False Discovery and Its Control in Low-Rank Estimation

Armeen Taeb ^a, Parikshit Shah ^b, and Venkat Chandrasekaran ^{a,c *}

^a Department of Electrical Engineering, California Institute of Technology

^b Yahoo Research and Wisconsin Institutes for Discovery at the University of Wisconsin

^c Department of Computing and Mathematical Sciences, California Institute of Technology

Abstract

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 to variable selection, we lack a formal framework to assess true/false discoveries in low-rank estimation; in particular, the key source of difficulty is that the standard notion of a discovery is a discrete one 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 analyze 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 to previous approaches via numerical experiments.

keywords: algebraic geometry, determinantal varieties, testing, model selection, regularization, stability selection

1 Introduction

Models 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. Similarly, the row/column spaces of matrices 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/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 as 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 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

*Correspondence email: ataeb@caltech.edu

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 perfectly estimate 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 coordinate subspaces that correspond to discoveries in variable selection. Therefore, the number of true discoveries would generically be zero. One method to improve upon 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/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/false discoveries in variable selection in terms of functionals of the projection matrices associated to the discovery/population subspaces described above; this expression varies smoothly with respect to the underlying subspaces, unlike dimensions of intersections of subspaces. Next, we interpret the discovery/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 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/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/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 dataset, and then declares as discoveries those variables that are selected most frequently. In analogy to 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 dataset, 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; *Shah and Samworth*, 2013), we provide a theoretical analysis of the performance of our algorithm.

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 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 to 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 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 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 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 3.2. 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.

Related work We are aware of prior work for low-rank estimation based on testing the significance level of the singular values of an observed matrix (see, for example, *Choi et al (2017)*, *Liu and Lin (2018)*, *Song and Shin (2018)*). However, in contrast to 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 the case, 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*; *Shah and Samworth, 2013*). However, the geometry underlying the collection of tangent spaces to low-rank matrices leads to a number of new challenges in our context.

Notation The Euclidean or ℓ_2 ball is denoted by $\mathcal{S}_{\ell_2}$, with the dimension of the space being clear from the context. For a subspace $\mathbb{V}$, we denote projection onto $\mathbb{V}$ by $\mathcal{P}_{\mathbb{V}}$. Given a self-adjoint linear map $M : \bar{\mathbb{V}} \rightarrow \bar{\mathbb{V}}$ 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 $\sigma_{\min}(\mathcal{P}_{\mathbb{V}}M\mathcal{P}_{\mathbb{V}}) = \inf_{x \in \mathbb{V} \setminus \{0\}} \frac{\|Mx\|_{\ell_2}}{\|x\|_{\ell_2}}$.

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/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 $\hat{\mathcal{S}} \subset \{1, \dots, p\}$, 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 present in the ‘true’ subset of significant variables $\mathcal{S}^* \subset \{1, \dots, p\}$ – the number of true discoveries is $|\hat{\mathcal{S}} \cap \mathcal{S}^*|$, while the number of false discoveries is $|\hat{\mathcal{S}} \cap \mathcal{S}^{*c}|$. We give next a geometric perspective on this combinatorial notion. As described in the introduction, one can associate to each subset $\mathcal{S} \subset \{1, \dots, p\}$ the coordinate aligned *subspace* $T(\mathcal{S}) = \{x \in \mathbb{R}^p \mid \text{support}(x) \subseteq \mathcal{S}\}$, where $\text{support}(x)$ denotes the locations of the nonzero entries of x . With this notation, the number of false discoveries in an estimate $\hat{\mathcal{S}}$ is given by:

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

Similarly, the number of true discoveries is given by $\text{trace} \left(\mathcal{P}_{T(\hat{\mathcal{S}})} \mathcal{P}_{T(\mathcal{S}^*)} \right)$. These latter reformulations in terms of projection operators have no obvious ‘discrete’ attribute to them. In particular, for any subspaces $\mathcal{W}, \tilde{\mathcal{W}}$, the expression $\text{trace}(\mathcal{P}_{\mathcal{W}} \mathcal{P}_{\tilde{\mathcal{W}}})$ is equal to the sum of the squares of the cosines of the principal angles between $\mathcal{W}$ and $\tilde{\mathcal{W}}$ (*Bhatia, 2013*); as a result, the quantity $\text{trace}(\mathcal{P}_{\mathcal{W}} \mathcal{P}_{\tilde{\mathcal{W}}})$ varies smoothly with respect to perturbations of $\mathcal{W}, \tilde{\mathcal{W}}$. The discrete nature of a discovery is embedded inside the encoding of the subsets $\hat{\mathcal{S}}, \mathcal{S}^*$ using the subspaces $T(\hat{\mathcal{S}}), T(\mathcal{S}^*)$. Consequently, to make progress towards a suitable definition of true/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 $T(\mathcal{S})$ associated to a subset $\mathcal{S} \subset \{1, \dots, p\}$ as a tangent space to an algebraic variety.

