ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)

False discovery and its control in low rank estimation

Armeen Taeb,

California Institute of Technology, Pasadena, USA

Parikshit Shah

Yahoo Research and Wisconsin Institutes for Discovery, Madison, USA

and Venkat Chandrasekaran

California Institute of Technology, Pasadena, USA

[Received October 2018. Final revision April 2020]

Summary. Models specified by low rank matrices are ubiquitous in contemporary applications. In many of these problem domains, the row–column space structure of a low rank matrix carries information about some underlying phenomenon, and it is of interest in inferential settings to evaluate the extent to which the row–column spaces of an estimated low rank matrix signify discoveries about the phenomenon. However, in contrast with variable selection, we lack a formal framework to assess true or false discoveries in low rank estimation; in particular, the key source of difficulty is that the standard notion of a discovery is a discrete notion that is ill suited to the smooth structure underlying low rank matrices. We address this challenge via a *geometric* reformulation of the concept of a discovery, which then enables a natural definition in the low rank case. We describe and analyse a generalization of the stability selection method of Meinshausen and Bühlmann to control for false discoveries in low rank estimation, and we demonstrate its utility compared with previous approaches via numerical experiments.

Keywords: Algebraic geometry; Determinantal varieties; Model selection; Regularization; Stability selection; Testing

1. Introduction

Models that are described by low rank matrices are ubiquitous in many contemporary problem domains. The reason for their widespread use is that low rank matrices offer a flexible approach to specify various types of low dimensional structure in high dimensional data. For example, low rank matrices are used to describe user preferences in collaborative filtering (Goldberg *et al.*, 1992), small collections of end member signatures in hyperspectral imaging (Manolakis, 2003), directions of moving targets in radar measurements (Fa and Lamare, 2011), low order systems in control theory (Liu and Vandenberghe, 2009), coherent imaging systems in optics (Pati and Kailath, 1994) and latent variable models in factor analysis (Shapiro, 1982). In many of these settings, the row–column space structure of a low rank matrix carries information about some underlying phenomenon of interest; for instance, in hyperspectral imaging for mineralogy problems, the column space represents the combined signatures of relevant minerals in a mixture.

Address for correspondence: Armeen Taeb, Department of Electrical Engineering, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA.
E-mail: ataeb@caltech.edu

 1369–7412/20/82000

Similarly, the row–column spaces of matrices that are obtained from radar measurements signify the directions of moving targets. Therefore, in inferential contexts in which low rank matrices are estimated from data, it is of interest to evaluate the extent to which the row–column spaces of the estimated matrices signify true or false *discoveries* about the relevant phenomenon.

In seeking an appropriate framework to assess discoveries in low rank estimation, it is instructive to consider the case of variable selection, which may be viewed conceptually as low rank estimation with diagonal matrices. Stated in terms of subspaces, the set of discoveries in variable selection is naturally represented by a subspace that is spanned by the standard basis vectors corresponding to the subset of variables that are declared significant. The number of true discoveries then corresponds to the dimension of the intersection between this 'discovery subspace' and the 'population subspace' (i.e. the subspace that is spanned by standard basis vectors corresponding to significant variables in the population), and the number of false discoveries is the dimension of the 'discovery subspace' minus the number of true discoveries. Generalizing this perspective to low rank estimation, it is perhaps appealing to declare that the number of true discoveries is the dimension of the intersection of the estimated row–column spaces and the population row–column spaces, and the number of false discoveries is the dimension of the remaining components of the estimated row–column spaces. The difficulty with this approach is that we cannot expect any inference procedure to estimate perfectly with positive probability even a one-dimensional subspace of the population row–column spaces as the collection of these spaces is not discrete; in particular, the set of all subspaces of a given dimension is the Grassmannian manifold, whose underlying smooth structure is unlike that of the finite collection of co-ordinate subspaces that correspond to discoveries in variable selection. Therefore, the number of true discoveries would generically be 0. One method to improve on this idea is to define the number of true discoveries as the dimension of the largest subspaces of the estimated row–column spaces that are within a specified angle of the population row–column spaces, and to treat the dimension of the remaining components of the estimated row–column spaces as the number of false discoveries. An unappealing feature of this second approach is that it depends on an extrinsic parameter, and minor perturbations of this parameter could result in potentially large changes in the number of true or false discoveries. In some sense, these preceding attempts fail as they are based on a sharp binary choice that declares components of the estimated row–column spaces exclusively as true or false discoveries, which is ill suited to the smooth structure underlying low rank matrices.

As our first contribution, we develop in Section 2 a *geometric* framework for evaluating false discoveries in low rank estimation. We begin by expressing the number of true or false discoveries in variable selection in terms of functionals of the projection matrices that are associated with the discovery or population subspaces that were described above; this expression varies smoothly with respect to the underlying subspaces, unlike dimensions of intersections of subspaces. Next, we interpret the discovery or population subspaces in variable selection as tangent spaces to algebraic varieties of sparse vectors. Finally, we note that tangent spaces with respect to varieties of low rank matrices encode the row–column space structure of a matrix and therefore offer an appropriate generalization of the subspaces that is discussed in the context of variable selection. Putting these observations together, we substitute tangent spaces with respect to varieties of low rank matrices into our reformulation of discoveries in variable selection in terms of projection matrices, which leads to a natural formalism of the number of true or false discoveries that is suitable for low rank estimation. We emphasize that, although our definition respects the smooth geometric structure underlying low rank matrices, one of its appealing properties is that it specializes transparently to the usual discrete notion of true or false discoveries in the setting of variable selection if the underlying low rank matrices are diagonal.

Our next contribution concerns the development of a procedure for low rank estimation that provides false discovery control. In Section 3, we generalize the 'stability selection' procedure of Meinshausen and Bühlmann (2010) for controlling false discoveries in variable selection. Their method operates by employing variable selection methods in conjunction with subsampling; in particular, one applies a variable selection algorithm to subsamples of a data set and then declares as discoveries those variables that are selected most frequently. In analogy with their approach, our algorithm—which we call 'subspace stability selection'—operates by combining existing low rank estimation methods in conjunction with subsampling. Our framework employs row–column space selection procedures (based on standard low rank estimation algorithms) on subsamples of a data set and then outputs as discoveries a set of row–column spaces that are 'close to' most of the estimated row–column spaces; the specific notion of distance here is based on our tangent space formalism. Building on the results in Meinshausen and Bühlmann (2010) and Shah and Samworth (2013), we provide a theoretical analysis of the performance of our algorithm. A key quantity in our results is the commutator between projection matrices associated with estimated tangent spaces and with the population tangent space, which highlights the distinction between the discrete nature of variable selection and the smooth geometry underlying low rank estimation.

Finally, in Section 4 we contrast subspace stability selection with previous methods in a range of low rank estimation problems involving simulated as well as real data. The tasks involving real data are on estimating user preference matrices for recommender systems and identifying signatures of relevant minerals in hyperspectral images. The estimates that are provided by subspace stability selection offer improvements in multiple respects. First, the row–column spaces of the subspace stability selection estimates are far closer to their population counterparts in comparison with other standard approaches; in other words, our experiments demonstrate that subspace stability selection provides estimates with far fewer false discoveries, without a significant loss in power (both false discovery and power are based on the definitions that are introduced in this paper). Second, in settings in which regularized formulations are employed, subspace stability selection estimates are much less sensitive to the specific choice of the regularization parameter. Finally, a common challenge with approaches that are based on cross-validation for low rank estimation is that they overestimate the complexity of a model, i.e. they produce higher rank estimates (indeed, a similar issue arises in variable selection, which was one of the motivations for the development of stability selection in Meinshausen and Bühlmann (2010)). We observe that the estimates that are produced by subspace stability selection have substantially lower rank than those produced by cross-validation, with a similar or improved prediction performance.

The outline of this paper is as follows. In Section 2, we briefly review the relevant concepts from algebraic geometry and then formulate a false discovery framework for low rank estimation. Our subspace stability selection algorithm is described in Section 3, with theoretical support presented in Section 3.1. In Section 4, we demonstrate the utility of our approach in experiments with synthetic and real data. We conclude with a discussion of further research directions in Section 5.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`.

1.1. Related work

We are aware of prior work for low rank estimation based on testing the level of significance of the singular values of an observed matrix (see, for example, Choi *et al.* (2017), Liu and Lin (2018) and Song and Shin (2018)). However, in contrast with our framework, these methods

do not directly control deviations of row–column spaces, which carry significant information about various phenomena of interest in applications. Further, these previous approaches have limited applicability as they rely on having observations of all the entries of a matrix; this is not so, for example, in low rank matrix completion problems which arise commonly in many domains. In comparison, our methodology is general purpose and is applicable to a broad range of low rank estimation problems. On the computational front, our algorithm and its analysis are a generalization of some of the ideas in Meinshausen and Bühlmann (2010) and Shah and Samworth (2013). However, the geometry underlying the collection of tangent spaces to low rank matrices leads to some new challenges in our context.

1.2. Notation

For a subspace $\mathbb{V}$, we denote projection onto $\mathbb{V}$ by $\mathcal{P}_{\mathbb{V}}$. Given a self-adjoint linear map $M : \bar{\mathbb{V}} \to \bar{\mathbb{V}}$ on a vector space $\bar{\mathbb{V}}$ and a subspace $\mathbb{V} \subset \bar{\mathbb{V}}$, the minimum singular value of M restricted to $\mathbb{V}$ is given by $\sigma_{\min}(\mathcal{P}_{\mathbb{V}} M \mathcal{P}_{\mathbb{V}}) = \inf_{x \in \mathbb{V} \setminus \{0\}} \|Mx\|_{l_2} / \|x\|_{l_2}$. We denote the Kronecker product between two matrices A and B by $A \otimes B$. Finally, the nuclear norm (the sum of singular values) is denoted by '$\|\cdot\|_*$', and the Frobenius norm is denoted by '$\|\cdot\|_{\mathrm{F}}$'.

2. A geometric false discovery framework

We describe a geometric framework for assessing discoveries in low rank estimation. Our discussion proceeds by first reformulating true or false discoveries in variable selection in geometric terms, which then enables a transparent generalization to the low rank case. We appeal to elementary ideas from algebraic geometry on varieties and tangent spaces (Harris, 1995). We also describe a procedure to obtain an estimate of a low rank matrix given an estimate of a tangent space.

2.1. False discovery in low rank estimation

The performance of a variable selection procedure $\hat{\mathcal{S}} \subset \{1, \ldots, p\}$, which estimates a subset of a collection of p variables as being significant, is evaluated by comparing the number of elements of $\hat{\mathcal{S}}$ that are also in the 'true' subset of significant variables $\mathcal{S}^* \subset \{1, \ldots, p\}$—the number of true discoveries is $|\hat{\mathcal{S}} \cap \mathcal{S}^*|$, whereas the number of false discoveries is $|\hat{\mathcal{S}} \cap \mathcal{S}^{*\mathrm{c}}|$. We give next a geometric perspective on this combinatorial notion. As described in Section 1, we can associate with each subset $\mathcal{S} \subset \{1, \ldots, p\}$ the co-ordinate aligned *subspace* $T(\mathcal{S}) = \{x \in \mathbb{R}^p | \mathrm{supp}(x) \subseteq \mathcal{S}\}$, where $\mathrm{supp}(x)$ denotes the locations of the non-zero entries of x. With this notation, the number of false discoveries in an estimate $\hat{\mathcal{S}}$ is given by

$$\#\text{false-discoveries} = |\hat{\mathcal{S}} \cap \mathcal{S}^{*\mathrm{c}}| = \dim\{T(\hat{\mathcal{S}}) \cap T(\mathcal{S}^*)^{\perp}\} = \mathrm{tr}(\mathcal{P}_{T(\hat{\mathcal{S}})} \mathcal{P}_{T(\mathcal{S}^*)^{\perp}}).$$

Similarly, the number of true discoveries is given by $\mathrm{tr}(\mathcal{P}_{T(\hat{\mathcal{S}})} \mathcal{P}_{T(\mathcal{S}^*)})$. These reformulations in terms of projection operators have no obvious 'discrete' attribute to them. In particular, for any subspaces $\mathcal{W}$ and $\tilde{\mathcal{W}}$, the expression $\mathrm{tr}(\mathcal{P}_{\mathcal{W}} \mathcal{P}_{\tilde{\mathcal{W}}})$ is equal to the sum of the squares of the cosines of the principal angles between $\mathcal{W}$ and $\tilde{\mathcal{W}}$ (Björck and Golub, 1973); as a result, the quantity $\mathrm{tr}(\mathcal{P}_{\mathcal{W}} \mathcal{P}_{\tilde{\mathcal{W}}})$ varies smoothly with respect to perturbations of $\mathcal{W}$ and $\tilde{\mathcal{W}}$. The discrete nature of a discovery is embedded inside the encoding of the subsets $\hat{\mathcal{S}}$ and $\mathcal{S}^*$ by using the subspaces $T(\hat{\mathcal{S}})$ and $T(\mathcal{S}^*)$. Consequently, to make progress towards a suitable definition of true or false discoveries in the low rank case, we require an appropriate encoding of row–column space structure via subspaces in the spirit of the mapping $\mathcal{S} \mapsto T(\mathcal{S})$. Towards this goal, we interpret next the subspace $T(\mathcal{S})$ that is associated with a subset $\mathcal{S} \subset \{1, \ldots, p\}$ as a tangent space to an algebraic variety.

Formally, for any integer $k \in \{1, \ldots, p\}$ let $\mathcal{V}_{\text{sparse}}(k) \subset \mathbb{R}^p$ denote the algebraic variety of elements of $\mathbb{R}^p$ with at most k non-zero entries. Then, for any point in $\mathcal{V}_{\text{sparse}}(k)$ consisting of exactly k non-zero entries at locations given by the subset $\mathcal{S} \subset \{1, \ldots, p\}$ (here $|\mathcal{S}| = k$), the tangent space at that point with respect to $\mathcal{V}_{\text{sparse}}(k)$ is given by $T(\mathcal{S})$. In other words, the tangent space at a smooth point of $\mathcal{V}_{\text{sparse}}(k)$ is completely determined by the locations of the non-zero entries of that point. This geometric perspective extends naturally to the low rank case.

Consider the *determinantal variety* $\mathcal{V}_{\text{low rank}}(r) \subset \mathbb{R}^{p_1 \times p_2}$ of matrices of size $p_1 \times p_2$ with rank at most r (here $r \in \{1, \ldots, \min(p_1, p_2)\}$). Then, for any matrix in $\mathcal{V}_{\text{low rank}}(r)$ with rank equal to r and with row and column spaces given by $\mathcal{R} \subset \mathbb{R}^{p_2}$ and $\mathcal{C} \subset \mathbb{R}^{p_1}$ respectively, the tangent space at that matrix with respect to $\mathcal{V}_{\text{low rank}}(r)$ is given by example 8.14 in Harris (1995):

$$T(\mathcal{C}, \mathcal{R}) \triangleq \{M_{\text{R}} + M_{\text{C}} | M_{\text{R}}, M_{\text{C}} \in \mathbb{R}^{p_1 \times p_2}, \text{row-space}(M_{\text{R}}) \subseteq \mathcal{R}, \text{column-space}(M_{\text{C}}) \subseteq \mathcal{C}\}. \quad (2.1)$$

The dimension of $T(\mathcal{C}, \mathcal{R})$ equals $r(p_1 + p_2) - r^2$ and the dimension of its orthogonal complement $T(\mathcal{C}, \mathcal{R})^{\perp}$ equals $(p_1 - r)(p_2 - r)$. Further, the projection operators onto $T(\mathcal{C}, \mathcal{R})$ and onto $T(\mathcal{C}, \mathcal{R})^{\perp}$ can be expressed in terms of the projection maps onto $\mathcal{C}$ and $\mathcal{R}$ as follows:

$$\begin{aligned} \mathcal{P}_{T(\mathcal{C},\mathcal{R})} &= \mathcal{P}_{\mathcal{C}} \otimes I + I \otimes \mathcal{P}_{\mathcal{R}} - \mathcal{P}_{\mathcal{C}} \otimes \mathcal{P}_{\mathcal{R}}, \\ \mathcal{P}_{T(\mathcal{C},\mathcal{R})^{\perp}} &= (I - \mathcal{P}_{\mathcal{C}}) \otimes (I - \mathcal{P}_{\mathcal{R}}) = \mathcal{P}_{\mathcal{C}^{\perp}} \otimes \mathcal{P}_{\mathcal{R}^{\perp}}. \end{aligned} \quad (2.2)$$

where '$\otimes$' denotes a Kronecker product. Consequently, the action of projection operators $\mathcal{P}_{T(\mathcal{C},\mathcal{R})}$ and $\mathcal{P}_{T(\mathcal{C},\mathcal{R})^{\perp}}$ on a matrix $M \in \mathbb{R}^{p_1 \times p_2}$ yields

$$\mathcal{P}_{T(\mathcal{C},\mathcal{R})}(M) = \mathcal{P}_{\mathcal{C}} M + M \mathcal{P}_{\mathcal{R}} - \mathcal{P}_{\mathcal{C}} M \mathcal{P}_{\mathcal{R}}$$

and

$$\mathcal{P}_{T(\mathcal{C},\mathcal{R})^{\perp}}(M) = \mathcal{P}_{\mathcal{C}^{\perp}} M \mathcal{P}_{\mathcal{R}^{\perp}}.$$

In analogy with the previous case with variable selection, the tangent space at a rank r matrix with respect to $\mathcal{V}_{\text{low rank}}(r)$ encodes—and is in one-to-one correspondence with—the row–column space structure at that point. Indeed, estimating the row–column spaces of a low rank matrix can be viewed equivalently as estimating the tangent space at that matrix with respect to a determinantal variety. With this notion in hand, we give our definition of true or false discoveries in low rank estimation.

