{\ell_\infty} &\leq 2\sigma\sqrt{\frac{2\log N}{n}} \left\|\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \\ &\leq C\lambda\sigma\sqrt{\frac{d_1 \log N}{n}} := \lambda I_3.\end{aligned}$$

with probability at least $1 - \frac{2}{N^2}$

In a similar fashion with probability at least $1 - \frac{2}{N^2}$, we have

$$\left\|\boldsymbol{Z}^{(\ell)^T}(\boldsymbol{y}_\perp - \boldsymbol{Y}_\perp^{(1)}\hat{\boldsymbol{\beta}}^{(1)})\right\|_{\ell_\infty} \leq \sigma^2\sqrt{\frac{L_N}{n}} := I_4.$$

Putting all this together, we need

$$(I_1 + I_3)\lambda + I_2 + I_4 < \lambda$$

to hold with high probability. It is easy to see that the affinity condition in Theorem 3.1 is equivalent to $I_1 + I_3 < 1 - \frac{1}{\sqrt{3}}$. Therefore it suffices to have $\lambda > \sqrt{3}(I_2 + I_4)$. The latter holds if

$$\lambda > L_N \frac{\sigma}{\sqrt{n}}.$$

The calculations have been performed assuming (8.17). Therefore, it suffices to have

$$\lambda > \sqrt{\frac{\sigma}{8d_1}} \max\Big(1, L_N \sqrt{\frac{d_1}{n}}\Big).$$

The simplifying assumption $d_1 \le n/L_N^2$ at the beginning of the paper concludes the proof.

8.3 The size of the solution to the projected problem

This section proves Lemma 8.3. We begin with two definitions.

Definition 8.7 (Inradius) *The inradius $r(\mathcal{P})$ of a convex body $\mathcal{P}$ is the radius of the largest Euclidean ball inscribed in $\mathcal{P}$.*

Definition 8.8 (Restricted isometry property (RIP)) *We say that $\boldsymbol{A} \in \mathbb{R}^{d\times N}$ obeys RIP(s,δ) if*

$$(1-\delta)\,\|\boldsymbol{x}\|_{\ell_2} \le \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} \le (1+\delta)\,\|\boldsymbol{x}\|_{\ell_2}$$

holds for all s-sparse vectors (vectors such that $\|\boldsymbol{x}\|_{\ell_0} \le s$).

We mentioned that Lemma 8.3 is essentially contained in the work of Wojtaszczyk and now make this clear. Below, $\hat{\boldsymbol{x}}$ is any optimal solution to

$$\min \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{8.19}$$

Theorem 8.9 *[45, Theorem 3.4] Suppose $\boldsymbol{A} \in \mathbb{R}^{d\times N}$ obeys RIP(s,δ) and $r(\mathcal{P}(\boldsymbol{A})) \ge \frac{\mu}{\sqrt{s}}$, where $\mathcal{P}(\boldsymbol{A})$ is the symmetrized convex hull of the columns of $\boldsymbol{A}$. Then there is a universal constant $C = C(\delta,\mu)$, such that for any solution $\boldsymbol{x}$ to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, we have*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_{\ell_2} \le C\left\|\boldsymbol{x} - \boldsymbol{x}_{(s)}\right\|_{\ell_2} + C\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_{(s)}\right\|_{\ell_2}. \tag{8.20}$$

Above, $\boldsymbol{x}_{(s)}$ is the best s-sparse approximation to $\boldsymbol{x}$ (the vector $\boldsymbol{x}$ with all but the s-largest entries set to zero).