Formally, for any integer $k \in \{1, \dots, p\}$ let $\mathcal{V}_{\text{sparse}}(k) \subset \mathbb{R}^p$ denote the algebraic variety of elements of $\mathbb{R}^p$ with at most k nonzero entries. Then for any point in $\mathcal{V}_{\text{sparse}}(k)$ consisting of exactly k nonzero entries at locations given by the subset $\mathcal{S} \subset \{1, \dots, p\}$ (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 nonzero 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 $r \in \{1, \dots, \min(p_1, p_2)\}$). 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:

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

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:

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

In analogy to 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/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 matrix in $\mathbb{R}^{p_1 \times p_2}$; in particular, $\dim(\mathcal{C}^*) = \dim(\mathcal{R}^*)$. Given observations of 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:

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

and the power of the estimator is defined as:

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

The expectations in both cases are with respect to randomness in the data employed by the estimator, and the tangent spaces $T(\hat{\mathcal{C}}, \hat{\mathcal{R}}), T(\mathcal{C}^*, \mathcal{R}^*)$ are as defined in (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 possess a number of favorable attributes. These definitions do not depend on a choice of basis for the tangent space $T(\mathcal{C}^*, \mathcal{R}^*)$. Further, for the reasons 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. Despite these definitions respecting the smooth structure underlying low-rank matrices, they specialize transparently to the usual discrete notion of true/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 $0 \leq \text{FD} \leq \dim(T(\mathcal{C}^*, \mathcal{R}^*)^\perp)$ and the power is bounded as $0 \leq \text{PW} \leq \dim(T(\mathcal{C}^*, \mathcal{R}^*))$, 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 $\text{FD} + \text{PW} = \mathbb{E}[\dim(T(\hat{\mathcal{C}}, \hat{\mathcal{R}}))]$, which is akin to the expected total discovery made by the estimator $(\hat{\mathcal{C}}, \hat{\mathcal{R}})$.

As a final remark, we note that the definition of FD may be modified to obtain an analog of the *false discovery rate* (Benjamini and Hockberg, 1995), which is of interest in contemporary multiple testing as well as in high-dimensional estimation:

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

We focus in the present paper on procedures that control the quantity FD by generalizing the stability selection approach of *Meinshausen and Bühlmann* (2010), and we discuss in Section 5 on potential future directions some challenges associated with controlling FDR in low-rank estimation.

2.2 From a Tangent Space to a Low-Rank Matrix

While the focus 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 datasets in Section 4) the ultimate object of interest is an estimate of a low-rank matrix. We present next a simple approach to estimate a low-rank matrix given an estimate of a tangent space. Specifically, letting $T(\mathcal{C}, \mathcal{R}) \subset \mathbb{R}^{p_1 \times p_2}$ be a tangent space that corresponds to column and row spaces $\mathcal{C} \subset \mathbb{R}^{p_1}, \mathcal{R} \subset \mathbb{R}^{p_2}$, and given a collection of observations $\mathcal{D}$, we wish to solve the following optimization problem:

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2}}{\operatorname{argmin}} \operatorname{Loss}(L; \mathcal{D}) \quad \text{subject to} \quad T(\text{column-space}(L), \text{row-space}(L)) \subseteq T(\mathcal{C}, \mathcal{R}) \quad (2.5)$$

in which the decision variable L is constrained to have a tangent space that lies within the prescribed tangent space $T(\mathcal{C}, \mathcal{R})$. This constraint may be simplified as follows. Suppose that the subspaces $\mathcal{R}, \mathcal{C}$ 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 one 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, the problem (2.5) may be reformulated as follows:

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2}, M \in \mathbb{R}^{k \times k}}{\operatorname{argmin}} \operatorname{Loss}(L; \mathcal{D}) \quad \text{subject to} \quad L = U_C M U_R'. \quad (2.6)$$

Note that the constraint here is linear in the decision variables L, M . Consequently, an appealing property of (2.6) is that if the loss function $\operatorname{Loss}(\cdot; \mathcal{D})$ is convex, then (2.6) is a convex optimization problem. For example, when $\operatorname{Loss}(\cdot; \mathcal{D})$ is the squared loss, an optimal solution can be obtained in closed form.