Definition 1. Let $\mathcal{C}^* \subset \mathbb{R}^{p_1}$ and $\mathcal{R}^* \subset \mathbb{R}^{p_2}$ denote the column and row spaces of a low rank matrix in $\mathbb{R}^{p_1 \times p_2}$; in particular, $\dim(\mathcal{C}^*) = \dim(\mathcal{R}^*)$. Given observations from a model that is parameterized by this matrix, let $(\hat{\mathcal{C}}, \hat{\mathcal{R}}) \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ be an estimator of the pair of subspaces $(\mathcal{C}^*, \mathcal{R}^*)$ with $\dim(\hat{\mathcal{C}}) = \dim(\hat{\mathcal{R}})$. Then the *expected false discovery* of the estimator is defined as

$$\text{FD} = \mathbb{E}[\text{tr}(\mathcal{P}_{T(\hat{\mathcal{C}}, \hat{\mathcal{R}})} \mathcal{P}_{T(\mathcal{C}^*, \mathcal{R}^*)^{\perp}})], \quad (2.3)$$

and the power of the estimator is defined as

$$\text{PW} = \mathbb{E}[\text{tr}(\mathcal{P}_{T(\hat{\mathcal{C}}, \hat{\mathcal{R}})} \mathcal{P}_{T(\mathcal{C}^*, \mathcal{R}^*)})]. \quad (2.4)$$

The expectations in both cases are with respect to randomness in the data that are employed by the estimator, and the tangent spaces $T(\hat{\mathcal{C}}, \hat{\mathcal{R}})$, $T(\mathcal{C}^*, \mathcal{R}^*)$ are as defined in expression (2.1).

With respect to our objective of identifying a suitable notion of discovery for low rank estimation, the definitions of FD and of PW have some favourable attributes. These definitions do not depend on a choice of basis for the tangent space $T(\mathcal{C}^*, \mathcal{R}^*)$. Further, for the reasons that were described above, small changes in row–column space estimates lead to small changes in the

performance of an estimator, as evaluated by FD and PW. Although these definitions respect the smooth structure underlying low rank matrices, they specialize transparently to the usual discrete notion of true or false discoveries in the setting of variable selection if the underlying low rank matrices are diagonal. We also have that the expected false discovery is bounded as $0 \leqslant \mathrm{FD} \leqslant \dim\{T(\mathcal{C}^*, \mathcal{R}^*)^\perp\}$ and the power is bounded as $0 \leqslant \mathrm{PW} \leqslant \dim\{T(\mathcal{C}^*, \mathcal{R}^*)\}$, which is in agreement with the intuition that the spaces $T(\mathcal{C}^*, \mathcal{R}^*)$ and $T(\mathcal{C}^*, \mathcal{R}^*)^\perp$ represent the total true and false discoveries respectively that can be made by any estimator. Similarly, we observe that $\mathrm{FD} + \mathrm{PW} = \mathbb{E}[\dim\{T(\hat{\mathcal{C}}, \hat{\mathcal{R}})\}]$, which is akin to the expected total discovery made by the estimator $(\hat{\mathcal{C}}, \hat{\mathcal{R}})$.

One can also arrive at the definitions (2.3) and (2.4) in an 'axiomatic' manner as follows. Suppose that we wish to identify a suitable notion of alignment between the estimate $T(\hat{\mathcal{C}}, \hat{\mathcal{R}})$ and the population $T(\mathcal{C}^*, \mathcal{R}^*)$ via a real-valued function $f(\cdot, \cdot)$ whose arguments consist of a pair of tangent spaces. First, f should remain invariant to simultaneous isometric linear transformations of the row–column spaces of the population and of the estimate; as a parallel, the appropriate invariance in variable selection is simultaneous relabelling of the variables in the estimate and the population. We conclude from this that f must be a function purely of the *principal angles* between its arguments, which correspond to the spectrum of the product of the associated projection matrices. Second, our definition of f should satisfy the condition that the sum $f\{T(\hat{\mathcal{C}}, \hat{\mathcal{R}}), T(\mathcal{C}^*, \mathcal{R}^*)^\perp\} + f\{T(\hat{\mathcal{C}}, \hat{\mathcal{R}}), T(\mathcal{C}^*, \mathcal{R}^*)\}$ equals $\dim\{T(\hat{\mathcal{C}}, \hat{\mathcal{R}})\}$—i.e. the sum of the false discovery and the true discovery must equal the total amount of discovery. Based on this requirement as well as the deduction from the first argument, one can arrive at the definitions (2.3) and (2.4) after taking expectations.

We note that the definition of FD may be modified to obtain an analogue of the *false discovery rate* (Benjamini and Hochberg, 1995), which is of interest in contemporary multiple testing as well as in high dimensional estimation:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\mathrm{tr}(\mathcal{P}_{T(\hat{\mathcal{C}}, \hat{\mathcal{R}})} \mathcal{P}_{T(\mathcal{C}^*, \mathcal{R}^*)^\perp})}{\dim\{T(\hat{\mathcal{C}}, \hat{\mathcal{R}})\}}\right].$$

We focus in the present paper on controlling the quantity FD and we discuss in Section 5 some challenges that are associated with controlling FDR in low rank estimation.

Finally, although the main focus of this paper is a false discovery framework for low rank estimation in which we seek reliable estimates of both the row and the column spaces, the geometric perspective outlined here can be adapted to settings in which one seeks only an estimate of the column space of a low rank matrix. (Such a problem arises in hyperspectral imaging, as illustrated in Section 4.) In such situations, the ideas described previously can be extended as follows:

$$\left.\begin{aligned} \widetilde{\mathrm{FD}} &= \mathbb{E}[\mathrm{tr}(\mathcal{P}_{\hat{\mathcal{C}}} \mathcal{P}_{\mathcal{C}^{*\perp}})]; \\ \widetilde{\mathrm{PW}} &= \mathbb{E}[\mathrm{tr}(\mathcal{P}_{\hat{\mathcal{C}}} \mathcal{P}_{\mathcal{C}^*})]; \\ \widetilde{\mathrm{FDR}} &= \mathbb{E}\left[\frac{\mathrm{tr}(\mathcal{P}_{\hat{\mathcal{C}}} \mathcal{P}_{\mathcal{C}^{*\perp}})}{\dim(\hat{\mathcal{C}})}\right]. \end{aligned}\right\} \tag{2.5}$$

Here $\mathcal{C}^* \subset \mathbb{R}^p$ represents the population column space and $\hat{\mathcal{C}} \subset \mathbb{R}^p$ is an estimator. These expressions can be derived by considering tangent spaces with respect to quotients of the determinantal variety under certain equivalence relations; supplementary material section A.9 provides the details.

2.2. From tangent space to parameter estimation

Although the primary emphasis of this paper is on a framework to evaluate and control the expected false discovery of tangent spaces estimated from data, in many practical settings (e.g. some of the prediction tasks with real data sets in Section 4), the ultimate object of interest is an estimate of a low rank matrix. One can obtain such an estimate by solving a subsequent matrix estimation problem in which the tangent space of the matrix is constrained to lie within the tangent space identified from our framework. Concretely, let $T(\mathcal{C}, \mathcal{R}) \subset \mathbb{R}^{p_1 \times p_2}$ be a tangent space that corresponds to column and row spaces $\mathcal{C} \subset \mathbb{R}^{p_1}$ and $\mathcal{R} \subset \mathbb{R}^{p_2}$, and, given a collection of observations $\mathcal{D}$, we wish to solve the following optimization problem:

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2}}{\arg\min}\ \mathrm{Loss}(L; \mathcal{D}) \qquad \text{subject to } T\{\text{column-space}(L), \text{row-space}(L)\} \subseteq T(\mathcal{C}, \mathcal{R}), \tag{2.6}$$

in which the decision variable L is constrained to have a tangent space that lies within the prescribed tangent space $T(\mathcal{C}, \mathcal{R})$. Furthermore, this constraint may be simplified as follows. Suppose that the subspaces $\mathcal{R}$ and $\mathcal{C}$ are of dimension k. Let $U_C \in \mathbb{R}^{p_1 \times k}$ and $U_R \in \mathbb{R}^{p_2 \times k}$ be any matrices with columns spanning the spaces $\mathcal{C}$ and $\mathcal{R}$ respectively. Then we can check that the set $\{U_C M U_R' \mid M \in \mathbb{R}^{k \times k}\}$ is precisely the collection of matrices whose tangent spaces are contained in $T(\mathcal{C}, \mathcal{R})$. Consequently problem (2.6) may be reformulated as

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2},\, M \in \mathbb{R}^{k \times k}}{\arg\min}\ \mathrm{Loss}(L; \mathcal{D}) \qquad \text{subject to } L = U_C M U_R'. \tag{2.7}$$

Note that the constraint here is linear in the decision variables L and M. Consequently, an appealing property of problem (2.7) is that, if the loss function $\mathrm{Loss}(\cdot; \mathcal{D})$ is convex, then problem (2.7) is a convex optimization problem. For example, when $\mathrm{Loss}(\cdot; \mathcal{D})$ is the squared loss, an optimal solution can be obtained in closed form.

In a similar fashion, in situations in which one is only concerned with estimating low rank matrices with an accurate column space, one can solve an analogue of problem (2.7) in which the decision variable satisfies the linear constraint that its column space lies inside a prescribed column space.

3. False discovery control via subspace stability selection

Building on the discussion in the preceding section, our objective is the accurate estimation of the tangent space that is associated with a low rank matrix, as this is in one-to-one correspondence with the row–column spaces of the matrix. In this section, we formulate an approach based on the stability selection procedure of Meinshausen and Bühlmann (2010) to estimate such a tangent space. We shall also describe how this method can be specialized for problems involving subspace estimation.

Stability selection is a general technique to control false discoveries in variable selection. The procedure can be paired with any variable selection procedure as follows: instead of applying a selection procedure (e.g. the lasso) to a collection of observations, we instead apply the procedure to many subsamples of the data and then choose those variables that are most consistently selected in the subsamples. The virtue of the subsampling and averaging framework is that it provides control over the expected number of falsely selected variables (see theorem 1 in Meinshausen and Bühlmann (2010) and theorem 1 in Shah and Samworth (2013)). We develop a generalization of this framework in which existing row–column space selection procedures (based on any low rank estimation procedure) are employed on subsamples of the data,

and then these spaces are aggregated to produce a tangent space that provides false discovery control.

3.1. Subsampling procedure

Although our framework is applicable to general subsamples of the data, we adopt the subsampling method that was outlined in Shah and Samworth (2013) in our experimental demonstrations and our theoretical analysis; in particular, given a data set $\mathcal{D}$ and a positive (even) integer B, we consider B subsamples or bags obtained from $B/2$ *complementary partitions* of $\mathcal{D}$ of the form $\{(\mathcal{D}_{2j-1}, \mathcal{D}_{2j}) : j = 1, 2, 3, \ldots, B/2\}$, where $|\mathcal{D}_{2j-1}| = |\mathcal{D}|/2$ and $\mathcal{D}_{2j} = \mathcal{D} \backslash \mathcal{D}_{2j-1}$.

3.2. Set-up for numerical demonstrations

For our numerical illustrations in this section, we consider the following stylized low rank matrix completion problem. The population parameter $L^* \in \mathbb{R}^{70 \times 70}$ is a rank 10 matrix with singular values (and associated multiplicities) given by 1 (times 3), 0.5 (times 5) and 0.1 (times 2), and with row–column spaces sampled uniformly at random according to the Haar measure. We are given noisy observations $Y_{i,j} = L^*_{i,j} + \epsilon_{i,j}$ with $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ and $(i, j) \in \Omega$, where $\Omega \subset \{1, \ldots, 70\}^2$ is chosen uniformly at random with $|\Omega| = 3186$. The variance σ^2 is chosen to set the signal-to-noise ratio SNR (defined as $\mathbb{E}[\|L^*\|_{\mathrm{F}}/\|\epsilon\|_{\mathrm{F}}]$) at a desired level, and this is specified later. As our subsamples, we consider a collection of $B = 100$ subsets each consisting of $|\Omega|/2 = 1593$ entries obtained from 50 random complementary partitions of the data. On each subsample—corresponding to a subset $S \subset \Omega$ of observations with $|S| = 1593$—we employ the convex program (Srebro and Shraibman, 2005; Candès and Recht, 2009)

$$\hat{L} = \underset{L \in \mathbb{R}^{70 \times 70}}{\arg\min} \sum_{\{i,j\} \in S} \|(L - Y)_{i,j}\|_{\mathrm{F}}^2 + \lambda \|L\|_*, \tag{3.1}$$

and we report the tangent space $T\{\text{column-space}(\hat{L}), \text{row-space}(\hat{L})\}$ as the estimate that is associated with the subsample. Here $\lambda > 0$ is a regularization parameter (to be specified later) and '$\|\cdot\|_*$' is the nuclear norm (the sum of the singular values), which is commonly employed to promote low rank structure in a matrix (Fazel, 2002). We emphasize that our development is relevant to general low rank estimation problems, and this problem is merely for illustration in the present section; for a more comprehensive set of experiments in more general settings, we refer the reader to Section 4.

3.3. Stable tangent spaces

The first step in stability selection is to combine estimates of significant variables that are obtained from different subsamples. This is accomplished by computing for each variable the frequency with which it is selected across the subsamples. We generalize this idea to our context via projection operators onto tangent spaces as follows.

Definition 2 (average projection operator). Suppose that $\hat{T}$ is an estimator of a tangent space of a low rank matrix, and suppose further that we are given a set of observations $\mathcal{D}$ and a corresponding collection of subsamples $\{\mathcal{D}_l\}_{l=1}^B$ with each $\mathcal{D}_l \subset \mathcal{D}$. Then the *average projection operator* of the estimator $\hat{T}$ with respect to the subsamples $\{\mathcal{D}_l\}_{l=1}^B$ is defined as

$$\mathcal{P}_{\text{avg}} \triangleq \frac{1}{B} \sum_{l=1}^{B} \mathcal{P}_{\hat{T}(\mathcal{D}_l)}, \tag{3.2}$$

where $\hat{T}(\mathcal{D}_l)$ is the tangent space estimate that is based on the subsample $\mathcal{D}_l$.