We now prove Lemma 8.3 and begin with $\lambda = 0$. The expected squared Euclidean norm of a column of $\boldsymbol{Y}_{\|}^{(1)}$ is equal to $1 + \sigma^2 d/n$. Rescaling $\boldsymbol{Y}_{\|}^{(1)}$ as $\boldsymbol{A} = (1+\sigma^2 d/n)^{-1/2}\boldsymbol{Y}_{\|}^{(1)}$, it is a simple calculation to show that with $s = \sqrt{d_1/L_{N_1/d_1}}$ (recall that L_{N_1/d_1} is a constant times $\log(N_1/d_1)$), $\boldsymbol{A}$ obeys RIP(s,δ) for a fixed numerical constant δ. For the same value of s, a simple rescaling of Lemma B.3 asserts that $r(\mathcal{P}(\boldsymbol{A})) \ge \mu/\sqrt{s}$ for a fixed numerical constant μ.

Now let $\boldsymbol{A}^\dagger$ be the pseudo-inverse of $\boldsymbol{A}$, and set $\boldsymbol{\beta} = \boldsymbol{A}^\dagger \boldsymbol{y}_\parallel$. First, $\|\boldsymbol{y}_\parallel\|_{\ell_2} \in [3/4, 5/4]$ and second $\|\boldsymbol{\beta}\|_{\ell_2} \le 1$ as shown in Lemma 8.2. Thus,

$$\left\|\tilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}\right\|_{\ell_2} \le C\,\|\boldsymbol{\beta}\|_{\ell_2} + C\left\|\boldsymbol{y}_\parallel - \boldsymbol{A}\boldsymbol{\beta}_{(s)}\right\|_{\ell_2} \le \frac{9}{4}C + C(1+\delta).$$

The second inequality comes from the RIP property $\|\boldsymbol{A}\boldsymbol{\beta}_{(s)}\|_{\ell_2} \le (1+\delta)\|\boldsymbol{\beta}_{(s)}\|_{\ell_2} \le (1+\delta)$. This completes the proof for $\lambda = 0$. For $\lambda > 0$, one simply applies the same argument with $\boldsymbol{y}_\parallel$ replaced by $\boldsymbol{Y}_\parallel^{(1)}\tilde{\boldsymbol{\beta}}^{(1)}$.

8.4 Proof of Theorem 3.2

Once we know there are no false discoveries, the many many-true-discovery result becomes quite intuitive. The reason is this: we already know that the residual sum of squares $\|\boldsymbol{y}_\parallel - \boldsymbol{Y}_\parallel^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_2}^2$ is much smaller than one provided λ is not too large—this is why we have the upper bound $f(t) \le \alpha_0/t$. Now recall that $\boldsymbol{y}_\parallel$ is a generic point in a d-dimensional subspace and, therefore, it cannot be well approximated as a short linear combination of points taken from the same subspace. It is possible—and not difficult—to make this mathematically precise and thus prove Theorem 3.2. Here, we take a shorter route re-using much of what we have already seen and/or established.

We rescale $\boldsymbol{Y}_\parallel^{(1)}$ as $\boldsymbol{A} = (1+\sigma^2 d/n)^{-1/2}\boldsymbol{Y}_\parallel^{(1)}$ to ensure that the expected squared Euclidean norm of each column is one, and set $s = \sqrt{d_1/L_{N_1/d_1}}$. We introduce three events E_1, E_2 and E_3.

- E_1: there are no false discoveries.
- E_2: with s as above, the two conditions in Theorem 8.9 hold.
- $E_3 = \{\|\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_1} \ge c_1\sqrt{d_1/\log\rho_1}\}$ for some numerical constant c_1.

Since $f(t) \ge \sigma/(\sqrt{2}t)$, there are no false discoveries with high probability. Further, in the proof of this theorem we established that E_1 and E_2 hold with high probability. Last, E_3 also has large probability.

Lemma 8.10 *Let $f(t)$ be as in Theorem 3.2, then the event E_3 has probability at least $1 - e^{-\gamma_3 d_1} - \frac{10}{N}$.*

Proof The proof is nearly the same as that of the lower bound in Lemma 8.2.[7] By definition, the point $\hat{\boldsymbol{\beta}}^{(1)}$ is a solution to

$$\min \|\boldsymbol{\beta}\|_{\ell_1} \quad \text{subject to} \quad \left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\boldsymbol{\beta}\right\|_{\ell_2} \le \tau$$

with $\tau = \|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_2}$. Hence, if we can take τ to be sufficiently smaller than one, then the argument in the proof of Lemma 8.2 can be copied to establish the claim.