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 associated to 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.

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, one instead applies the procedure to many subsamples of the data and then chooses 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.

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

Setup 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 $1(x3), 0.5(x5)$, and $0.1(x2)$, 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 $\Omega \subset \{1, \dots, 70\}^2$ is chosen uniformly at random with $|\Omega| = 3186$. The variance σ^2 is chosen to set the signal-to-noise (SNR) ratio 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 following convex program (Candés and Recht, 2009; Srebro and Shraibman, 2005)

$$\hat{L} = \underset{L \in \mathbb{R}^{70 \times 70}}{\operatorname{argmin}} \sum_{\{i,j\} \in S} \|(L - Y)_{i,j}\|_F^2 + \lambda \|L\|_*, \quad (3.1)$$

and we report the tangent space $T(\text{column-space}(\hat{L}), \text{row-space}(\hat{L}))$ as the estimate associated to 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 for general low-rank estimation problems, and this problem is merely for illustrative purposes in the present section; for a more comprehensive set of experiments in more general settings, we refer the reader to Section 4.

3.1 Stable Tangent Spaces

The first step in stability selection is to combine estimates of significant variables 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 $\hat{T}$ 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 $\{\mathcal{D}_i\}_{i=1}^B$ with each $\mathcal{D}_i \subset \mathcal{D}$. Then the average projection operator of the estimator $\hat{T}$ with respect to the subsamples $\{\mathcal{D}_i\}_{i=1}^B$ is defined as:*

$$\mathcal{P}_{\text{avg}} \triangleq \frac{1}{B} \sum_{i=1}^B \mathcal{P}_{\hat{T}(\mathcal{D}_i)}, \quad (3.2)$$

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

The operator $\mathcal{P}_{\text{avg}} : \mathbb{R}^{p_1 \times p_2} \rightarrow \mathbb{R}^{p_1 \times p_2}$ is self-adjoint, and its eigenvalues lie in the interval $[0, 1]$ as each $\mathcal{P}_{\hat{T}(\mathcal{D}_i)}$ 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 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.

Illustration: the value of averaging projection maps – Consider the stylized low-rank matrix completion problem described at the beginning of Section 3. To support the intuition that the average projection matrix $\mathcal{P}_{\text{avg}}$ has reduced in energy in directions corresponding to $T^{\star\perp}$ (i.e., the orthogonal complement of the population tangent space), we compare the quantities $\mathbb{E}[\text{trace}(\mathcal{P}_{\text{avg}} \mathcal{P}_{T^{\star\perp}})]$ and $\mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D})} \mathcal{P}_{T^{\star\perp}})]$, where the expectation is computed over 10 instances. The first of these quantities measures the average false discovery over the subsampled models, while the second 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 Figure 1, $\mathbb{E}[\text{trace}(\mathcal{P}_{\text{avg}} \mathcal{P}_{T^{\star\perp}})]$ is smaller than $\mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D})} \mathcal{P}_{T^{\star\perp}})]$ 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 spanned by $T^{\star\perp}$.

While the average projection aggregated over many subsamples appears to be better ‘aligned’ with $T^{\star\perp}$, 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 $\hat{T}$ 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 $\{\mathcal{D}_i\}_{i=1}^B$ with each $\mathcal{D}_i \subset \mathcal{D}$. For a parameter $\alpha \in (0, 1)$, the set of stable tangent spaces is defined as*

$$\mathcal{T}_\alpha \triangleq \left\{ T \mid \sigma_{\min}(\mathcal{P}_T \mathcal{P}_{\text{avg}} \mathcal{P}_T) \geq \alpha \text{ and } T \text{ is a tangent space to a determinantal variety} \right\} \quad (3.3)$$

Figure 1: The quantities $\mathbb{E} [\text{trace} (\mathcal{P}_{\hat{T}(\mathcal{D})} \mathcal{P}_{T^{\star\perp}})]$ (in blue) and $\mathbb{E} [\text{trace} (\mathcal{P}_{\text{avg}} \mathcal{P}_{T^{\star\perp}})]$ (in red) as a function of λ for $\text{SNR} = 1.6$ and $\text{SNR} = 0.8$ in the synthetic matrix completion setup. The cross-validated choice of λ is shown as the dotted black line. Here ‘N-S’ denotes no subsampling and ‘W-S’ denotes with subsampling.

where $\mathcal{P}_{\text{avg}}$ is computed based on 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 nonzero 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 simultaneously diagonalizable). As a consequence, the set $\mathcal{T}_\alpha$ has a straightforward characterization for variable selection problems; it consists of subspaces 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 obtained from the subsampled data reduces energy in the directions contained in $T^{\star\perp}$, each element of $\mathcal{T}_\alpha$ is also far from being closely aligned with $T^{\star\perp}$ (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 3.3 we describe efficient methods to choose an element of $\mathcal{T}_\alpha$.