Here $\mathcal{P}_{\text{avg}} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^{p_1 \times p_2}$ is self-adjoint, and its eigenvalues lie in the interval $[0, 1]$ as each $\mathcal{P}_{\hat{T}(\mathcal{D}_l)}$ is self-adjoint with eigenvalues equal to 0 or 1. To draw a comparison with variable selection, the tangent spaces in that case correspond to subspaces that are spanned by coordinate vectors in $\mathbb{R}^p$ (with p being the total number of variables of interest) and the average projection operator is a diagonal matrix of size $p \times p$, with each entry on the diagonal specifying the fraction of subsamples in which a particular variable is selected. The virtue of averaging over tangent spaces estimated across a large number of subsamples is that most of the 'energy' of the average projection operator $\mathcal{P}_{\text{avg}}$ tends to be better aligned with the underlying population tangent space. We illustrate this point next with an example.

3.3.1. Illustration: the value of averaging projection maps

Consider the stylized low rank matrix completion problem that was described in Section 3.2. To support the intuition that the average projection matrix $\mathcal{P}_{\text{avg}}$ has reduced in energy in directions corresponding to $T^{*\perp}$ (i.e. the orthogonal complement of the population tangent space), we compare the quantities $\mathbb{E}[\text{tr}(\mathcal{P}_{\text{avg}}\mathcal{P}_{T^{*\perp}})]$ and $\mathbb{E}[\text{tr}(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^{*\perp}})]$, where the expectation is computed over 100 instances. Generically speaking, the operator $\mathcal{P}_{\text{avg}}$ is not a projection operator onto a tangent space and thus the quantity $\mathbb{E}[\text{tr}(\mathcal{P}_{\text{avg}}\mathcal{P}_{T^{*\perp}})]$ is not a valid false discovery; rather it evaluates the average false discovery over the subsampled models. The second quantity, $\mathbb{E}[\text{tr}(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^{*\perp}})]$, is based on employing the nuclear norm regularization procedure on the full set of observations. The variance σ is selected so that $\text{SNR} = \{0.8, 1.6\}$. As is evident from Fig. 1, $\mathbb{E}[\text{tr}(\mathcal{P}_{\text{avg}}\mathcal{P}_{T^{*\perp}})]$ is smaller than $\mathbb{E}[\text{tr}(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^{*\perp}})]$ for the entire range of λ, with the gap being larger in the low SNR-regime. In other words, averaging the subsampled tangent spaces reduces energy in the directions that are spanned by $T^{*\perp}$.

While the average projection aggregated over many subsamples appears to have less energy in $T^{*\perp}$, this operator is not a proper projection. Thus it still remains for us to identify a single tangent space as our estimate from $\mathcal{P}_{\text{avg}}$. We formulate the following criterion to establish a measure of closeness between a single tangent space and the aggregate over subsamples.

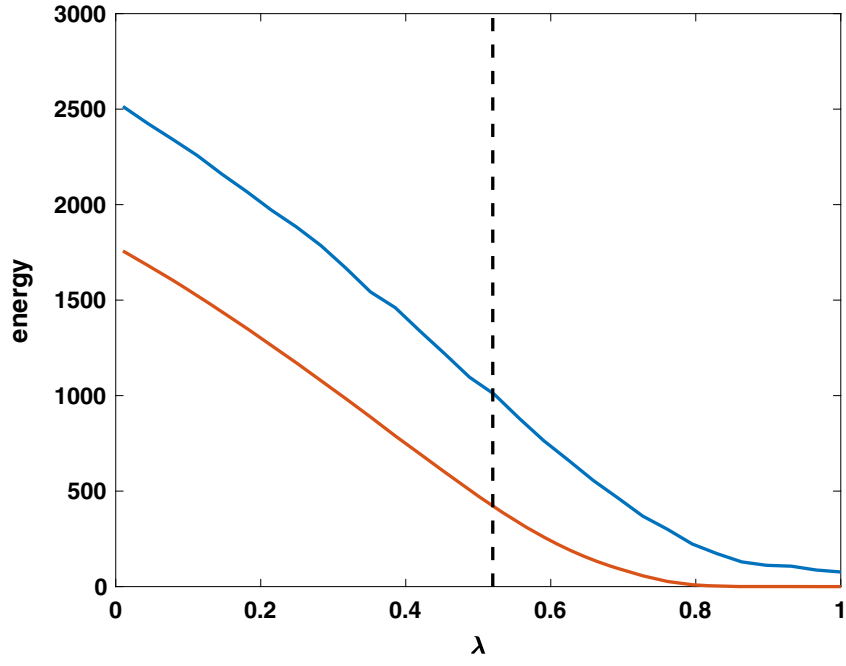
Definition 3 (stable tangent spaces). Suppose that $\hat{T}$ is an estimator of a tangent space of a low rank matrix, and suppose further that we are given a set of observations $\mathcal{D}$ and a corresponding collection of subsamples $\{\mathcal{D}_l\}_{l=1}^{B}$ with each $\mathcal{D}_l \subset \mathcal{D}$. For a parameter $\alpha \in (0, 1)$, the set of stable tangent spaces is defined as

$$\mathcal{T}_\alpha \triangleq \{T | \sigma_{\min}(\mathcal{P}_T \mathcal{P}_{\text{avg}} \mathcal{P}_T) \geqslant \alpha \text{ and } T \text{ is a tangent space to a determinantal variety}\} \tag{3.3}$$

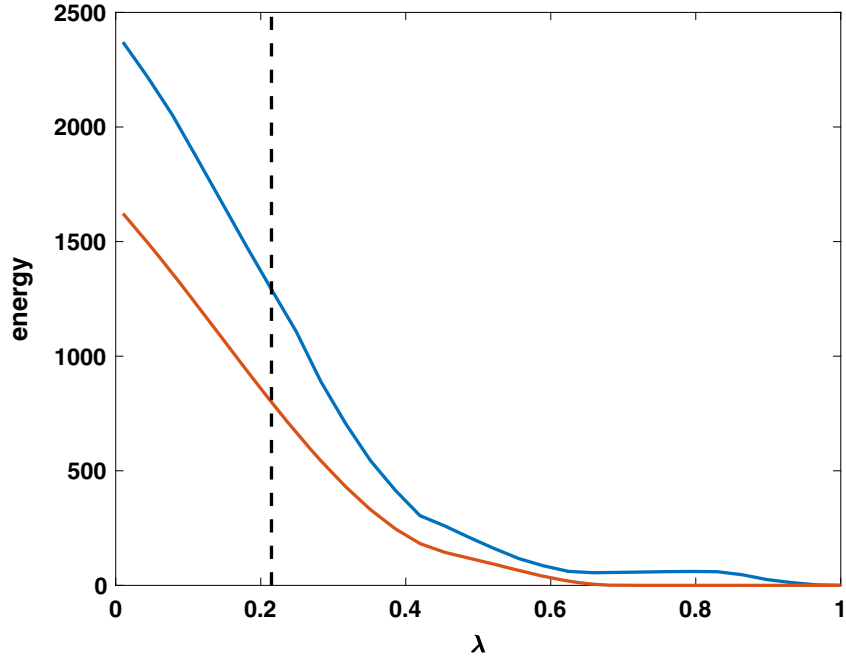
where $\mathcal{P}_{\text{avg}}$ is computed on the basis of definition 2.

As the spectrum of $\mathcal{P}_{\text{avg}}$ lies in the range $[0, 1]$, this is also the only meaningful range of values for α. The set $\mathcal{T}_\alpha$ consists of all those tangent spaces T to a determinantal variety such that the Rayleigh quotient of every non-zero element of T with respect to $\mathcal{P}_{\text{avg}}$ is at least α. To contrast again with variable selection, we note that both $\mathcal{P}_T$ and $\mathcal{P}_{\text{avg}}$ are diagonal matrices in that case (and thus are simultaneously diagonalizable). As a consequence, the set $\mathcal{T}_\alpha$ has a straightforward characterization for variable selection problems; it consists of subspaces that are spanned by any subset of standard basis vectors corresponding to variables that are selected as significant in at least an α-fraction of the subsamples.

As averaging the tangent spaces that are obtained from the subsampled data reduces energy in the directions that are contained in $T^{*\perp}$, each element of $\mathcal{T}_\alpha$ is also far from being closely aligned with $T^{*\perp}$ (for large values of α). We build on this intuition by proving next that a tangent space estimator that selects any element of $\mathcal{T}_\alpha$ provides false discovery control at a level that is a function of α. In Section 3.5 we describe efficient methods to choose an element of $\mathcal{T}_\alpha$.



(a)



(b)

Fig. 1. The quantities $\mathbb{E}[\mathrm{tr}(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^*\perp})]$ (——) (no subsampling) and $\mathbb{E}[\mathrm{tr}(\mathcal{P}_{\mathrm{avg}}\mathcal{P}_{T^*\perp})]$ (——) (with subsampling) as a function of λ for (a) SNR = 0.8 and (b) SNR = 1.6 in the synthetic matrix completion set-up: ¦, cross-validated choice of λ

As a final remark, the ideas that are described here can be readily applied to subspace estimation problems. Specifically, we define the average projection operator $\mathcal{P}^{\mathcal{C}}_{\text{avg}}$ (analogous to expression (3.2)) as the average of projection matrices onto column space estimates that are obtained from $n/2$ subsamples. Then, the stable subspace set (3.3) is modified to be the collection of subspaces $\mathcal{C} \in \mathbb{R}^p$ that satisfy the criterion $\sigma_{\min}(\mathcal{P}_{\mathcal{C}}\mathcal{P}^{\mathcal{C}}_{\text{avg}}\mathcal{P}_{\mathcal{C}}) \geqslant \alpha$.

3.4. False discovery control of stable tangent spaces: theoretical analysis

Consider a low rank matrix $L^* \in \mathbb{R}^{p_1 \times p_2}$ with associated tangent space T^*, and suppose that we are given independent and identically distributed (IID) observations from a model parameterized by L^*. The objective is to obtain an accurate estimate of T^*. We intentionally keep our discussion broad so that our results are relevant for a wide range of low rank estimation problems, e.g. low rank matrix completion or factor analysis. Let $\hat{T}$ denote a tangent space estimator that operates on samples drawn from the model parameterized by L^*. Let $\mathcal{D}(n)$ denote a data set consisting of n IID observations from this model; we assume that n is even and that we are given B subsamples $\{\mathcal{D}_l\}_{l=1}^B$ via complementary partitions of $\mathcal{D}(n)$.

We present a general result that bounds the expected false discovery of stable tangent spaces under the sole assumption that the data set provided consists of IID observations. Under additional assumptions that take the form of 'better than random guessing' and a geometric analogue of exchangeability, we specialize our result to obtain a more refined bound that is similar in spirit to the bound of Meinshausen and Bühlmann (2010). Finally, inspired by theorem 1 of Shah and Samworth (2013), we also specialize our result to produce a bag-independent false discovery bound that is valid for any $B \geqslant 2$. The results in this section extend naturally to settings in which one only seeks accurate estimates of the column space of a matrix; for precise statements in that setting, see supplementary material section A.10.

Our theoretical findings are centred on the following intuition: for subspace stability selection to be effective, the tangent space estimates across subsamples should contain many directions around T^* (i.e. the signal component) and the remaining components (i.e. the noise) should be evenly spread over all the other directions. Owing to the smooth structure underlying low rank matrices, there are 'many' directions in which deviations about T^* can occur in a low rank estimation procedure (a significant contrast with variable selection where the collection of tangent spaces is a discrete set); thus, the requirement on the noise portion of the estimates from the subsamples is a stringent one. This situation is alleviated if the noise components in the subsamples are concentrated around $T^{*\perp}$, i.e. the tangent space estimates across subsamples contain directions that mostly lie close to T^* or $T^{*\perp}$. Mathematically, this intuition can be quantified via *commutators*. The commutator between self-adjoint operators A and B is denoted $[A, B] = AB - BA$, and this map evaluates how far away A and B are from commuting with each other. For projection operators $\mathcal{P}_{T_1}$ and $\mathcal{P}_{T_2}$ associated with subspaces T_1 and T_2, the singular values of $[\mathcal{P}_{T_1}, \mathcal{P}_{T_2}]$ are $\pm\frac{1}{2}\sin(2\theta_i)$ where $\{\theta_i\}$ are the principal angles between T_1 and T_2 (Galántai, 2008). Consequently, $\|[\mathcal{P}_{T_1}, \mathcal{P}_{T_2}]\|_{\text{F}}^2 = \frac{1}{2}\Sigma_i \sin(2\theta_i)^2$ and $\|[\mathcal{P}_{T_1}, \mathcal{P}_{T_2}]\|_2^2 = \frac{1}{4}\max_i \sin(2\theta_i)^2$. A small commutator between the tangent space estimates from subsamples and $T^{*\perp}$ ensures that the tangent space estimates consist of components that are closely aligned with T^* or with $T^{*\perp}$. (As a contrast, in variable selection the associated projection operators commute; in particular, $\theta_i \in \{0, \frac{\pi}{2}\}$ in variable selection.)

Theorem 1 (false discovery control of subspace stability selection). Consider the set-up that was described above. Let $\hat{T}(\mathcal{D}_l)$ denote the tangent space estimates that are obtained from each of the subsamples, and let $\mathcal{P}_{\text{avg}}$ denote the associated average projection operator computed

via expression (3.2) over B complementary bags. Fix any $\alpha \in (\frac{1}{2}, 1)$ and let T denote any selection of an element of the associated set $\mathcal{T}_\alpha$ of stable tangent spaces. Then for any fixed orthonormal basis $\{M_i\}_{i=1}^{\dim(T^{*\perp})}$ for $T^{*\perp}$, we have that

$$\mathbb{E}[\mathrm{tr}(\mathcal{P}_T \mathcal{P}_{T^{*\perp}})] \leqslant F + \kappa_{\mathrm{bag}}(\alpha) + 2(1-\alpha)\mathbb{E}[\dim(T)], \tag{3.4}$$

where for a basis-dependent bound take $F = \Sigma_{i=1}^{\dim(T^{*\perp})} \mathbb{E}[\|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M_i)\|_{\mathrm{F}}]^2$ and

$$\kappa_{\mathrm{bag}}(\alpha) = \sum_{i=1}^{\dim(T^{*\perp})} (2/B) \sum_{j=1}^{B/2} \mathbb{E}[\max_{k\in\{0,1\}} \mathrm{tr}([\mathcal{P}_T, \mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}] \times [\mathcal{P}_{\mathrm{span}(M_i)}, \mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}])],$$

and for a basis-independent bound take $F = \mathbb{E}[\mathrm{tr}(\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}} \mathcal{P}_{T^{*\perp}})^{1/2}]^2$ and

$$\kappa_{\mathrm{bag}}(\alpha) = (2/B) \sum_{j=1}^{B/2} \mathbb{E}[\max_{k\in\{0,1\}} \mathrm{tr}([\mathcal{P}_T, \mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}] \times [\mathcal{P}_{T^{*\perp}}, \mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}])].$$

The expectations are with respect to randomness in the data and the set $\mathcal{D}(n/2)$ denotes $n/2$ IID observations drawn from the model parameterized by L^*.

The proof of theorem 1 is presented in the on-line supplementary material section A.1. The result states that the expected false discovery of a stable tangent space is bounded by a sum of three quantities. The first term F characterizes the quality of the estimator employed on subsamples consisting of $n/2$ observations. The terms $\kappa_{\mathrm{bag}}(\alpha)$ and $2(1-\alpha)\mathbb{E}[\dim(T)]$ are functions of the user-specified parameter α, the number of bags B and the product of commutators. In proposition 1, we show that α close to 1 leads to a small $\kappa_{\mathrm{bag}}(\alpha)$, and thus, as expected, a smaller expected false discovery. Further, one must select $\alpha > \frac{1}{2}$ for bound (3.4) to be non-vacuous as we always have that $\mathbb{E}[\mathrm{tr}(\mathcal{P}_T \mathcal{P}_{T^{*\perp}})] \leqslant \mathbb{E}[\dim(T)]$.