Moving forward, Lemma 8.5 gives

$$\begin{aligned}\left\|\boldsymbol{y} - \boldsymbol{Y}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} &\le \frac{32}{3}\lambda\sqrt{\frac{d_1}{\log\rho_1}} + C\sigma \\ &\le \frac{32}{3}\frac{\alpha_0}{\text{Val(Step 1)}}\sqrt{\frac{d_1}{\log\rho_1}} + C\sigma \\ &\le 110\alpha_0 + C\sigma.\end{aligned}$$

[7] Notice that the properties needed for this lemma to hold are the same as that of Lemma 8.2 and, therefore, E_3 holds under the conditions of the no-false discovery theorem. Hence, the guaranteed probabilities of success of our two main theorems are the same.

The second inequality follows from the definition of the regularization parameter since $\lambda \le \alpha_0/\mathrm{Val}(\text{Step 1})$ and the third from (8.3) in Lemma 8.2. We have proved the claim provided α_0 and σ are sufficiently small (this is the place where the assumption about σ comes into play). ∎

Lemma 8.11 *On the event $E_2 \cap \{\|\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_0} \le s\}$, where s is as above,*

$$\left\|\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \le c_2$$

for some numerical constant c_2.

Proof We apply Theorem 8.9 one more time with $\boldsymbol{A}$ as before, $\boldsymbol{y} = (1+\sigma^2 d/n)^{-1/2}\boldsymbol{y}_{\|}$, $\hat{\boldsymbol{x}} = \bar{\boldsymbol{\beta}}^{(1)}$ and $\boldsymbol{x} = \hat{\boldsymbol{\beta}}^{(1)}$. By assumption, $\boldsymbol{x}_{(s)} = \boldsymbol{x}$, and

$$\left\|\bar{\boldsymbol{\beta}}^{(1)} - \hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \le C(1+\sigma^2 d/n)^{-1/2}\left\|\boldsymbol{y}_{\|} - \boldsymbol{Y}_{\|}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2} \le C$$

(the second inequality comes from Lemma 8.5). The proof follows by applying the triangular inequality and Lemma 8.3 with $\lambda = 0$. ∎

The proof of Theorem 3.2 is now straightforward. With s as above, there is nothing to do if $\|\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_0} \ge s$, so we assume $\|\hat{\boldsymbol{\beta}}\|_{\ell_0} \le s$. On this event and $E_1 \cap E_2 \cap E_3$,

$$\frac{\left\|\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_1}}{\left\|\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_2}} \ge c_3\sqrt{\frac{d_1}{\log\rho_1}}.$$

Cauchy-Schwartz asserts that $\|\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_1} \le \sqrt{\|\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_0}}\|\hat{\boldsymbol{\beta}}^{(1)}\|_{\ell_2}$ and, therefore,

$$\left\|\hat{\boldsymbol{\beta}}^{(1)}\right\|_{\ell_0} \ge C d_1/\log\rho_1.$$

A Standard inequalities in probability

This section collects standard inequalities that shall be used throughout. The first concerns tails of chi-square random variables: a chi-square χ^2_n with n degrees of freedom obeys

$$\mathbb{P}\{\chi^2_n \ge (1+\epsilon)n\} \le \exp\Big(-\frac{(1-\log 2)}{2}\,n\epsilon^2\Big). \tag{A.1}$$

The second concerns the size of the dot product between a fixed vector and Gaussian random verctors.

Lemma A.1 *Suppose $\boldsymbol{A}$ in $\mathbb{R}^{d\times N}$ has iid $\mathcal{N}(0,1)$ entries and let $\boldsymbol{z} \in \mathbb{R}^d$ a unit-norm vector. Then*

$$\left\|\boldsymbol{A}^T\boldsymbol{z}\right\|_{\ell_\infty} \le 2\sqrt{2\log N}$$

with probability at least $1 - \frac{2}{N^2}$. (This also applies if $\boldsymbol{z}$ is a random vector independent from $\boldsymbol{A}$.)