3.2 False Discovery Control of Stable Tangent Spaces: Theoretical Analysis

Setup: Suppose we have a population matrix parameter L^* that is low-rank with associated tangent space T^* . We are given i.i.d observations from a model parametrized by L^* and the objective is to obtain an accurate estimate of T^* . We intentionally keep our discussion broad so our results are relevant for a wide range of low-rank estimation problems, e.g., low-rank matrix completion, factor analysis, etc. Let $\hat{T}$ denote a tangent space estimator that operates on samples drawn from the model parametrized by L^* . Let $\mathcal{D}(n)$ denote a dataset consisting of n i.i.d observations from this models; we assume that n is even and that we are given B subsamples $\{\mathcal{D}_i\}_{i=1}^B$ via complementary partitions of $\mathcal{D}(n)$.

Theorem 4 (False Discovery Control of Subspace Stability Selection). *Consider the setup described above. Let $\hat{T}(\mathcal{D}_i)$ denote the tangent space estimates obtained from each of the subsamples, and let $\mathcal{P}_{\text{avg}}$ denote the associated average projection operator computed via (3.2). Fix any $\alpha \in (0, 1)$ and let T denote any selection of an element of the associated set $\mathcal{T}_\alpha$ of stable tangent spaces. Then we have that*

$$\mathbb{E} [\text{trace} (\mathcal{P}_T \mathcal{P}_{T^{\star\perp}})] \leq \mathbb{E} \left[\left\| \mathcal{P}_{\hat{T}(\mathcal{D}(n/2))} \mathcal{P}_{T^{\star\perp}} \right\|_{\star}^{1/2} \right]^2 + 2 \dim(T^{\star\perp}) \sqrt{1 - \alpha}. \quad (3.4)$$

Here the expectation is with respect to randomness in the observations. The set $\mathcal{D}(n/2)$ denotes a collection of $n/2$ i.i.d. observations drawn from the model parametrized by L^* .

Before giving the proof of this result, a number of remarks are in order.

Remark 1: Theorem 4 states that the expected false discovery of a stable tangent space is bounded by two quantities. The first term $\mathbb{E} \left[\left\| \mathcal{P}_{T^* \perp} \mathcal{P}_{\hat{T}(\mathcal{D}(n/2))} \right\|_{\star}^{1/2} \right]^2$ characterizes the quality of the estimator $\hat{T}$ on subsamples consisting of $n/2$ observations; specifically, this term is equal to $\mathbb{E} \left[\sqrt{\sum_i \cos(\theta_i)} \right]^2$ where θ_i are the principal angles between $T^* \perp$ and $\hat{T}(\mathcal{D}(n/2))$ (Bhatia, 2013). The second quantity in (3.4) is due to the user specified parameter α , and as expected, choosing α closer to 1 leads to a smaller amount of false discovery.

Remark 2: As with the results of Meinshausen and Bühlmann (2010), Theorem 4 is agnostic to the number of bags that are used in the subsampling procedure. Nevertheless, one can obtain the following bag-dependent bound based on an intermediate step of the proof of Theorem 4 presented in the sequel:

$$\begin{aligned} \mathbb{E} [\text{trace}(\mathcal{P}_T \mathcal{P}_{T^* \perp})] &\leq \mathbb{E} \left[\min_{i=1,2,\dots,B/2} \prod_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^* \perp} \right\|_{\star}^{1/2} \right. \\ &\quad \left. + \sum_{k \in \{0,1\}} \sqrt{\dim(T^* \perp)} \sqrt{\text{trace}(\mathcal{P}_{T^* \perp} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k}) \perp} \mathcal{P}_T \mathcal{P}_{T^* \perp})} \right] \end{aligned} \quad (3.5)$$

It is straightforward to check that when (3.5) reduces to (3.4) when $B = 2$ with some algebraic manipulations. However, for $B > 2$, (3.5) may produce a tighter bound on the expected false discovery. It is of interest to consider additional assumptions on the population model (similar in spirit to the developments in Shah and Samworth (2013)) to produce a more refined bag-dependent bound.