Remark 1. The quantities $\Sigma_{i=1}^{\dim(T^{*\perp})} \mathbb{E}[\|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M_i)\|_{\mathrm{F}}]^2$ and $\mathbb{E}[\mathrm{tr}(\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}} \mathcal{P}_{T^{*\perp}})^{1/2}]^2$ for F highlight the role of bagging in reducing variance. For ease of exposition, we define $\beta \in \mathbb{R}^{\dim(T^{*\perp})}$ as $\beta_i =^{\triangle} \|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M_i)\|_{\mathrm{F}}$, so that $\Sigma_{i=1}^{\dim(T^{*\perp})} \mathbb{E}[\|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M_i)\|_{\mathrm{F}}]^2 = \mathrm{tr}(\mathbb{E}[\beta]\mathbb{E}[\beta]')$ and $\mathrm{tr}(\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}} \mathcal{P}_{T^{*\perp}}) = \mathrm{tr}(\beta\beta')$. Jensen's inequality yields $\mathbb{E}[\mathrm{tr}(\beta\beta')^{1/2}]^2 \leqslant \mathbb{E}[\mathrm{tr}(\beta\beta')]$, so the improvement of bagging over just using a subsample $\mathcal{D}(n/2)$ once is given by $\mathrm{var}\{\mathrm{tr}(\beta\beta')^{1/2}\}$. Next, by appealing to the positive definiteness of a covariance matrix, we have that $\mathrm{tr}(\mathbb{E}[\beta]\mathbb{E}[\beta]') \leqslant \mathbb{E}[\mathrm{tr}(\beta\beta')]$; in this case, the variance reduction is given by $\mathrm{tr}\{\mathrm{cov}(\beta)\}$. In both these cases, the variance is maximally reduced under conditions that follow from the Bhatia–Davis inequality. Specifically, given a fixed $\mathbb{E}[\mathrm{tr}(\beta\beta')^{1/2}]$, the Bhatia–Davis inequality states that $\mathrm{var}\{\mathrm{tr}(\beta\beta')^{1/2}\}$ is enhanced when the distribution of $\mathrm{tr}(\beta\beta')^{1/2}$ concentrates around 0 and $\sqrt{\dim(T^{*\perp})}$ (i.e. most discoveries are either true or false). Similarly, given a fixed $\mathbb{E}[\beta]$, $\mathrm{tr}\{\mathrm{cov}(\beta)\}$ is enhanced when the distribution of each β_i concentrates around 0 or 1 (i.e. the estimate $\hat{T}\{\mathcal{D}(n/2)\}$ is mostly aligned with or orthogonal to each $M_i \in T^{*\perp}$). Such concentration of β_i can be precisely translated to the commutators $\|\mathbb{E}[[\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}, \mathcal{P}_{\mathrm{span}(M_i)}]]\|_{\mathrm{F}}$ being small, which is exploited in proposition 2 to bound F. In Section 4, we use this intuition to describe synthetic experiments that illustrate the improvement (in terms of expected false discovery) of a stable tangent space over using the original estimator without subsampling.

Remark 2. The terms $\Sigma_{i=1}^{\dim(T^{*\perp})} \mathbb{E}[\|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M_i)\|_{\mathrm{F}}]^2$ and $\mathbb{E}[\mathrm{tr}(\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}} \mathcal{P}_{T^{*\perp}})^{1/2}]^2$ for F are incomparable in general. The term $\mathbb{E}[\|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M_i)\|_{\mathrm{F}}]^2$ depends on the specific choice of basis, and it is useful in scenarios in which a particular choice of $\{M_i\}_{i=1}^{\dim(T^{*\perp})}$ is natural, such as in variable selection problems in which the standard basis has a clear interpreta-

tion. In contrast, $\mathbb{E}[\mathrm{tr}(\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}\mathcal{P}_{T^{*\perp}})^{1/2}]^2$ is basis independent and is more useful in problem settings in which no particular choice of a basis is natural.

Remark 3. The quantity $\kappa_{\mathrm{bag}}(\alpha)$ depends on commutators of projection operators associated with various tangent spaces. As such, this quantity is closer to 0 if the principal angles between T and $\hat{T}\{\mathcal{D}(n/2)\}^{\perp}$ and between $T^{*\perp}$ and $\hat{T}\{\mathcal{D}(n/2)\}$ are close to 0 or $\pi/2$. Note that in variable selection problems all the underlying projection matrices commute, and as a result we have that $\kappa_{\mathrm{bag}}(\alpha) = 0$. In this sense, $\kappa_{\mathrm{bag}}(\alpha)$ highlights the distinction between low rank estimation and variable selection.

Remark 4. Building on the previous remark, the commutativity property in the variable selection setting enables additional simplifications of our bounds. Although the bound (3.4) is valid for variable selection, exploiting the fact that the projection matrices commute in that case and with the choice of the standard basis for $\{M_i\}_{i=1}^{\dim(T^{*\perp})}$, we obtain additional simplifications. Specifically, letting $\{M_i\}_{i=1}^{\dim(T^{*\perp})}$ be the subset of the standard basis that lies in $T^{*\perp}$ and noting that κ_{bag} vanishes, one can modify the proof of theorem 1 to obtain the following bound:

$$\mathbb{E}[\mathrm{tr}(\mathcal{P}_T\mathcal{P}_{T^{*\perp}})] \leqslant \sum_{i=1}^{\dim(T^{*\perp})} \frac{\mathbb{E}[\|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M_i)\|_{\mathrm{F}}]^2}{2\alpha - 1} = \sum_{i=1}^{\dim(T^{*\perp})} \frac{\mathbb{P}[i\text{th null selected by } \hat{T}\{\mathcal{D}(n/2)\}]}{2\alpha - 1}. \tag{3.5}$$

This improved bound follows from a careful accounting of the first and third terms in bound (3.4); see the supplementary material section A.3. The equality here is a consequence of the observations that $\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}$ is a diagonal projection matrix and that each M_i is an element of the standard basis. Thus, we recover the interpretation that the overall expected false discovery for the special case of variable selection can be bounded in terms of the probability that the procedure $\hat{T}$ selects null variables on subsamples. The final expression (3.5) matches theorem 1 of Shah and Samworth (2013) (in particular, it holds for any $B \geqslant 2$). As a final comparison between the low rank estimation and variable selection settings, the dependence on α in inequality (3.5) is multiplicative as opposed to additive as in inequalities (3.4). In particular, in the low rank case even if the estimator $\hat{T}$ performs exceedingly well on the subsamples, the expected false discovery may still be large depending on the choice of α and $\dim(T^{*\perp})$; in contrast, for variable selection if the estimator $\hat{T}$ performs exceedingly well on the subsamples, the expected false discovery is small provided that α is chosen to be close to 1. This distinction is fundamental to the geometry underlying the sparse and determinantal varieties. Specifically, in the low rank case even if $\mathcal{P}_{\mathrm{avg}} \approx \mathcal{P}_{T^*}$ the set of stable tangent spaces $\mathcal{T}_\alpha$ necessarily includes many tangent spaces that are near the population tangent space T^* but are not perfectly aligned with it. This is because the collection of row–column spaces forms a Grassmannian manifold rather than a finite or discrete set. In contrast, if $\mathcal{P}_{\mathrm{avg}} \approx \mathcal{P}_{T^*}$ in variable selection, the only elements of the set of stable tangent spaces (for large α) are those corresponding to subsets of the true significant variables.

Next we provide a bound on both the basis-independent and basis-dependent versions of $\kappa_{\mathrm{bag}}(\alpha)$, which leads to a bag-independent bound on the expected false discovery by combining with theorem 1.

Proposition 1 (bounding $\kappa_{\mathrm{bag}}(\alpha)$ and a bag-independent result). Consider the set-up of theorem 1. Then the following bound holds for both the basis-independent and the basis-dependent versions of $\kappa_{\mathrm{bag}}(\alpha)$: $\kappa_{\mathrm{bag}}(\alpha) \leqslant 2\sqrt{1-\alpha}\,\mathbb{E}[\dim(T)]$. Further, letting the average number

of discoveries from $n/2$ observations be denoted by $q := \mathbb{E}[\dim[\hat{T}\{\mathcal{D}(n/2)\}]]$, we also have that $\mathbb{E}[\dim(T)] \leqslant q/\alpha$. Thus, we obtain the following false discovery bound for any $B \geqslant 2$:

$$\mathbb{E}[\mathrm{tr}(\mathcal{P}_T \mathcal{P}_{T^{*\perp}})] \leqslant F + 2\{1-\alpha+\sqrt{(1-\alpha)}\}\mathbb{E}[\dim(T)] \leqslant F + \frac{2q}{\alpha}\{1-\alpha+\sqrt{(1-\alpha)}\} \tag{3.6}$$

for either the basis-dependent or the basis-independent form of F from theorem 1.

Remark 5. The proof of the result is presented in the on-line supplementary material section A.2. This bound highlights the role of α, where $\kappa_{\mathrm{bag}}(\alpha)$ becomes smaller as α is chosen close to 1. The bag-independent bound (3.6) on expected false discovery of a stable tangent space holds for any $B \geqslant 2$, and thus can be looser than bound (3.4). In particular, bound (3.6) is relevant for $\alpha \gtrapprox 0.9$ (as the bound otherwise exceeds q), which is more stringent than the condition $\alpha > \frac{1}{2}$ in theorem 1. Despite the more restrictive range of values for α, these bag-independent results may nonetheless have utility in regimes in which the signal strength is high so that larger values of α may be considered.

Next we describe a more refined false discovery bound under additional assumptions on the estimator $\hat{T} = (\hat{C}, \hat{R})$.

Assumption 1.

$$\frac{\mathbb{E}[\mathrm{tr}(\mathcal{P}_{T^{*\perp}} \mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}})]}{\dim(T^{*\perp})} \leqslant \frac{\mathbb{E}[\mathrm{tr}(\mathcal{P}_{T^*} \mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}})]}{\dim(T^*)}$$

Assumption 2. The distribution of $\|\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}(M)\|_{\mathrm{F}}$ is the same for all rank 1 $M \in T^{*\perp}$, $\|M\|_{\mathrm{F}} = 1$

In words, assumption 1 states that the estimator's normalized power is greater than its normalized expected false discovery and assumption 2 states that the energy of any normalized rank 1 element in $T^{*\perp}$ onto tangent spaces obtained from subsamples consisting of $n/2$ observations is identically distributed. In the case of variable selection, assumption 1 reduces precisely to the 'better than random guessing' assumption employed by Meinshausen and Bühlmann (2010), namely that the probability that the procedure $\hat{T}$ selects a null variable when employed on the subsamples is better than random guessing. As a second condition, Meinshausen and Bühlmann (2010) required that the random variables $\{\mathbb{I}_{k \in \hat{T}\{\mathcal{D}(n/2)\}}\}$ are exchangeable. This assumption implies that the distribution of $\mathbb{I}_{k \in \hat{T}\{\mathcal{D}(n/2)\}}$ is the same for all null k. Our assumption 2 when specialized to variable selection reduces to the weaker requirement that each of the random variables $\mathbb{I}_{k \in \hat{T}\{\mathcal{D}(n/2)\}}$ has the same distribution. In supplementary material section A.4, we show that assumptions 1 and 2 are satisfied by some natural ensembles and estimators in low rank estimation problems. We prove next a bound on the expected false discovery under these additional assumptions.

Proposition 2 (refined false discovery control). Consider the set-up of theorem 1. Suppose additionally that assumptions 1 and 2 are satisfied. For any $M \in T^{*\perp}$ with $\mathrm{rank}(M) = 1$, $\|M\|_{\mathrm{F}} = 1$, the false discovery of a stable tangent space T is bounded by

$$\mathbb{E}[\mathrm{tr}(\mathcal{P}_T \mathcal{P}_{T^{*\perp}})] \leqslant \frac{q^2}{p_1 p_2} + f(\kappa_{\mathrm{indiv}}) + \frac{2q}{\alpha}\{1-\alpha+\sqrt{(1-\alpha)}\}, \tag{3.7}$$

where $\kappa_{\mathrm{indiv}} := \mathbb{E}[\|[\mathcal{P}_{\mathrm{span}(M)}, \mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}]\|_{\mathrm{F}}]$ and $f(\kappa_{\mathrm{indiv}}) = p_1 p_2 \kappa_{\mathrm{indiv}}^2 + 2q\kappa_{\mathrm{indiv}}$.

Remark 6. The proof of proposition 2 can be found in the on-line supplementary material section A.5. It proceeds by showing that, in the bag-dependent setting (although the specific choice does not matter because of assumption 2), $F \leqslant q^2/(p_1 p_2) + f(\kappa_{\text{indiv}})$ and employs the bounds $\kappa_{\text{bag}} \leqslant 2\sqrt{(1-\alpha)}\mathbb{E}[\dim(T)]$ and $\mathbb{E}[\dim(T)] \leqslant q/\alpha$ (from proposition 1). Consider the term $q^2/(p_1 p_2)$ in this result, where q can be approximated by $q \approx \text{tr}(\mathcal{P}_{\text{avg}})$. Suppose that typical outputs of the estimates obtained from subsamples $\{\mathcal{D}_l\}_{l=1}^B$ have rank k which is far smaller than the ambient dimensions, i.e. $k \ll \min\{p_1, p_2\}$, yielding $\text{tr}(\mathcal{P}_{\text{avg}}) = \mathcal{O}\{k(p_1 + p_2)\}$; as a result, $q^2/(p_1 p_2)$ is much smaller than q. The second term is an increasing function of the commutator-dependent quantity κ_{indiv}. To bound κ_{indiv} we note that it suffices to consider a single $M \in T^{*\perp}$ with $\text{rank}(M) = 1$, $\|M\|_{\text{F}} = 1$. A natural data-driven heuristic to obtain such an M is to consider a rank 1 matrix that is 'least aligned' with $\mathcal{P}_{\text{avg}}$, i.e. in some sense choosing the opposite of a stable tangent space. Concretely, letting u and v be the singular vectors corresponding to the smallest singular values of $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ respectively, we propose setting $\tilde{M} = uv'$. This choice can be justified theoretically provided that the estimator $\hat{T}\{\mathcal{D}(n/2)\}$ has good power; see supplementary material section A.6. We then obtain the following data-driven approximation $\kappa_{\text{indiv}} = (1/B)\Sigma_{l=1}^B \|[\mathcal{P}_{\hat{T}(\mathcal{D}_l)}, \mathcal{P}_{\text{span}(\tilde{M})}]\|_{\text{F}}$. Finally, the third term can be controlled by choosing α sufficiently close to 1.

Remark 7. For the case of variable selection, $\kappa_{\text{indiv}} = 0$, so $F \leqslant q^2/\text{total variables}$. Plugging this into expression (3.5), we obtain the bound on the expected false discovery of $\mathbb{E}[\#\text{ discoveries in } \hat{T}\{\mathcal{D}(n/2)\}]^2/\{2(1-\alpha)(\#\text{ total variables})\}$. This bound was obtained by Shah and Samworth (2013) as a consequence of their theorem 1 and it holds for any $B \geqslant 2$ (an identical bound was also obtained by Meinshausen and Bühlmann (2010), although that result requires averaging over all subsamples).

3.5. Subspace stability selection algorithm

As described in the previous subsection, every tangent space in $\mathcal{T}_\alpha$ provides control on the expected false discovery. The goal then is to select an element of $\mathcal{T}_\alpha$ to optimize power. A natural approach to achieve this objective is to choose a tangent space of largest dimension from $\mathcal{T}_\alpha$ to maximize the total discovery.