Lemma A.2 (Sub-Gaussian rows [42]) *Let $\boldsymbol{A}$ be an $N \times d$ matrix ($N \geq d$) whose rows are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^d$. Then for every $t \geq 0$,*

$$\sigma_{\min}(\boldsymbol{A}) \geq \sqrt{N} - C\sqrt{d} - t$$

with probability at least $1 - e^{-ct^2}$. Here, $\sigma_{\min}$ is the minimum singular value of $\boldsymbol{A}$ and $C = C_K$, $c = c_K > 0$ depend only on the sub-Gaussian norm $K = \max_i \|A_i\|_{\Psi_2}$ of the rows (see [42]).

Lemma A.3 *With probability at least $1 - e^{-d_1/2}$,*

$$\sigma_{\min}(\boldsymbol{Y}_{\|}^{(1)}) \geq \sqrt{\left(1 + \sigma^2 \frac{d_1}{n}\right)}\left(\sqrt{\frac{N_1}{d_1}} - 2\right).$$

Proof This is a trivial consequence of Lemma A.2 above with $t = \sqrt{d_1}$. ∎

Lemma A.4 *If σ and d_1 are as in Section 1.2, all the columns in $\boldsymbol{Y}_{\|}^{(1)}$ and $\boldsymbol{y}_{\|}$ have Euclidean norms in $[3/4, 5/4]$ with probability at least $1 - \frac{1}{N^2}$. (For a single column, the probability is at least equal to $1 - \frac{1}{N^3}$.)*

Proof A column of $\boldsymbol{Y}_{\|}^{(1)}$ or $\boldsymbol{y}_{\|}$ is of the form $\boldsymbol{a} = \boldsymbol{x} + \boldsymbol{z}_{\|}$ where $\boldsymbol{x}$ is uniform on the unit sphere of S_1 and $\boldsymbol{z} \sim \mathcal{N}(0, (\sigma^2/n)\boldsymbol{I}_n)$. We have

$$\|\boldsymbol{x}\|_{\ell_2} - \left\|\boldsymbol{z}_{\|}\right\|_{\ell_2} \leq \|\boldsymbol{a}\|_{\ell_2} \leq \|\boldsymbol{x}\|_{\ell_2} + \left\|\boldsymbol{z}_{\|}\right\|_{\ell_2}.$$

The result follows from $\|\boldsymbol{x}\|_{\ell_2} = 1$ and $\|\boldsymbol{z}_{\|}\|_{\ell_2} \leq \frac{1}{4}$, which holds with high probability. The latter is a consequence of (A.1) since $\|\boldsymbol{z}_{\|}\|_{\ell_2}^2$ (properly normalized) is a chi-square with d_1 degrees of freedom and the bounds on σ and d_1 from (1.2). ∎

B Geometric Lemmas

Consider the linear program

$$\boldsymbol{x}^\star \in \arg\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{B.1}$$

and its dual

$$\boldsymbol{\nu}^\star \in \arg\max_{\boldsymbol{\nu}} \ \langle \boldsymbol{y}, \boldsymbol{\nu} \rangle \quad \text{subject to} \quad \left\|\boldsymbol{A}^T\boldsymbol{\nu}\right\|_{\ell_\infty} \leq 1. \tag{B.2}$$