Consider the following optimization problem for each $r = 1, \ldots, \min\{p_1, p_2\}$:

$$T_{\text{OPT}}(r) = \underset{T \text{ tangent space to a point in } \mathcal{V}_{\text{low rank}}(r)}{\arg\max} \sigma_{\min}(\mathcal{P}_T \mathcal{P}_{\text{avg}} \mathcal{P}_T). \tag{3.8}$$

A conceptually appealing approach to select an optimal tangent space is via the optimization problem

$$T_{\text{OPT}} \in \underset{T \in T_{\text{OPT}}(r) \cap \mathcal{T}_\alpha}{\arg\max} r, \tag{3.9}$$

where, by construction, the set $T_{\text{OPT}}(r) \cap \mathcal{T}_\alpha$ is non-empty if $\mathcal{T}_\alpha$ is a non-empty set. In the case of variable selection, this procedure would result in the selection of all those variables that are estimated as being significant in at least an α-fraction of the bags, which is in agreement with the procedure of Meinshausen and Bühlmann (2010). In our setting of low rank estimation, however, we are not aware of a computationally tractable approach to solve problem (3.8). The main source of difficulty lies in the geometry underlying the collection of tangent spaces to determinantal varieties. In particular, solving problem (3.8) in the case of variable selection is easy because the operators $\mathcal{P}_T$ and $\mathcal{P}_{\text{avg}}$ are both diagonal (and hence trivially simultaneously diagonalizable) in that case; as a result, we can decompose problem (3.8) into a set of one-variable

problems. In contrast, the operators $\mathcal{P}_T$ and $\mathcal{P}_{\text{avg}}$ are not simultaneously diagonalizable in the low rank case, and consequently there does not appear to be any clean separability in problem (3.8) in general with determinantal varieties.

We describe next a heuristic to approximate expression (3.8). Our approximation entails computing optimal row space and column space approximations from the bags separately rather than in a combined fashion via tangent spaces. Specifically, suppose that $\{(\hat{\mathcal{C}}(\mathcal{D}_l), \hat{\mathcal{R}}(\mathcal{D}_l))\}_{l=1}^B$ denote the row–column space estimates from B subsamples $\{\mathcal{D}_l\}_{l=1}^B \subset \mathcal{D}$ of the data. We average the projection operators that are associated with these row–column spaces:

$$\begin{aligned} \mathcal{P}_{\text{avg}}^{\mathcal{C}} &= \frac{1}{B}\sum_{l=1}^{B} \mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}_l)}, \\ \mathcal{P}_{\text{avg}}^{\mathcal{R}} &= \frac{1}{B}\sum_{l=1}^{B} \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_l)}. \end{aligned} \tag{3.10}$$

Note that the average projection operator $\mathcal{P}_{\text{avg}}$ based on estimates from subsamples of tangent spaces to determinantal varieties is a self-adjoint map on the space $\mathbb{R}^{p_1 \times p_2}$ whereas the averages $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ are self-adjoint maps on the spaces $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$ respectively. On the basis of these separate column space and row space averages, we approximate expression (3.8) as

$$\begin{aligned} T_{\text{approx}}(r) = T\Bigg\{ &\underset{\mathcal{C} \subset \mathbb{R}^{p_1} \text{ subspace of dimension } r}{\arg\max} \sigma_{\min}(\mathcal{P}_{\mathcal{C}} \mathcal{P}_{\text{avg}}^{\mathcal{C}} \mathcal{P}_{\mathcal{C}}), \\ &\times \underset{\mathcal{R} \subset \mathbb{R}^{p_2} \text{ subspace of dimension } r}{\arg\max} \sigma_{\min}(\mathcal{P}_{\mathcal{R}} \mathcal{P}_{\text{avg}}^{\mathcal{R}} \mathcal{P}_{\mathcal{R}}) \Bigg\}. \end{aligned} \tag{3.11}$$

The advantage of this formulation is that the inner optimization problems of identifying the best row space and column space approximations of rank r can be computed tractably. In particular, the optimal column space and row space approximations of dimension r are equal to the span of the eigenvectors corresponding to the r largest eigenvalues of $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ respectively. We have that $\sigma_{\min}(\mathcal{P}_{T_{\text{approx}}(r)} \mathcal{P}_{\text{avg}} \mathcal{P}_{T_{\text{approx}}(r)}) \leqslant \sigma_{\min}(\mathcal{P}_{T_{\text{OPT}}(r)} \mathcal{P}_{\text{avg}} \mathcal{P}_{T_{\text{OPT}}(r)})$ and we expect this inequality to be strict in general, even though tangent spaces to determinantal varieties are in one-to-one correspondence with the underlying row–column spaces. To see why this is so, consider a column space and row space pair $(\mathcal{C}, \mathcal{R}) \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, with $\dim(\mathcal{C}) = \dim(\mathcal{R}) = r$. The collection of matrices $\mathcal{M}_{\mathcal{C}} \subseteq \mathbb{R}^{p_1 \times p_2}$ with column space contained in $\mathcal{C}$ has dimension $p_2 r$ and the collection of matrices $\mathcal{M}_{\mathcal{R}} \subseteq \mathbb{R}^{p_1 \times p_2}$ with row space contained in $\mathcal{R}$ has dimension $p_1 r$. However, the tangent space $T(\mathcal{C}, \mathcal{R}) \subset \mathbb{R}^{p_1 \times p_2}$, which is the sum of $\mathcal{M}_{\mathcal{C}}$ and $\mathcal{M}_{\mathcal{R}}$, has dimension $p_1 r + p_2 r - r^2$. In other words, the spaces $\mathcal{M}_{\mathcal{C}}$ and $\mathcal{M}_{\mathcal{R}}$ do not have a transverse intersection (i.e. $\mathcal{M}_{\mathcal{C}} \cap \mathcal{M}_{\mathcal{R}} \neq \{0\}$), and therefore optimal tangent space estimation does not appear to be decoupled into (separate) optimal column space estimation and optimal row space estimation. Although this heuristic is only an approximation, it does yield good performance in practice, as described in the illustrations in the next subsection as well as in the experiments with real data in Section 4. Further, our final estimate of a tangent space still involves the solution of problem (3.9) by using approximation (3.11) instead of (3.8). Consequently, we continue to retain our guarantees from Section 3.1 on false discovery control. The full procedure is presented in algorithm 1 in Table 1.

The tuning parameter $\alpha \in [0, 1]$ in algorithm 1 plays an important role in how much signal is selected by subspace stability selection. In our experience, the output of subspace stability selection is quite robust to α in moderate to high SNR settings. As a result, in all our experiments we

Table 1. Algorithm 1: subspace stability selection algorithm

Step 1: *input*, a set of observations $\mathcal{D}$, a collection of subsamples $\{\mathcal{D}_l\}_{l=1}^{B} \subset \mathcal{D}$, a row–column space (equivalently, tangent space) estimation procedure $(\hat{\mathcal{C}}, \hat{\mathcal{R}})$ and a parameter $\alpha \in (0, 1)$
Step 2: *obtain tangent space estimates*; for each bag $\{\mathcal{D}_l,\ l = 1, 2, \ldots, B\}$, obtain row–column space estimates $\{(\hat{\mathcal{C}}(\mathcal{D}_l), \hat{\mathcal{R}}(\mathcal{D}_l))\}_{l=1}^{B}$ and set $\hat{T}(\mathcal{D}_l) = T\{\hat{\mathcal{C}}(\mathcal{D}_l), \hat{\mathcal{R}}(\mathcal{D}_l)\}$
Step 3: *compute average projection operators*; compute the average tangent space projection operator $\mathcal{P}_{\text{avg}}$ according to expression (3.2) and the average row–column space projection operators $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ according to expression (3.1)
Step 4: *compute optimal row–column space approximations*; compute ordered singular vectors $\{u_1, u_2, \ldots, u_{p_1}\} \subset \mathbb{R}^{p_1}$ and $\{v_1, v_2, \ldots, v_{p_2}\} \subset \mathbb{R}^{p_2}$ of $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ respectively; for each $r = 1, \ldots, \min\{p_1, p_2\}$, set $\bar{\mathcal{C}}(r) = \text{span}(u_1, \ldots, u_r)$ and $\bar{\mathcal{R}}(r) = \text{span}(v_1, \ldots, v_r)$
Step 5: *tangent space selection via* expression (3.9); let r_{S3} denote the largest r such that $T\{\bar{\mathcal{C}}(r), \bar{\mathcal{R}}(r)\} \in \mathcal{T}_\alpha$
Step 6: *output*, tangent space $T_{\text{S3}} = T\{\bar{\mathcal{C}}(r_{\text{S3}}), \bar{\mathcal{R}}(r_{\text{S3}})\}$

select α to equal 0.70. For detailed analysis on the sensitivity to α see the on-line supplementary material section A.7.

3.5.1. Computational cost of algorithm 1

We do not account for the cost of obtaining the row–column space estimates $\{(\hat{\mathcal{C}}(\mathcal{D}_l), \hat{\mathcal{R}}(\mathcal{D}_l))\}_{l=1}^{B}$ on each subsample in step 2 of algorithm 1 and focus exclusively on the cost of combining these estimates via steps 3–5. In step 3, the computational complexity of computing the averages $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ requires $\mathcal{O}[B \max\{p_1, p_2\}^2]$ operations and computing the average $\mathcal{P}_{\text{avg}}$ requires $\mathcal{O}(Bp_1^2 p_2^2)$ operations. Step 4 entails the computation of two singular value decompositions of matrices of size $p_1 \times p_1$ and $p_2 \times p_2$, which leads to a cost of $\mathcal{O}[\max\{p_1, p_2\}^3]$ operations. Finally, in step 5, to check membership in $\mathcal{T}_\alpha$ we multiply three maps of size $p_1 p_2 \times p_1 p_2$ and compute the singular value decomposition of the result, which requires a total of $\mathcal{O}(p_1^3 p_2^3)$ operations. Thus, the computational cost of algorithm 1 to aggregate estimates produced by B bags is $\mathcal{O}[\max\{Bp_1^2, Bp_2^2, Bp_1^2 p_2^2, p_1^3, p_2^3, p_1^3 p_2^3\}]$.

Although the scaling of algorithm 1 is polynomial in the size of the inputs, when either p_1 or p_2 is large the overall cost due to terms such as $p_1^3 p_2^3$ may be prohibitive. In particular, the reason for the expensive terms $Bp_1^2 p_2^2$ and $p_1^3 p_2^3$ in the final expression is computations involving projection maps onto tangent spaces (which belong to $\mathbb{R}^{p_1 p_2}$). We describe next a modification of algorithm 1 so that the resulting procedure consists of only computations involving projection maps onto row and column spaces (which belong to $\mathbb{R}^{p_2}$ and $\mathbb{R}^{p_1}$ respectively).

3.5.2. Modification of algorithm 1 and associated cost

The inputs to this modified procedure are the same as those of the original procedure. We modify step 3 of algorithm 1 by computing only the average row–column space projection maps $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$. Let $\mathcal{P}_{\text{avg}}^{\mathcal{C}} = U\Gamma U'$ and let $\mathcal{P}_{\text{avg}}^{\mathcal{R}} = V\Delta V'$ be the singular value decomposition computations of step 4. We modify step 5 of algorithm 1 to choose the largest r'_{S3} so that $\Gamma_{r'_{\text{S3}}, r'_{\text{S3}}} \geqslant \alpha$ and $\Delta_{r'_{\text{S3}}, r'_{\text{S3}}} \geqslant \alpha$. One can check that the cost that is associated with this modified procedure is $\mathcal{O}[\max\{Bp_1^2, Bp_2^2, p_1^3, p_2^3\}]$.

This modified method has the property that the row and column spaces are individually well aligned with the corresponding averages from the subsamples; the following result shows that the resulting tangent space belongs to a set of stable tangent spaces.

Proposition 4 (modified algorithm 1 satisfies subspace stability selection criterion). Let $T_{\text{S3-modified}}$ be the output of the modified algorithm 1 with input parameter α. Then, $T_{\text{S3-modified}} \in \mathcal{T}_{1-4(1-\alpha)}$.

Proposition 4 guarantees that our modification of algorithm 1 continues to provide false discovery control. We use this modified approach in some of our larger experiments in Section 4. The proof of this proposition can be found in the on-line supplementary material section A.8.

Finally we remark that, in subspace estimation problems (see Section 2.1), the subspace stability selection can be readily employed to find a stable tangent space. In particular, recall from Section 3.1 that the stability selection criterion (3.3) reduces to finding $\mathcal{C}$ such that $\sigma_{\min}(\mathcal{P}_{\mathcal{C}}\mathcal{P}_{\text{avg}}^{\mathcal{C}}\mathcal{P}_{\mathcal{C}}) \geqslant \alpha$. Naturally, a projection operator $\mathcal{P}_{\mathcal{C}}$ that satisfies the criterion above can be obtained via singular value thresholding. Furthermore, this subspace estimate is optimal according to problem (3.9).

3.6. Further illustrations

In the remainder of this section, we explore various facets of algorithm 1 via illustrations on the synthetic matrix completion problem set-up that was described at the beginning of Section 3. For further demonstrations of the utility of subspace stability selection with real data, we refer the reader to the experiments of Section 4.

3.6.1. Illustration: α versus r_{S3}

The threshold parameter α determines the eventual optimal rank r_{S3}, with larger values of α yielding a smaller r_{S3}. To understand this relationship better, we plot in Fig. 2 $\sigma_{\min}(\mathcal{P}_{T_{\text{S3}}}\mathcal{P}_{\text{avg}}\mathcal{P}_{T_{\text{S3}}})$ as a function of r_{S3} for a large range of values of the regularization parameter λ and SNR $= \{0.4, 0.8, 1.2, 50\}$. Each curve in the plots corresponds to a particular value of r_{S3}, with the full curves representing $r_{\text{S3}} = 1, \ldots, 10$ and the dotted curves representing $r_{\text{S3}} = 11, \ldots, 70$. As smaller values of r_{S3} lead to larger values of $\sigma_{\min}(\mathcal{P}_{T_{\text{S3}}}\mathcal{P}_{\text{avg}}\mathcal{P}_{T_{\text{S3}}})$, the curves are ordered such that the top curve corresponds to $r_{\text{S3}} = 1$ and the bottom curve corresponds to $r_{\text{S3}} = 70$. We first observe that, for a fixed r_{S3}, the associated curve is generally decreasing as a function of λ. For large values of λ, both signal and noise are substantially reduced because of a significant amount of regularization. Conversely, for small values of λ, both signal and noise are present to a greater degree in the estimates on each subsample; however, the averaging procedure reduces the effect of noise, which results in high quality aggregated estimates for smaller values of λ. Next, we observe that the curves that are indexed by the r_{S3}-cluster in the high SNR-regime, with the first three corresponding to $r_{\text{S3}} = 1, 2, 3$, the next five corresponding to $r_{\text{S3}} = 4, \ldots, 8$, the next two corresponding to $r_{\text{S3}} = 9, 10$ and finally the remaining curves corresponding to $r_{\text{S3}} > 10$. This phenomenon is due to the clustering of the singular values of the underlying population L^*. In contrast, for low values of SNR, the clustering is less pronounced as the components of L^* with small singular values are overwhelmed by noise.

3.6.2. Illustration: subspace stability selection reduces false discovery

Next, we demonstrate that subspace stability selection produces a tangent space which is different and usually of a higher quality (e.g. smaller expected false discovery) than the base estimator applied to the full data set. We choose the noise level so that SNR takes one of the values in $\{1.5, 2, 2.5, 3\}$. In contrast, we employ procedure (3.1) on a subset of 2231 observations (the training set) of the full set of 3186 observations and the remaining subset of 955 observations

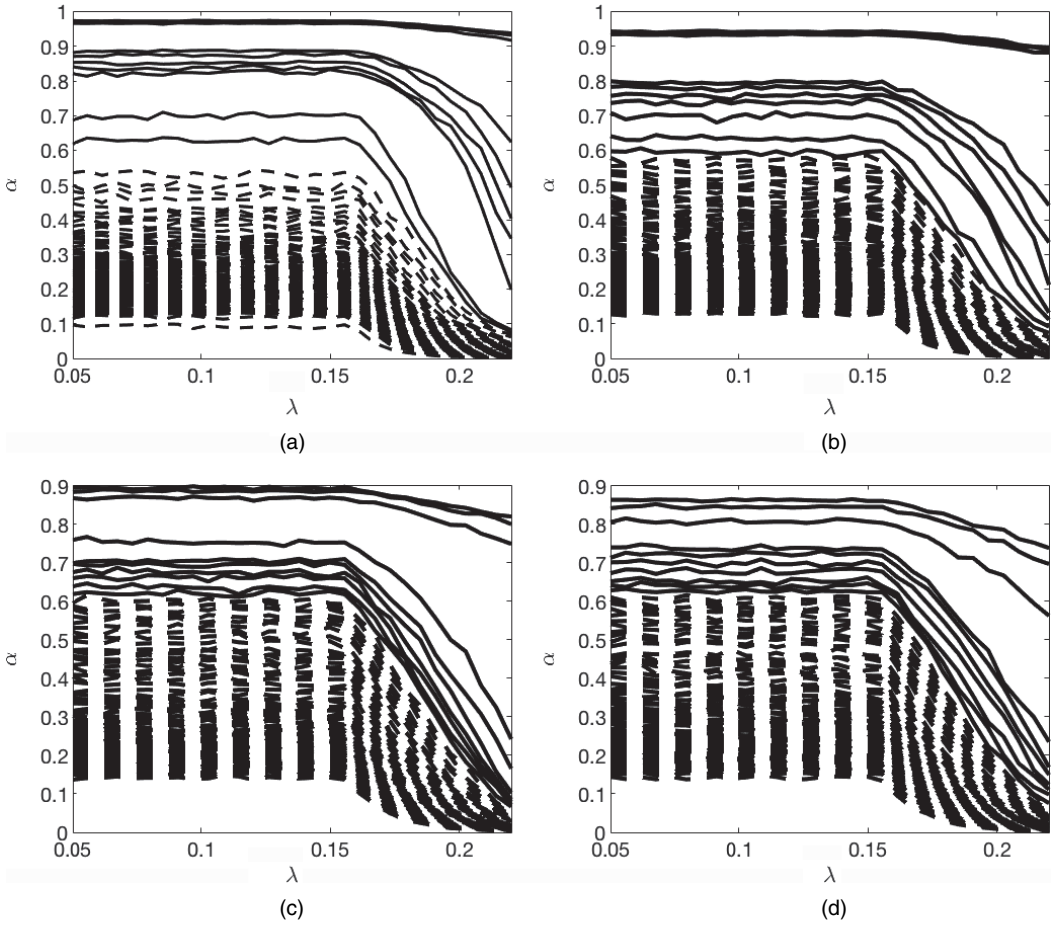


Fig. 2. Relationship between r_{S3} and α in algorithm 1 for a large range of λ and SNR (a) 0.4, (b) 0.8, (c) 1.2 and (d) 50

constitute the test set. We use cross-validation to identify an optimal choice λ^* of the regularization parameter. The estimate that is produced by procedure (3.1) on the training set for this choice of λ^* is recorded as the output of the non-subsampled approach. In contrast, estimator (3.1) with the choice λ^* is used in conjunction with $\alpha = 0.7$ to produce a subspace stability selection tangent space via algorithm 1. For each of the four choices of SNR, we ran 100 experiments and averaged to find an empirical approximation to the expected false discovery (2.3). Table 2 compares the expected false discovery (with 1σ-statistics) of the non-subsampled approach with that of the subspace stability selection procedure for the various problem settings. Evidently, subspace stability selection yields a much smaller amount of false discovery compared with not employing subsampling.

At this stage, it is natural to wonder whether the source of the improved false discovery control that is provided by subspace stability selection over not using subsampling is simply because the non-subsampled approach provides estimates with a larger rank. In particular, as an extreme hypothetical example, the zero-dimensional space is a stable tangent space and has zero expected false discovery, and more generally lower rank tangent space estimates are likely to

Table 2. False discovery of subspace stability selection *versus* a non-subsampled approach on the stylized matrix completion problem†

Method	*Results for the following SNRs:*			
	SNR = 1.5	*SNR = 2*	*SNR = 2.5*	*SNR = 3*
No subsampling	1274.6 ± 78.8	1532.8 ± 68.5	1573.5 ± 71.2	1417 ± 63.5
Subspace stability selection	107.6 ± 11.5	89.7 ± 16.9	87.9 ± 18.7	87.9 ± 19.4

†The maximum possible amount of false discovery is $\dim(T^{*\perp}) = (70-10)^2 = 3600$.

have smaller expected false discovery. Thus, is subsampling better primarily because it produces lower rank estimates? To address this point in our stylized set-up, we consider a population L^* with associated incoherence parameter equal to 0.8. (The incoherence of a matrix M is $\max_i \max\{\|\mathcal{P}_{\text{col-space}(M)}(e_i)\|_2^2, \|\mathcal{P}_{\text{row-space}(M)}(e_i)\|_2^2\}$ where e_i is the ith standard basis vector, and it plays a prominent role in various analyses of the low rank matrix completion problem (Candès and Recht, 2009).) We sweep over the regularization parameter λ, and we compare the following two estimates: first, the estimate $\hat{L}$ obtained via expression (3.1) and then truncated to its first three singular values, and subsampled estimates obtained via algorithm 1 with r_{S3} set to 3. The choice of 3 here is motivated by the fact that the population low rank matrix L^* has three large components. We perform this comparison for $\text{SNR} = \{0.8, 1.6\}$ and describe the results in the plots in Fig. 3. In the high SNR-regime, the performances of the subsampled and the non-subsampled approaches are similar. However, in the low SNR-regime, subspace stability selection yields a tangent space with far less false discovery across the entire range of regularization parameters. Further, subspace stability selection provides a fundamentally different solution that cannot be reproduced simply by selecting the 'right' regularization penalty in expression (3.1) applied to the entire data set.

Similar behaviour is also observed when the solution $\hat{L}$ is truncated at a different rank. As an example, with $\text{SNR} = 0.8$, we choose λ via cross-validation and truncate $\hat{L}$ at rank $r = 1, 2, \ldots, 5$ and compare its false discovery estimate with the estimate that is produced by subspace stability selection with $r_{S3} = r$ (shown in Table 3).

3.6.3. Illustration: stability of tangent spaces to small changes in regularization parameter

Finally, we note that, in settings in which regularization is employed, the estimate can be extremely sensitive to the choice of regularization parameter. For example, in nuclear-norm-regularized formulations such as expression (3.1), small changes to the parameter λ can often lead to substantial changes in the optimal solution. A virtue of subspace stability selection is that the estimates that it provides are generally very stable to small perturbations of λ. To formalize this discussion, given two tangent spaces T and $\tilde{T}$, we consider the quantity

$$\mu(T, \tilde{T}) \triangleq 1 - \frac{\text{tr}(\mathcal{P}_T \mathcal{P}_{\tilde{T}})}{\max\{\dim(T), \dim(\tilde{T})\}}$$

which measures the degree to which T and $\tilde{T}$ are misaligned. If $T = \tilde{T}$, then $\mu(T, \tilde{T}) = 0$ and, in contrast, $T \subseteq \tilde{T}^{\perp}$ would yield $\mu(T, \tilde{T}) = 1$. Hence, larger values of $\mu(T, \tilde{T})$ are indicative of greater deviations between T and $\tilde{T}$. We use this metric to compare the stability of the non-subsampled

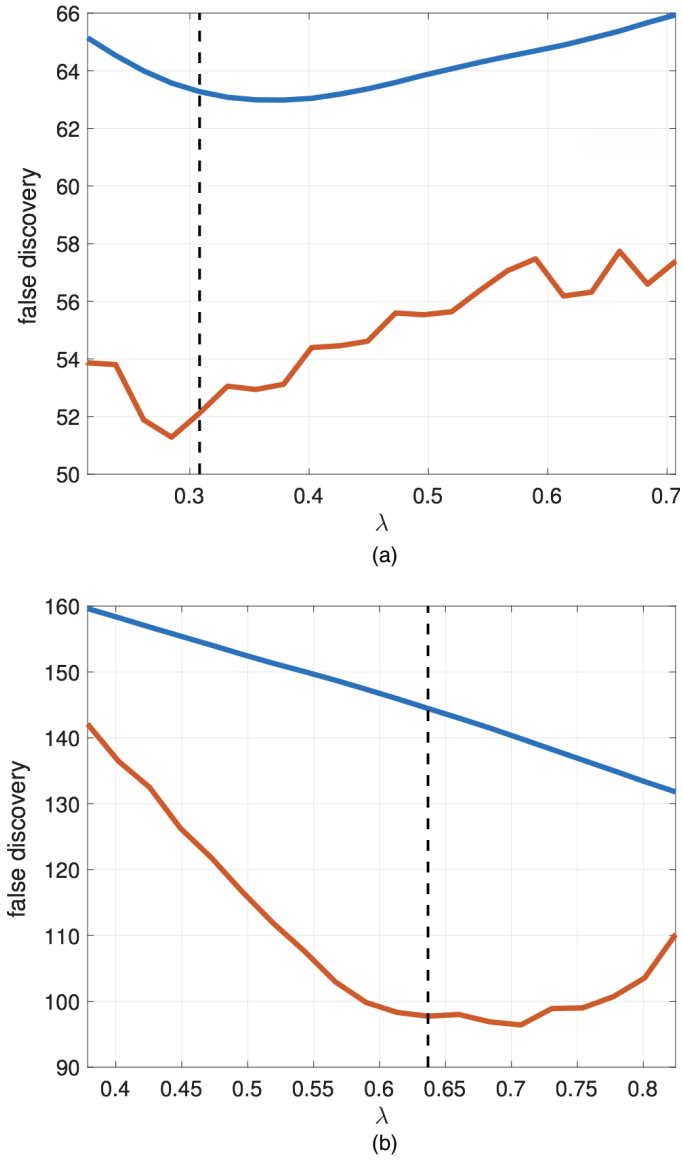


Fig. 3. False discovery of subspace stability selection *versus* a non-subsampled approach with (a) SNR = 1.6 and (b) SNR = 0.8 (here, we choose a rank 3 approximation of the non-subsampled approach and $r_{S3} = 3$ in algorithm 1 of subspace stability selection; the maximum possible amount of false discovery is $\dim(T^{*\perp}) = (70 - 10)^2 = 3600$): ——, no subsampling; ——, subspace stability selection

approach with subspace stability selection. In our stylized set-up, we choose the noise level so that $\mathrm{SNR} = 4$ and we select $\lambda = 0.03$ (based on cross-validation). Letting T be the tangent space of estimator (3.1) with $\lambda = 0.03$ and $\tilde{T}$ with $\lambda = 0.05$, we find that $\mu(T, \tilde{T}) = 0.23$. Setting $\alpha = 0.7$ with $B = 100$ complementary bags and computing the same metrics for the outputs of subspace stability selection, we find that $\mu(T, \tilde{T}) = 0.003$. This contrast is observed for many other SNR-levels.

Table 3. False discovery of subspace stability selection *versus* a non-subsampled approach with SNR = 0.8 and rank of the estimate set to vary from 1 to 5†

Method	*Results for the following ranks:*				
	rank = 1	*rank = 2*	*rank = 3*	*rank = 4*	*rank = 5*
No subsampling	20.4	48.1	89.7	146.7	218.8
Subspace stability selection	12.4	25.6	44.3	70.4	109

†The maximum possible amount of false discovery is $\dim(T^{*\perp}) = 3600$.

4. Experiments

In this section, we demonstrate the utility of subspace stability selection in providing false discovery control with both synthetic and real data. We consider the following types of low rank estimation problems.

4.1. Low rank linear measurements and matrix completion

We consider noisy linear functions of a low rank matrix $L^* \in \mathbb{R}^{p_1 \times p_2}$ of the form $Y_i \approx \langle \mathcal{A}_i, L^* \rangle$, $i = 1, \ldots, n$, where each $\mathcal{A}_i \in \mathbb{R}^{p_1 \times p_2}$. In the linear measurement setting, $\mathcal{A}_i$ is an arbitrary sensing matrix and, in the matrix completion setting, $\mathcal{A}_i$ consists of 0s everywhere except a single entry which is equal to 1. The matrix completion problem is similar to that considered in the stylized demonstrations of Section 3.1. One point of departure from that discussion in the present section is that, in experiments where the dimensions p_1 and p_2 are large, employing the nuclear norm regularized estimator (3.1) on each subsample is impractical. Instead, we use on each subsample the following non-convex formulation:

$$(\hat{U}, \hat{V}) = \underset{U \in \mathbb{R}^{p_1 \times k}, V \in \mathbb{R}^{p_2 \times k}}{\arg\min} \sum_{i \in S} (Y_i - \langle \mathcal{A}_i, UV' \rangle)^2 + \lambda(\|U\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2). \tag{4.1}$$

where $\|U\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2$ is a surrogate for the nuclear norm penalty (3.1), $\lambda > 0$ is a regularization parameter and $S \subset \{1, \ldots, p_1\} \times \{1, \ldots, p_2\}$ is the set of observed indices. By construction, $\hat{L} = \hat{U}\hat{V}'$ is constrained to have rank at most k, and this rank can be adjusted by appropriately tuning λ. Fixing U and V the above problem is convex in V and U respectively, and thus a commonly employed approach in practice is alternating least squares (ALS).

4.2. Factor analysis

We observe samples $\{Y^{(i)}\}_{i=1}^n \subset \mathbb{R}^p$ of a random vector and we identify a factor model that best explains these observations, i.e. a model in which the co-ordinates of the observed vector are independent conditioned on a small number $k \ll p$ of latent variables. In other words, our objective is to approximate the sample covariance of $\{Y^{(i)}\}_{i=1}^n$ by a covariance matrix that is decomposable as the sum of a diagonal matrix and a positive semidefinite low rank matrix. Using the Woodbury inversion lemma, we have that the precision matrix can be decomposed as a diagonal matrix minus a positive semidefinite low rank matrix. The virtue of working with precision matrices is that the log-likelihood function is concave with respect to this parameterization. On each subsample, we use the following estimator (Shapiro, 1982):

$$(\hat{D},\hat{L}) = \underset{L\in\mathbb{S}^p, D\in\mathbb{S}^p}{\arg\min} -\log\{\det(D-L)\} + \operatorname{tr}\left\{\left(\frac{1}{|S|}\sum_{i\in S} Y^{(i)}Y^{(i)\prime}\right)(D-L)\right\} + \lambda\operatorname{tr}(L), \quad (4.2)$$
$$\text{subject to } D-L\succ 0,\ L\succeq 0,\ D \text{ is diagonal.}$$

Here $\operatorname{tr}(\cdot)$ is the restriction of the nuclear norm to symmetric positive semidefinite matrices.