Lemma B.1 *Any dual feasible point obeys*

$$\|\boldsymbol{\nu}\|_{\ell_2} \leq \frac{1}{r(\mathcal{P}(\boldsymbol{A}))}.$$

Proof Put $r = r(\mathcal{P}(\boldsymbol{A}))$ for short. By definition, there exists $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_{\ell_1} \le 1$ such that $\boldsymbol{A}\boldsymbol{x} = r\boldsymbol{\nu}$. Now,

$$r\|\boldsymbol{\nu}\|_{\ell_2} = \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{\nu}\rangle = \langle \boldsymbol{x}, \boldsymbol{A}^T\boldsymbol{\nu}\rangle \le \|\boldsymbol{x}\|_{\ell_1}\, \|\boldsymbol{A}^T\boldsymbol{\nu}\|_{\ell_\infty} \le 1.$$

■

Strong duality $\|\boldsymbol{x}^\star\|_{\ell_1} = \langle \boldsymbol{y}, \boldsymbol{\nu}^\star\rangle \le \|\boldsymbol{y}\|_{\ell_2}\|\boldsymbol{\nu}^\star\|_{\ell_2}$ also gives:

Lemma B.2 *Any optimal solution $\boldsymbol{x}^\star$ to (B.1) obeys*

$$\|\boldsymbol{x}^\star\|_{\ell_1} \le \frac{\|\boldsymbol{y}\|_{\ell_2}}{r(\mathcal{P}(\boldsymbol{A}))}.$$

Lemma B.3 *Assume $\rho_1 = N_1/d_1 \ge \rho^\star$. Then*

$$r(\mathcal{P}(\boldsymbol{Y}_\parallel^{(1)})) \ge \frac{3}{16}\sqrt{\frac{\log(N_1/d_1)}{d_1}},$$

with probability at least $1 - \frac{1}{N^2} - e^{-\sqrt{N_1 d_1}}$.

Proof Suppose $\boldsymbol{A} \in \mathbb{R}^{d\times N}$ has columns chosen uniformly at random from the unit sphere of $\mathbb{R}^d$ with $\rho = N/d \ge \rho_0$. Then [37, Lemma 7.4]

$$\mathbb{P}\left\{ r(\mathcal{P}(\boldsymbol{A})) < \frac{1}{4}\sqrt{\frac{\log(N/d)}{d}} \right\} \le e^{-\sqrt{Nd}}.$$

The claim in the lemma follows from the lower bound on the Euclidean norm of the columns of $\boldsymbol{Y}_\parallel^{(1)}$ (Lemma A.4) together with the fact that they have uniform orientation. ■

Corollary B.4 *With high probability as above, any dual feasible point to (8.12) obeys*

$$\|\boldsymbol{\nu}\|_{\ell_2}^2 \le \frac{256}{9}\frac{d_1}{\log(N_1/d_1)}.$$

C Sharpening Lemma 8.3 Asymptotically

Here, we assume that the ratio $\rho_1 = N_1/d_1$ is fixed and $N_1 \to \infty$. In this asympotic setting, it is possible to sharpen Lemma 8.3. Our arguments are less formal than in the rest of the paper.

Let $\boldsymbol{x}_0 \in \mathbb{R}^N$ be an unknown vector, and imagine we observe

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{z},$$

where $\boldsymbol{A}$ is a $d \times N$ matrix with i.i.d. $\mathcal{N}(0, 1/d)$ entries, and $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$. Let $\hat{\boldsymbol{x}}$ be the solution to

$$\min \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 + \lambda\|\boldsymbol{x}\|_{\ell_1}.$$

Then setting $\delta = d/N$, the main result in [6, 7] states that almost surely,

$$\lim_{N\to\infty} \frac{1}{N}\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_{\ell_2}^2 = \mathbb{E}\left\{\left[\eta(X_0 + \tau_* Z; \alpha\tau_*) - X_0\right]^2\right\} = \delta(\tau_*^2 - \sigma^2),$$

where $Z \sim \mathcal{N}(0,1)$ and the random variable X_0 has the empirical distribution of the entries of $\boldsymbol{x}_0$. In addition, Z and X_0 are independent. We refer to [6,7] for a precise statement. Above, α and τ_* are solutions to

$$\lambda = \alpha\tau_*\left[1 - \frac{1}{\delta}\mathbb{E}\{\eta'(X_0 + \tau_* Z; \alpha\tau_*)\}\right] \tag{C.1}$$

$$\tau_*^2 = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}[\eta(X_0 + \tau_* Z; \alpha\tau_*) - X_0]^2. \tag{C.2}$$

Here, $\eta(\boldsymbol{x}, \theta)$ is applying a soft-thresholding rule elementwise. For a scalar t, this rule is of the form

$$\eta(t, \theta) = \operatorname{sgn}(t)\max(|t| - \theta, 0).$$

We apply this in the setting of Lemma 8.3 with $\boldsymbol{x}_0 = \boldsymbol{0}$, $X_0 = 0$. Here, $\boldsymbol{A} = \boldsymbol{U}_1^T \boldsymbol{Y}_{\|}^{(1)}$ and with abuse of notation $\boldsymbol{y} := \boldsymbol{U}_1^T \boldsymbol{y}_{\|}$. In the asymptotic regime the vector $\boldsymbol{y}$ and the columns of $\boldsymbol{A}$ are both random Gaussian vectors with variance of each entry equal to $1/d_1 + 1/n$. Since the LASSO solution is invariant by rescaling of the columns and we are interested in bounding its norm, we assume without loss of generality that $\boldsymbol{y}$ and $\boldsymbol{A}$ have $\mathcal{N}(0, 1/d)$ entries, i.e. the variance of the noise $\boldsymbol{z}$ above is $1/d$. With this, the above result simplifies to

$$\lim_{N\to\infty} \frac{1}{N}\,\|\hat{\boldsymbol{x}}\|_{\ell_2}^2 = \mathbb{E}\big\{\big[\eta(\tau_* Z; \alpha\tau_*)\big]^2\big\} = \delta(\tau_*^2 - \sigma^2), \quad \sigma^2 = 1/d.$$

To find α and τ_*, we solve

$$\lambda = \alpha\tau_*\left[1 - \frac{1}{\delta}\mathbb{E}\{\eta'(\tau_* Z; \alpha\tau_*)\}\right], \quad \tau_*^2 = \sigma^2 + \frac{1}{\delta}\,\mathbb{E}[\eta(\tau_* Z; \alpha\tau_*)]^2.$$

Now notice that

$$\mathbb{E}\{\eta'(\tau_* Z; \alpha\tau_*)\} = 2\mathbb{P}\{Z \geq \alpha\}, \quad \mathbb{E}[\eta(\tau_* Z; \alpha\tau_*)]^2 = \tau_*^2 \mathbb{E}[\eta(Z; \alpha)]^2.$$

The equations then become

$$\lambda = \alpha\tau_*\left[1 - \frac{2}{\delta}\mathbb{P}\{Z \geq \alpha\}\right], \quad \tau_*^2 = \frac{\sigma^2}{1 - \frac{1}{\delta}\,\mathbb{E}[\eta(Z; \alpha)]^2}.$$

Eliminating τ_* and solving for α yields

$$\lambda\sqrt{(1 - \frac{1}{\delta}\,\mathbb{E}[\eta(Z; \alpha)]^2)} = \alpha\sigma\left[1 - \frac{2}{\delta}\mathbb{P}\{Z \geq \alpha\}\right].$$

This one-dimensional nonlinear equation can be solved with high accuracy. Plugging in the solution in the expression for τ_* bounds the ℓ_2 norm of the solution.