4.3. Synthetic simulations

We explore the role of the commutator in the false discovery bound of theorem 1 in a stylized matrix denoising problem. Specifically, we generate a population low rank matrix $L^*\in\mathbb{R}^{p\times p}$ with $p=200$, with $\operatorname{rank}(L^*)=6$, the non-zero singular values set to $\{120,100,80,30,20,10\}$ and the row and column spaces are sampled uniformly from the Steifel manifold. Once L^* has been generated, we also choose a basis for the orthogonal complements of the row–column spaces of L^* and we let $U^*QV^{*\prime}$ be the full SVD of L^*, i.e. $U^*,V^*\in\mathbb{R}^{p\times p}$ are orthogonal matrices and $Q\in\mathbb{R}^{p\times p}$ is a diagonal matrix that is zero padded. We obtain n noisy measurements of L^* of the form $Y_i=L^*+\delta(\gamma U^*D_iV^{*\prime}+\epsilon_i)$ for $j=1,2,\ldots,n$, where D_i is a diagonal matrix with IID standard Gaussian entries on the diagonal and $\epsilon_i\in\mathbb{R}^{p\times p}$ is a matrix with IID standard Gaussian entries. The parameter $\delta>0$ controls the signal-to-noise ratio and the parameter $\gamma>0$ controls the commutator term inside theorem 1. In particular, larger values of γ lead to a smaller commutator term since the measurements Y_i and L^* are all closer to being simultaneously diagonalizable. Geometrically, this corresponds to the principal angles between $T^{*\perp}$ and $\hat{T}\{\mathcal{D}(n/2)\}$ concentrating around 0 and $\pi/2$. We vary γ in the range $\{10,30\}$ and, for each γ, we chose δ so that $\mathrm{SNR}=0.15$ (here $\mathrm{SNR}=\mathbb{E}[\|L^*\|_2/\|\delta[\gamma U^*D_iV^*+\epsilon_i]\|_2]$). We obtain $n=2p$ measurements, and the estimator that we employ on a subsample computes best rank k approximation of the average over the data in the subsample (where k is selected *a priori*). In our first illustration, the estimator computes rank 6 approximations. We apply subspace stability selection with $\alpha\in[0.75,0.97]$ and $B=100$ complementary bags, and we obtain an empirical approximation of the expected false discovery over 100 trials. Since the population model is known, the quantities inside theorem 1 are readily obtainable. We set the orthonormal basis elements $\{M_i\}_{i=1}^{\dim(T^{*\perp})}$ needed to compute the term F in bound (3.4) to be $\{U^*_{:,6+i}V^{*\prime}_{:,6+j}\}_{i,j=1}^{p-6}$. Figs 4(a) and 4(b) compare the expected false discovery achieved by subspace stability selection with the bound of theorem 1, the average number of discoveries of subspace stability selection (i.e. $\mathbb{E}[\dim(T)]$), and simply computing a rank 6 approximation of the entire data without any subsampling. Figs 4(c) and 4(d) show a similar set of illustrations but with the estimator computing a rank 10 approximation. A number of points are worth noting from these plots. First, subspace stability selection performs far better than simply using computing low rank approximations on the entire data set; in particular, when the estimator computes rank 6 approximations and with $\gamma=10$, subspace stability selection chooses a rank 3 model for $\alpha=0.9$ and expected false discovery about 32 whereas a rank 6 approximation on the entire data set without subsampling yields an expected false discovery around 515. For comparison, the total amount of possible false discovery is $\dim(T^{*\perp})=37636$. Second, relative to the value of $\dim(T^{*\perp})$, the results provided by theorem 1 are very effective as they yield an expected false discovery bound between 300 and 1100 depending on the choice of α. Specifically, these bounds are also smaller than the average number of discoveries made by subspace stability selection as well as the expected false discovery of an estimator that operates on all the data with no subsampling. As a final remark, we also note that smaller values of the commutator (larger choice of γ) lead to better bounds on the expected false discovery, as predicted by theorem 1.

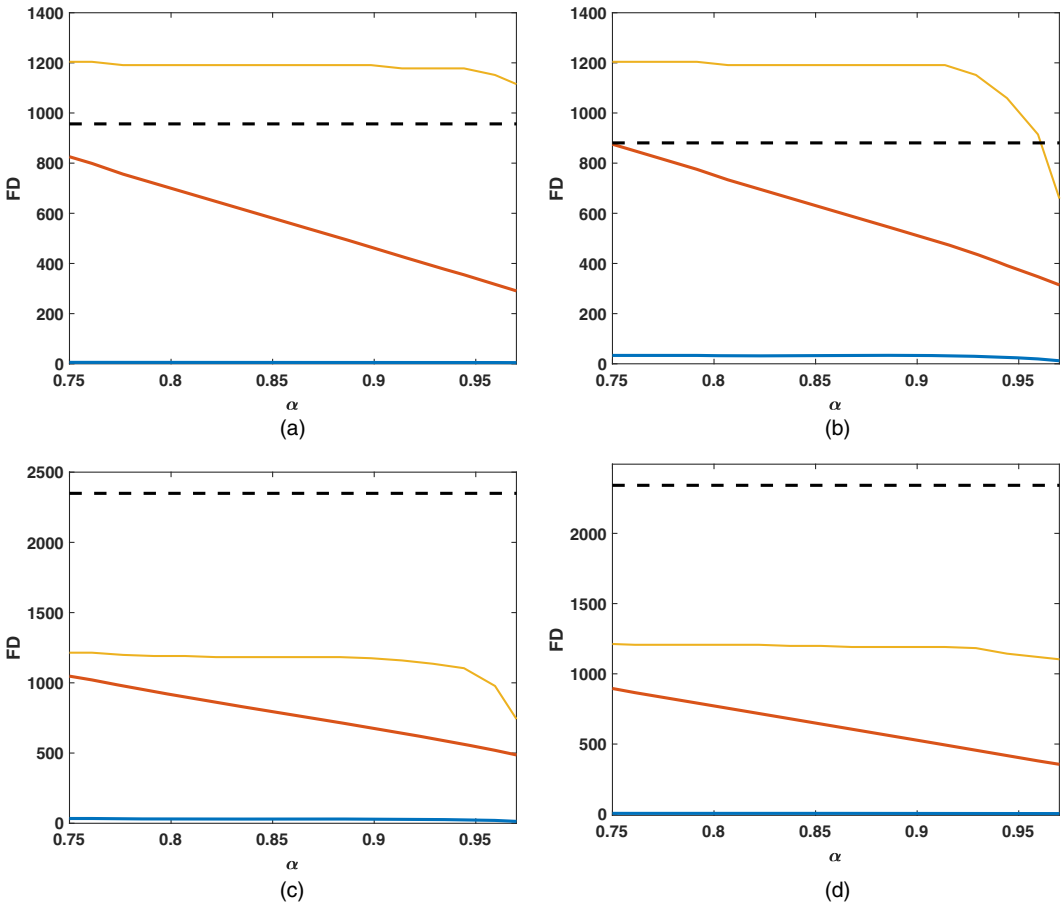


Fig. 4. False discovery of subspace stability selection as a function of α for the matrix completion setting (——, false discovery obtained by subspace stability selection; ——, theorem 1 bound; ——, proposition 1; ——, average dimension of the selected tangent space; - - - - - -, false discovery from using entire data) (subspace stability selection has small but non-zero false discoveries; as an example, for $\gamma = 20$, rank selected = 6, and $\alpha = 0.9$, subspace stability selection chooses typically a rank 3 model with 32.1 false discoveries; here $\dim(T^{*\perp}) = 37636$): (a) $\gamma = 30$, rank sel. = 6, $\|[\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}, \mathcal{P}_{T^{*\perp}}]\|_F \approx 52$; (b) $\gamma = 10$, rank sel. = 6, $\|[\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}, \mathcal{P}_{T^{*\perp}}]\|_F \approx 226$; (c) $\gamma = 30$, rank sel. = 10, $\|[\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}, \mathcal{P}_{T^{*\perp}}]\|_F \approx 81$; (d) $\gamma = 10$, rank sel. = 10, $\|[\mathcal{P}_{\hat{T}\{\mathcal{D}(n/2)\}}, \mathcal{P}_{T^{*\perp}}]\|_F \approx 291$

Next, we explore the false discovery and power attributes of subspace stability selection in various noise and rank regimes. We consider the linear Gaussian measurement setting that was described earlier with $p = 60$, the rank of L^* in the set $\{1, 2, 3, 4\}$, the non-zero singular values set to 1, and the row and column spaces sampled uniformly from the Steifel manifold. The measurements matrices $\{\mathcal{A}_i\}_{i=1}^n$ consist of IID entries drawn from $\mathcal{N}(0, 1)$. We obtain noisy measurements $Y_i = \langle \mathcal{A}_i, L^* + \epsilon \rangle$, $i = 1, \ldots, n$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The observation noise level σ^2 is tuned so that SNR (here $\mathbb{E}[\langle \mathcal{A}_i, L^* \rangle / \epsilon]$) lies in the set $\{1, 2, 3, 4, 5\}$. A fraction $n = 6p^2/10$ are used as training data for estimator (4.1) with λ chosen via hold-out validation with a validation set of size $3p^2/20$ and the rank constraint k set to 10. With this choice of λ, we evaluate the expectation and standard deviations of false discovery and the power empirically over 100 trials. As a point of comparison, we set $\alpha = 0.7$ with $B = 100$ complementary bags and compute

the same metrics based on subspace stability selection. We repeat a similar experiment in the matrix completion setting where $L^* \in \mathbb{R}^{p \times p}$ with $p = 100$, rank in the set $\{1, 2, 3, 4\}$ and row and column spaces chosen uniformly from a Steifel manifold. We select a fraction 7/10 of the total entries uniformly chosen at random as the observation set Ω so that $|\Omega| = 7p^2/10$. These observations are corrupted with Gaussian noise with variance selected so that SNR is in the range $\{0.5, 0.875, 1.25, 1.625, 2.00\}$. We use these observations as input to estimator (4.1), with λ selected on the basis of hold-out validation on an $n_{\text{test}} = 7/20p^2$ validation set.

Fig. 5 compares the performance of the non-subsampled approach and subspace stability selection computed empirically over 100 iterations for all the problem settings. For settings where either the false discovery standard deviation normalized by expected value or the power standard deviation normalized by expected value is greater than 0.01, we plot the expected value with a cross and the 1σ around the mean with a rectangle. Several settings in Fig. 5 experience a significant loss in power by using the subspace stability selection procedure. Those precisely correspond to models with high rank and low SNR-regime where some components of the signal are overwhelmed by noise. To control false discoveries in these settings, subspace stability selection filters out some of the signal and as a result yields a small power.

4.4. Experimental results on real data sets

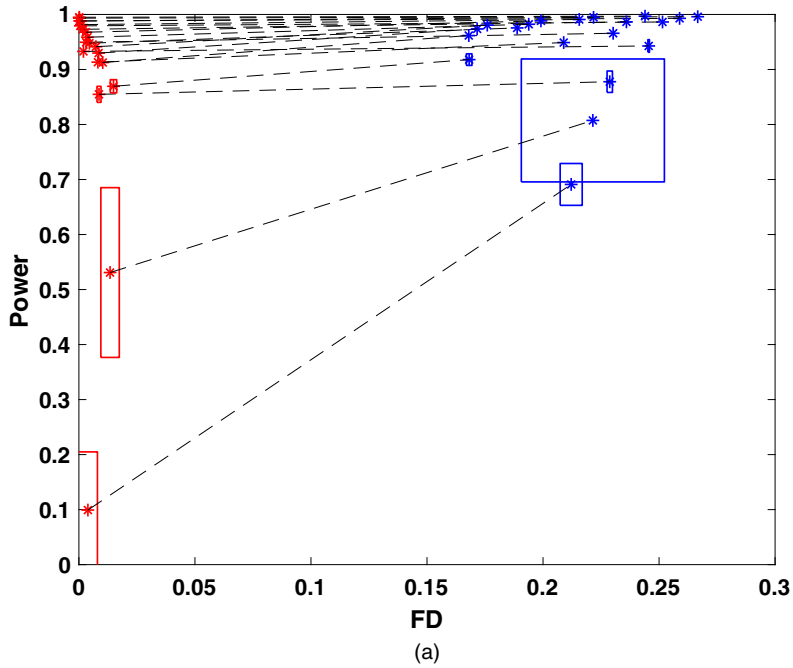
4.4.1. Collaborative filtering

In collaborative filtering, one is presented with partially filled user preference matrices in which rows are indexed by users and columns by items, with each entry specifying a user's preference for an item. The objective is to infer the unobserved entries. As discussed in Section 1, such user preference matrices are often well approximated as low rank, and therefore a popular approach to collaborative filtering is to frame it as a problem of low rank matrix completion and to solve this problem on the basis of either the convex relaxation (3.1) or the non-convex approach (4.1) via ALS. We describe experimental results on two popular data sets in collaborative filtering:

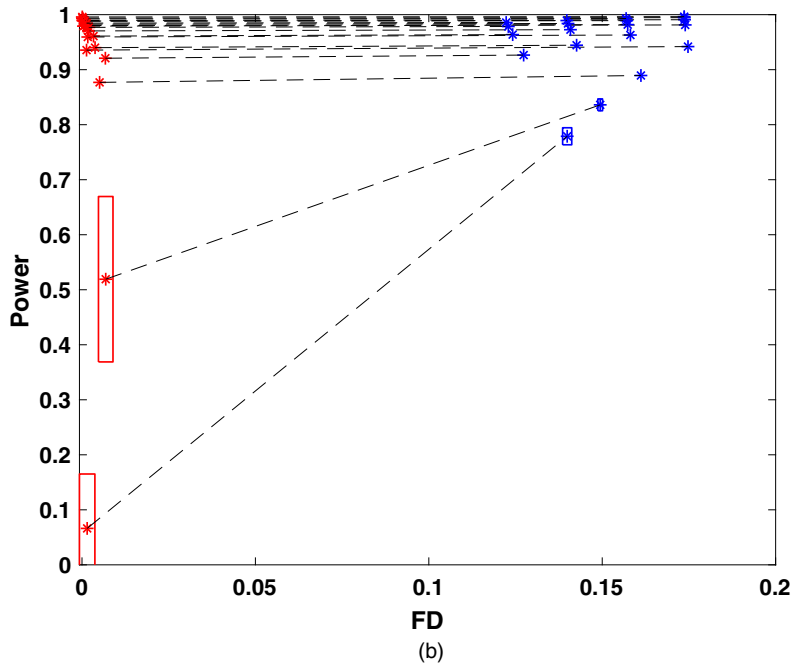
(a) the Amazon book crossing data set (obtained from `http://www2.informatik.uni-freiburg.de/~cziegler/BX/`) of which we consider a portion consisting of $p_1 = 1245$ users and $p_2 = 1054$ items with approximately 6% of the ratings (integer values from 1 to 10) observed, and
(b) the Amazon video games data set (obtained from `http://jmcauley.ucsd.edu/data/amazon/`) of which we consider a portion consisting of $p_1 = 482$ users and $p_2 = 520$ items with approximately 3.5% of the ratings (integer values from 1 to 5) observed.

In each case, we partition the data set as follows: we set aside 85% of the observations as a training set, 10% of the observations as a hold-out validation set and the remaining 5% as an evaluation set to assess the performance of our learned models.