Now we explain how these relationships can be used to show $\|\hat{\boldsymbol{x}}\|_{\ell_2} \leq 1$ for $\rho \geq \rho^\star$ as $\lambda \to 0$. The argument for any $\lambda > 0$ follows along similar steps, which we avoid here. As λ tends to zero we must have

$$0 = 1 - \frac{2}{\delta}\mathbb{P}\{Z \geq \alpha\} \Rightarrow \mathbb{P}\{Z \geq \alpha\} = \frac{\delta}{2} \Rightarrow \alpha = \sqrt{2}\operatorname{erfc}^{-1}(\delta)$$

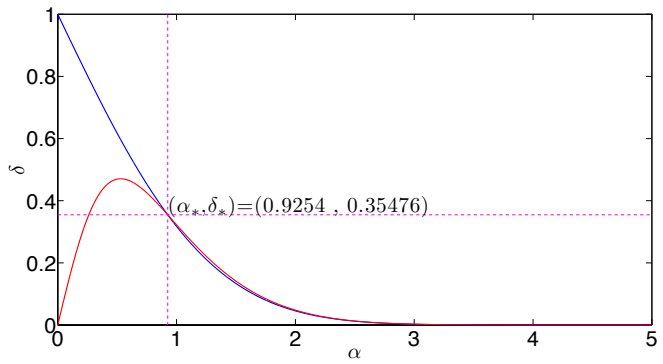


Figure 12: Left-hand side (blue) and right-hand side (red) of (C.3). The two curves intersect at $(\alpha_*, \delta_*) = (0.9254, 0.35476)$.

where erfc^{-1} is the inverse of $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$. With this, we obtain

$$\|\hat{\boldsymbol{x}}\|_{\ell_2}^2 = N\delta(\tau_*^2 - \sigma^2) = N\delta\sigma^2 \frac{\mathbb{E}[\eta(Z;\alpha)]^2}{\delta - \mathbb{E}[\eta(Z;\alpha)]^2} = \frac{\mathbb{E}[\eta(Z;\alpha)]^2}{\delta - \mathbb{E}[\eta(Z;\alpha)]^2}.$$

Some algebraic manipulations give

$$\mathbb{E}[\eta(Z;\alpha)]^2 = (\alpha^2+1)\mathrm{erfc}(\alpha/\sqrt{2}) - \alpha\sqrt{\frac{2}{\pi}}e^{-\alpha^2/2} = (\alpha^2+1)\delta - \alpha\sqrt{\frac{2}{\pi}}e^{-\alpha^2/2},$$

where $\alpha = \sqrt{2}\mathrm{erfc}^{-1}(\delta)$. For the bound to be less than 1 it suffices to have $\mathbb{E}[\eta(Z;\alpha)]^2 \le \delta/2$. After simplification, this is equivalent to

$$\delta = \mathrm{erfc}(\alpha/\sqrt{2}) \le \sqrt{\frac{2}{\pi}} \frac{\alpha}{\alpha^2 + 1/2} e^{-\alpha^2/2}. \tag{C.3}$$

The two functions on both sides of the above inequality are shown in Figure 12. As can be seen, for $\delta \le 0.35476$ we have the desired inequality. This is equivalent to $N_1/d_1 = \rho_1 \ge \rho^\star = 2.8188$.

Acknowledgements

E. C. is partially supported by AFOSR under grant FA9550-09-1-0643 and by ONR under grant N00014-09-1-0258 and by a gift from the Broadcom Foundation. M.S. is supported by a Benchmark Stanford Graduate Fellowship. We thank René Vidal for helpful discussions as well as Ery Arias-Castro, Rina Foygel and Lester Mackey for a careful reading of the manuscript and insightful comments. E. C. would like to thank Chiara Sabatti for invaluable feedback on an earlier version of the paper. He also thanks the organizers of the 41st annual Meeting of Dutch Statisticians and Probabilists held in November 2012 where these results were presented. A brief summary of this work was submitted in August 2012 and presented at the NIPS workshop on Deep Learning in December 2012.

References

[1] P.K. Agarwal and N.H. Mustafa. k-means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 155–165, 2004.

[2] E. Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *Information Theory, IEEE Transactions on*, 57(3):1692–1706, 2011.

[3] E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electronic Journal of Statistics*, 5:1537–1587, 2011.

[4] L. Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 2011.

[5] M.F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1068–1077. Society for Industrial and Applied Mathematics, 2009.

[6] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[7] M. Bayati and A. Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997 –2017, April 2012.

[8] S.R. Becker, E.J. Candès, and M.C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, pages 1–54, 2011.

[9] T.E. Boult and L. Gottesfeld Brown. Factorization-based segmentation of motions. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 179–186, 1991.

[10] P.S. Bradley and O.L. Mangasarian. k-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

[11] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[12] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Transactions on Information Theory*, (99):1–1, 2010.

[13] E. J.. Candès and T. Tao. The Dantzig Selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[14] G. Chen and G. Lerman. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Foundations of Computational Mathematics*, 9(5):517–558, 2009.

[15] G. Chen and G. Lerman. Spectral Curvature Clustering (SCC). *International Journal of Computer Vision*, 81(3):317–330, 2009.

[16] Y. Chen, N.M. Nasrabadi, and T.D. Tran. Hyperspectral image classification using dictionary-based sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, (99):1–13, 2011.

[17] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pages 2790–2797.

[18] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence, arXiv preprint arXiv:1203.1005*, 2012.

[19] B. Eriksson, L. Balzano, and R. Nowak. High-rank matrix completion and subspace clustering with missing data. *Arxiv preprint arXiv:1112.5629*, 2011.

[20] A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[21] Y.P.C. Kotropoulos and G.R. Arce. ℓ_1-graph based music structure analysis. In *International Society for Music Information Retrieval Conference, ISMIR 2011.*

[22] G. Lerman and T. Zhang. Robust recovery of multiple subspaces by geometric ℓ_p minimization. *The Annals of Statistics*, 39(5):2686–2715, 2011.

[23] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, Jan. 2013.

[24] P.L. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[25] L. Lu and R. Vidal. Combined central and subspace clustering for computer vision applications. In *Proceedings of the 23rd international conference on Machine learning*, pages 593–600. ACM, 2006.

[26] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, 2007.

[27] Y. Ma and R. Vidal. Identification of deterministic switched arx systems via identification of algebraic varieties. *Hybrid Systems: Computation and Control*, pages 449–465, 2005.

[28] Y. Ma, A.Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM review*, 50(3):413–458, 2008.

[29] B. McWilliams and G. Montana. Subspace clustering of high-dimensional data: a predictive approach. *Arxiv preprint arXiv:1203.1065*, 2012.

[30] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[31] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[32] N. Ozay, M. Sznaier, and C. Lagoa. Model (in) validation of switched arx systems with unknown switches and its application to activity monitoring. In *49th IEEE Conference on Decision and Control (CDC)*, pages 7624–7630. IEEE, 2010.

[33] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps. GPCA with denoising: A moments-based convex approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3209–3216. IEEE, 2010.

[34] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

[35] M. Rosenbaum and A.B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[36] M. Rosenbaum and A.B. Tsybakov. Improved matrix uncertainty selector. *arXiv preprint arXiv:1112.4413*, 2011.

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

[38] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[39] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

[40] stats.stanford.edu/~candes/RSC.

[41] users.ece.gatech.edu/~sasif/homotopy.

[42] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Chapter 5 of the book Compressed Sensing, Theory and Applications, ed. Y. Eldar and G. Kutyniok, 2012.

[43] R. Vidal. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2):52–68, 2011.

[44] R. Vidal, Y. Ma, and S. Sastry. Generalized Principal Component Analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

[45] P. Wojtaszczyk. Stability and instance optimality for Gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, 10(1):1–13, 2010.

[46] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. *ECCV 2006*, pages 94–106, 2006.

[47] A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari. Guess who rated this movie: Identifying users through subspace clustering. *Arxiv preprint arXiv:1208.1544*, 2012.

[48] T. Zhang, A. Szlam, and G. Lerman. Median k-flats for hybrid linear modeling with many outliers. In *IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 234–241, 2009.

[49] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, pages 1–24, 2012.

[50] F. Zhou, F. Torre, and J. K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *8th IEEE International Conference on Automatic Face and Gesture Recognition, 2008. FG'08.*, pages 1–7, 2008.