As these problems are relatively large in size, we employ ALS on the non-convex formulation (4.1) with $k = 80$ (the upper bound on the rank) and we apply the modification of algorithm 1 for subspace stability selection. Finally, to obtain estimates of low rank matrices (as this is the eventual object of interest in collaborative filtering) we use formulation (2.7) given estimates of tangent spaces. We set $\alpha = 0.7$ and $B = 100$ complementary bags. Fig. 6 illustrates the mean-squared error (MSE) of ALS and subspace stability selection on the hold-out set for these two data sets for a range of values of the regularization parameter λ. For both data sets, we observe that subspace stability selection yields models with better MSE on the hold-out set over the entire range of regularization parameters. On the book crossings data set, we further note that, at the cross-validated λ, the rank of the estimate that is obtained from the non-subsampled



(a)



(b)

Fig. 5. False discovery *versus* power with (a) matrix completion and (b) linear measurements over 20 different problem instances (varying rank and noise level) (×, performance of the non-subsampled approach; ×, subspace stability selection with $\alpha = 0.7$): for the instances where the standard deviation divided by the mean is greater than 0.01, we show a 1σ-rectangle around the mean; the lines connect dots corresponding to the same problem instance; both the false discovery and the power are normalized by dividing expressions (2.3) and (2.4) by $\dim(T^{*\perp})$ and $\dim(T^{*})$ respectively

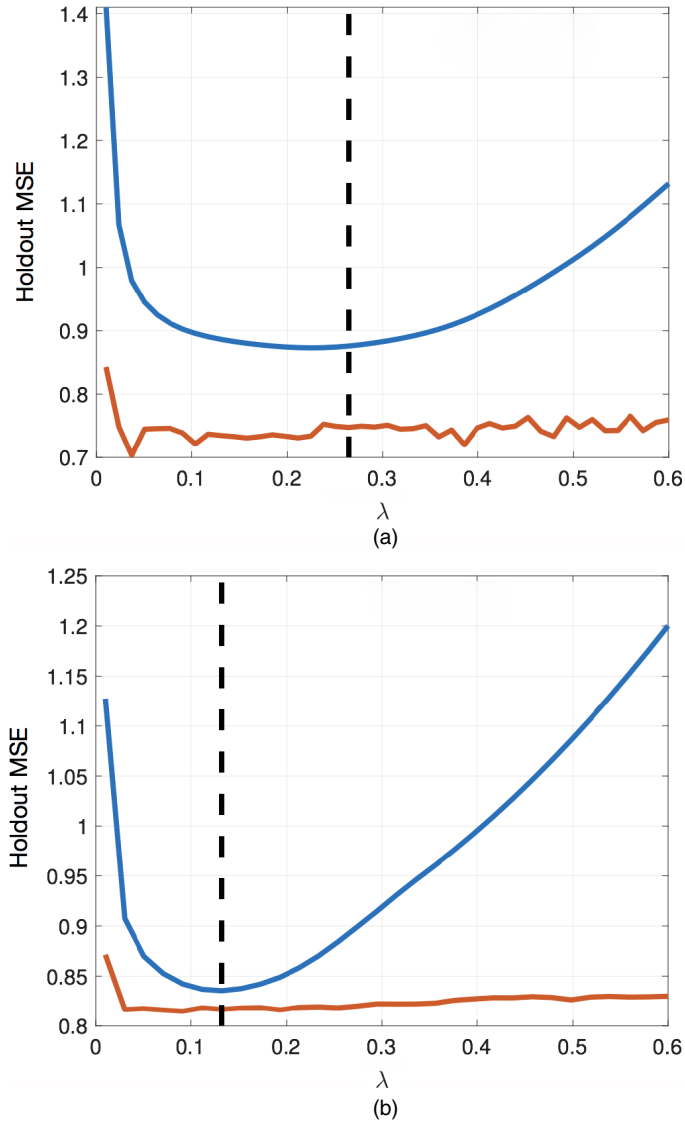


Fig. 6. Collaborative filtering—MSE on the hold-out set of the non-subsampled approach (——) and subspace stability selection (——) (⁞, cross-validated choice of λ with the non-subsampled approach): (a) Amazon video games; (b) Amazon book crossing

approach is 80 (i.e. the maximum allowable rank) with the first three singular values equal to 4329, 135.4 and 63.1. The MSE of this model on the evaluation set is 0.83. In contrast, at the cross-validated λ-subspace stability selection yields a rank 2 model with an MSE of 0.81 on the evaluation set. Thus, we obtain a much simpler model with subspace stability selection that also offers better predictive performance. Similarly, for the Amazon video games data set, the rank of the estimate that is obtained from the non-subsampled approach is 39 with the first five singular values equal to 1913.5, 49.4, 43.6, 28.4 and 27.4, with an MSE of 0.87 on the evaluation set. In contrast, subspace stability selection yields a rank 4 solution with a much smaller MSE of 0.74 on the evaluation set. Finally, we observe for both data sets that subspace stability

selection is much more stable across the range of regularization parameters. Thus, subspace stability selection is far less sensitive to the particular choice of λ, which removes the need for fine tuning λ.

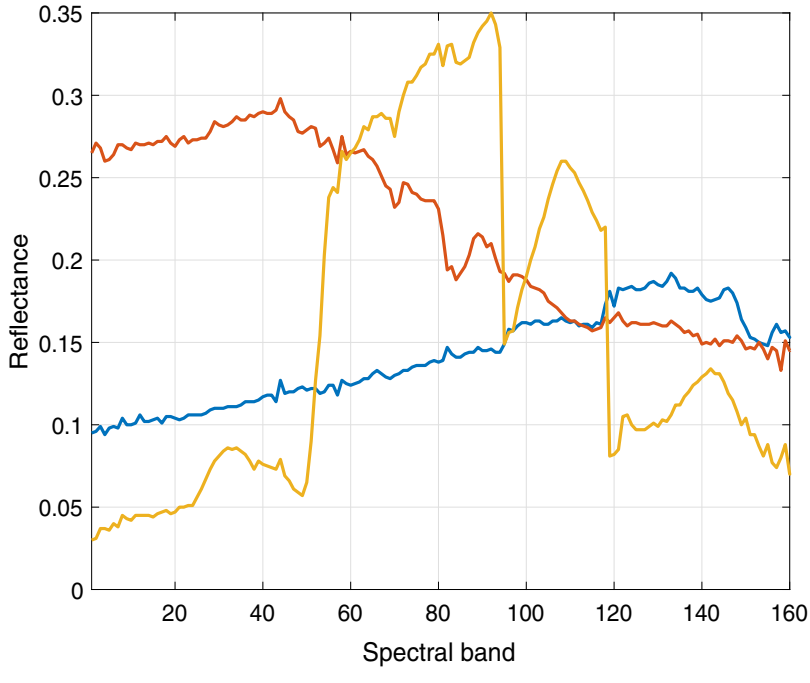
4.4.2. *Hyperspectral unmixing*

Here we give an illustration with real hyperspectral imaging data in which the underlying population parameters are known on the basis of extensive prior experiments. In this problem, we are given a hyperspectral image $Y \in \mathbb{R}^{p_1 \times p_2}$ consisting of p_1 frequency bands and p_2 pixels, where $Y_{i,j}$ is the reflectance of the jth image pixel to the ith frequency band. The spectral unmixing problem aims to find $W \in \mathbb{R}^{p_1 \times k}$ (called the end member matrix) and $H \in \mathbb{R}^{k \times p_2}$ (called the abundance matrix) so that $Y \approx WH$, where $k \ll \min(p_1, p_2)$ is the number of end members (Manolakis, 2003). Of particular interest is the k-dimensional column space of W, which corresponds to the space that is spanned by the k end members that are present in the image. We discuss two natural hyperspectral unmixing problems that arise commonly in practice. We focus on the urban data set (obtained from `http://www.escience.cn/people/feiyunZHU/Dataset_GT.html`): a hyperspectral image consisting of 307×307 pixels, each of which corresponds to a $2\,\text{m} \times 2\,\text{m}$ area with 210 wavelengths ranging from 400 nm to 2500 nm. Following previous analyses of this data set, we remove 48 noisy channels to obtain 162 wavelengths and select a 30×25 patch (equal to 750 pixels) that is shown in Fig. 7(a). In the patch selected, there are a total of three end members (shown in Fig. 7(b)), with one strong signal and two weak signals.

In many settings, obtaining a complete hyperspectral image of a scene may be costly, and it is of interest to reconstruct a hyperspectral image accurately from partial observations. This problem may be naturally formulated as one of low rank matrix completion. As with other application domains in which problems are reformulated as low rank matrix completion, ALS applied to the non-convex formulation (4.1) is commonly employed. To simulate such a hyperspectral unmixing problem, we randomly subsampled 10% of the hyperspectral data in the patch as training data. We further selected another 10% of the remaining data as a hold-out validation set. We compare the amount of false discovery of a non-subsampled approach and subspace stability approach, with k conservatively chosen to be equal to 20 in the ALS procedure in each case. Because the scale of this problem is large, we use the modification of algorithm 1 (with $\alpha = 0.7$ and $B = 100$ complementary bags) described in Section 3.1 for subspace stability selection. As the column space of the low rank estimate is the principal object of interest for end member detection, the quantities of interest for evaluating performance are based on expression (2.5): $\overline{\text{FD}} = \mathbb{E}[\text{tr}(\mathcal{P}_{\text{col-space}(W^*)^\perp}\mathcal{P}_{\text{col-space}(\hat{W})})]$ and $\overline{\text{PW}} = \mathbb{E}[\text{tr}(\mathcal{P}_{\text{col-space}(W^*)}\mathcal{P}_{\text{col-space}(\hat{W})})]$. Here, the expectation is with respect to the randomness in the selection of the 10% training data, $W^* \in \mathbb{R}^{162 \times 3}$ is the matrix consisting of the spectra of the three end members in Fig. 7(b)) and $\hat{W}$ is the estimated matrix. We find a cross-validated choice of $\lambda = 1$ from one random selection of training data. With this λ and over 100 random trials in the selection of training data, non-subsampled ALS produces on average rank 20 estimate with $\overline{\text{FD}} = 0.1 \dim\{\text{col-space}(W^{*\perp})\}$ and $\overline{\text{PW}} = 0.97 \dim(\text{col-space}(W^*))$. In contrast, for the same $\lambda = 1$, subspace stability selection (operating on tangent spaces $T_n(\text{col-space}(\hat{W}))$ produces on average rank 2.86 with $\overline{\text{FD}} = 0.0007 \dim(\text{col-space}\{W^{*\perp})\}$ and $\overline{\text{PW}} = 0.91 \dim\{\text{col-space}(W^*)\}$. Furthermore, even if λ is set sufficiently large (e.g. $\lambda = 29$) so that the non-subsampled ALS estimate has on average rank equal to 2.52, the false discovery estimate is $\overline{\text{FD}} = 0.007 \dim\{\text{col-space}(W^*)^\perp\}$, which is still far larger than the amount of false discovery of subspace stability selection.

A different type of hyperspectral unmixing problem arises if the observations are corrupted by noise. In particular, based on the decomposition $Y \approx WH$, the outer product YY' is well

(a)



(b)

Fig. 7. (a) Urban hyperspectral image and (b) spectra of three materials in the image (the data and the population spectra are obtained from http://www.escience.cn/people/feiyunZHU/Dataset_GT.html): ——, asphalt; ——, root; ——, grass

approximated by a low rank matrix. Thus, another natural approach for end member detection is to perform factor analysis by viewing each column of Y (i.e. an entire collection of wavelengths corresponding to each pixel) as an observation and approximating the sample covariance of these observations as the sum of diagonal and low rank matrices. The row–column spaces of the low rank component (which is symmetric and hence the row and column spaces are the same) serve as estimates of the subspace that is spanned by the end members. We obtain $\{Y^{(i)}\}_{i=1}^{750} \subset \mathbb{R}^{162}$ spectral observations of the 750 total pixels by applying white noise to the population parameters with the noise level chosen so that $\mathrm{SNR} = 0.78$. We then set aside 80% of the data as training data for estimator (4.2), which is solved by using the LogDetPPA solver of Toh *et al.* (2006). We set aside the remaining 20% as a hold-out validation set. Employing estimator (4.2) without subsampling and with λ chosen via cross-validation and expectations computed over 100 yields false discovery $\mathrm{FD} = 0.04 \dim(T^{*\perp})$ and power $\mathrm{PW} = 0.48 \dim(T^*)$. (Here T^* represents the population tangent space.) In contrast, subspace stability selection with $\alpha = 0.7$ and $B = 100$ complementary bags yields a tangent space estimate with a false discovery $\mathrm{FD} = 0.015 \dim(T^{*\perp})$ and power $\mathrm{PW} = 0.69 \dim(T^*)$. Evidently, subspace stability selection yields a substantial decrease in the amount of false discovery as well as an improvement in power.

5. Conclusions and future directions

In this paper, we describe a geometric framework for assessing false discoveries in low rank estimation. The framework proposed has many appealing properties including that it is a natural generalization of false discovery in variable selection. We further describe the subspace stability selection algorithm to provide false discovery control in the low rank setting. This procedure is a generalization of the stability selection method of Meinshausen and Bühlmann (2010). The method is general and we demonstrate its utility with both synthetic and real data sets in a range of low rank estimation tasks.

There are several interesting directions for further investigation that arise from our work. First, within the context of theorem 1 on the expected false discovery of a stable tangent space produced by subspace stability selection, it would be useful to carry out a more refined bag-dependent analysis in the spirit of Shah and Samworth (2013). Second, while algorithm 1 from Section 3.3 outputs an estimate that does provide false discovery control, it is unclear whether this is the most powerful procedure possible. In particular, it is of interest to obtain an optimal solution to problem (3.9), or to prove that algorithm 1 computes a nearly optimal solution. Third, algorithm 1 requires a user-specified α to produce an estimate that provides a false discovery bound as stated in theorem 1. In exploratory settings, one may wish to examine the data first, and to choose α to obtain a desired amount of discovery while still retaining some false discovery guarantees. This viewpoint, considered by Goeman and Solari (2011), reverses the traditional role of the analyst and the inference procedure. Building on their perspective, it would be of interest to develop false discovery bounds for subspace stability selection that remain valid despite *post hoc* selection of α. Fourth, a significant topic of contemporary interest in variable selection—especially when there are a large number of possible predictors—is to control for the false discovery rate. In Section 2 we gave a formulation of the false discovery rate in the low rank setting, and it is natural to seek procedures that provide false discovery rate control in settings with high dimensional matrices. One obstacle that arises with this effort is that every proof of false discovery rate control of a variable selection method (of which we are aware) relies strongly on the simultaneous diagonalizability of the projection matrices that are associated with the population tangent space and the estimated tangent space (when translated

to the geometric viewpoint of our paper). Finally, the geometric framework that is developed in this paper for assessing false discovery is potentially relevant beyond the specific setting of low rank estimation. For example, our set-up extends naturally to latent variable graphical model selection (Chandrasekaran *et al.*, 2012) as well as low rank tensor estimation (Kolda and Bader, 2009), both of which are settings in which the underlying geometry is similar to that of low rank estimation. More broadly, the perspective that is presented here may be useful in addressing many other structured estimation problems.

Acknowledgements

This research was funded by National Science Foundation grant CCF-1350590, Air Force Office of Scientific Research grant FA9550-16-1-0210 and Sloan and Resnick Fellowships.

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Björck, A. and Golub, G. (1973) Numerical methods for computing angles between linear subspaces. *Math. Computns*, **27**, 579–594.

Candès, E. and Recht, B. (2009) Exact matrix completion via convex optimization. *Foundns Computnl Math.*, **9**, 717–772.

Chandrasekaran, V., Parillo, P. A. and Willsky, A. S. (2012) Latent variable graphical model selection via convex optimization. *Ann. Statist.*, **40**, 1935–1967.

Choi, Y., Taylor, J. and Tibshirani, R. (2017) Selecting the number of principal components: estimation of the true rank of a noisy matrix. *Ann. Statist.*, **45**, 2590–2617.

Fa, R. and Lamare, R. (2011) Reduced-rank STAP algorithms using joint iterative optimization of filters. *IEEE Trans. Aer. Electron. Syst.*, **47**, 1668–1684.

Fazel, M. (2002) Matrix rank minimization with applications. *PhD Thesis*. Department of Electrical Engineering, Stanford University, Stanford.

Goldberg, D., Nichols, D., Oki, B. and Terry, D. (1992) Using collaborative filtering to weave an information tapestry. *Communs ACM*, **35**, no. 12, 61–70.

Harris, J. (1995) *Algebraic Geometry: a First Course*. Berlin: Springer.

Kolda, T. and Bader, B. (2009) Tensor decompositions and applications. *SIAM Rev.*, **51**, 455–500.

Liu, Z. and Lin, X. (2018) A geometric perspective on the power of principal component association tests in multiple phenotype studies. *J. Am. Statist. Ass.*, **114**, 975–990.

Liu, Z. and Vandenberghe, L. (2009) Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrx Anal. Appl.*, **31**, 1235–1256.

Manolakis, D. (2003) Detection algorithms for hyperspectral imaging applications: a signal processing perspective. In *Proc. Wrkshp Advances in Techniques for Analysis of Remotely Sensed Data*, pp. 378–384.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.

Pati, Y. and Kailath, T. (1994) Phase-shifting masks for microlithography: automated design and mask requirements. *J. Opt. Soc. Am.* A, **11**, 2438–2452.

Shah, R. D. and Samworth, J. (2013) Variable selection and error control: another look at stability selection. *J. R. Statist. Soc.* B, **75**, 55–80.

Shapiro, A. (1982) Weighted minimum trace factor analysis. *Psychometrika*, **47**, 243–264.

Song, J. and Shin, S. (2018) Stability approach to selecting the number of principal components. *Computnl Statist.*, **33**, 1923–1938.

Srebro, N. and Shraibman, A. (2005) Rank, trace-norm and max-norm. In *Proc. 18th A. Conf. Learning Theory*, pp. 545–560.

Toh, K. C., Todd, M. J. and Tutuncu, R. H. (2006) SDPT3—a MATLAB software package for semidefinite-quadratic-linear programming. (Available from `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`.)

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material: False discovery and its control in low rank estimation'.