Remark 3: The bound in (3.4) is also valid in the setting of variable selection. However, by exploiting the fact that projection matrices of tangent spaces to varieties of sparse vectors are simultaneously diagonalizable, one is able to choose a basis (the standard basis) that simultaneously diagonalizes all these matrices, which leads to certain simplifications as well as an eventual tighter bound on the expected false discovery in variable selection. In particular, letting $\{w_j\}_{j=1}^{\dim(T^* \perp)}$ be the vectors of the standard basis that belong to $T^* \perp$, one can modify the proof of Theorem 4 to obtain the following bound:

$$\begin{aligned} \mathbb{E} [\text{trace}(\mathcal{P}_T \mathcal{P}_{T^* \perp})] &\leq \frac{\sum_{j=1}^{\dim(T^* \perp)} \mathbb{E} \left[\left\| \mathcal{P}_{\text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}(n/2))} \right\|_{\star}^{1/2} \right]^2}{2\alpha - 1} \\ &= \frac{\sum_{j=1}^{\dim(T^* \perp)} \text{Prob}[\text{null variable corresponding to } w_j \text{ selected by } \hat{T}(\mathcal{D}(n/2))]^2}{2\alpha - 1}. \end{aligned} \quad (3.6)$$

The proof of the bound in the first line is given in the supplementary material. The second line here follows from the observations that $\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}$ is a diagonal projection matrix and that $\mathcal{P}_{\text{span}(w_j)}$ is also diagonal with only one non-zero element on the diagonal. Thus, the simultaneous diagonalizability property also leads to the conceptually appealing 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 (3.6) matches precisely one of the main results of Meinshausen and Bühlmann (2010), although that work also goes further in its analysis by imposing additional assumptions on the underlying model from which the data are generated in order to obtain bounds on the term $\mathbb{E} \left[\left\| \mathcal{P}_{\text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}(n/2))} \right\|_{\star}^{1/2} \right]$. As a final comparison between Theorem 4 for low-rank estimation and the bound (3.6) for variable selection, note that the dependence on α in (3.6) is multiplicative as opposed to additive as in (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 α is close to 1. This distinction is fundamental to the geometry underlying the sparse and determinantal varieties. Specifically, in the low-rank case even if $\mathcal{P}_{\text{avg}} \approx \mathcal{P}_{T^*}$ 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 due to the fact that the collection of row/column spaces forms a Grassmannian manifold rather than a finite/discrete set. On the other hand, if $\mathcal{P}_{\text{avg}} \approx \mathcal{P}_{T^*}$ in variable selection, the only elements of the set of stable tangent spaces (for suitable α) are those corresponding to subsets of the true significant variables.

Proof of Theorem 1 Notice that for any tangent space $\hat{T}(\mathcal{D}_j)$ estimated on a subsample, we have that

$$\begin{aligned} \text{trace}(\mathcal{P}_T \mathcal{P}_{T^{\star\perp}}) &= \text{trace}(\mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_j)} \mathcal{P}_{T^{\star\perp}}) + \text{trace}(\mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp} \mathcal{P}_{T^{\star\perp}}) \\ &\leq \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_j)} \mathcal{P}_{T^{\star\perp}} \right\|_\star + \left\| \mathcal{P}_{T^{\star\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp} \right\|_\star \end{aligned} \quad (3.7)$$

This inequality follows from the property that $\text{trace}(AB) \leq \|A\|_2 \|B\|_\star$, the cyclicity of the trace functional, and the idempotence of projection maps. As $\mathcal{D}_j$ was arbitrary, we can minimize over the entire collection as follows:

$$\begin{aligned} \text{trace}(\mathcal{P}_T \mathcal{P}_{T^{\star\perp}}) &\stackrel{(a)}{\leq} \min_{i=1,2,\dots,B/2} \min_{k=\{0,1\}} \left\{ \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + \left\| \mathcal{P}_{T^{\star\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp} \right\|_\star \right\} \\ &\stackrel{(b)}{\leq} \min_{i=1,2,\dots,B/2} \min_{k=\{0,1\}} \left\{ \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + \sqrt{\dim(T^{\star\perp})} \left\| \mathcal{P}_{T^{\star\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp} \right\|_F \right\} \\ &\stackrel{(c)}{=} \min_{i=1,2,\dots,B/2} \min_{k=\{0,1\}} \left\{ \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + \sqrt{\dim(T^{\star\perp})} \sqrt{\text{trace}(\mathcal{P}_{T^{\star\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp} \mathcal{P}_T \mathcal{P}_{T^{\star\perp}})} \right\} \\ &\stackrel{(d)}{\leq} \min_{i=1,2,\dots,B/2} \min_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + \sum_{k \in \{0,1\}} \sqrt{\dim(T^{\star\perp})} \sqrt{\text{trace}(\mathcal{P}_{T^{\star\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp} \mathcal{P}_T \mathcal{P}_{T^{\star\perp}})} \\ &\stackrel{(e)}{\leq} \frac{2}{B} \sum_{i=1}^{B/2} \min_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + \frac{2\sqrt{\dim(T^{\star\perp})}}{B} \sum_{i=1}^B \sqrt{\text{trace}(\mathcal{P}_{T^{\star\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_i)^\perp} \mathcal{P}_T \mathcal{P}_{T^{\star\perp}})} \\ &\stackrel{(f)}{\leq} \frac{2}{B} \sum_{i=1}^{B/2} \min_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + 2 \sqrt{\dim(T^{\star\perp})} \sqrt{\text{trace}(\mathcal{P}_{T^{\star\perp}} \mathcal{P}_T (\mathcal{I} - \mathcal{P}_{\text{avg}}) \mathcal{P}_T \mathcal{P}_{T^{\star\perp}})} \\ &\stackrel{(g)}{\leq} \frac{2}{B} \sum_{i=1}^{B/2} \min_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + 2 \dim(T^{\star\perp}) \sqrt{\|\mathcal{P}_T (\mathcal{I} - \mathcal{P}_{\text{avg}}) \mathcal{P}_T\|_2} \\ &\stackrel{(h)}{\leq} \frac{2}{B} \sum_{i=1}^{B/2} \min_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star + 2 \dim(T^{\star\perp}) \sqrt{1 - \alpha} \\ &\stackrel{(h)}{\leq} \frac{2}{B} \sum_{i=1}^{B/2} \left[\prod_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{T^{\star\perp}} \right\|_\star^{1/2} \right] + 2 \dim(T^{\star\perp}) \sqrt{1 - \alpha}. \end{aligned}$$

Here (a) follows from the fact that the choice of subsample in (3.7) is arbitrary; (b) from the inequality $\|A\|_\star \leq \|A\|_F \sqrt{\text{rank}(A)}$; (c) from the idempotence of projection maps; (d) from the property $\min\{a+b, c+d\} \leq \min\{a, c\} + b + d$ for $a, b, c, d \geq 0$; (e) from the fact that the minimum over a collection is bounded above by the average of the collection; (f) from the concavity of the square root; (g) from the cyclicity of the trace functional, the idempotence of projection maps, and the inequality $\text{trace}(AB) \leq \|A\|_2 \|B\|_\star$; (h) due to $T \in \mathcal{T}_\alpha$ as defined in Definition 3; and finally (i) from the fact that the minimum of two positive quantities is bounded above the product of their square roots. Taking expectations on both sides, observing that the quantities $\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-1})} \mathcal{P}_{T^{\star\perp}}\|_\star^{1/2}$ and $\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2i})} \mathcal{P}_{T^{\star\perp}}\|_\star^{1/2}$ are statistically independent due to complementary partitioning, and noting that $\|\mathcal{P}_{\hat{T}(\mathcal{D}_i)} \mathcal{P}_{T^{\star\perp}}\|_\star^{1/2}$ is identically distributed for all i gives us the desired result. $\square$

3.3 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 $r = 1, \dots, \min\{p_1, p_2\}$:

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

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

$$T_{\text{OPT}} \in \underset{T \in T_{\text{OPT}}(r) \cap \mathcal{T}_\alpha}{\operatorname{argmax}} r. \quad (3.9)$$

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 the problem (3.8). The main source of difficulty lies in the geometry underlying the collection of tangent spaces to determinantal varieties. In particular, solving (3.8) in the case of variable selection is easy because the operators $\mathcal{P}_T, \mathcal{P}_{\text{avg}}$ are both diagonal (and hence trivially simultaneously diagonalizable) in that case; as a result, one can decompose (3.8) into a set of one-variable problems. In contrast, the operators $\mathcal{P}_T, \mathcal{P}_{\text{avg}}$ are not simultaneously diagonalizable in the low-rank case, and consequently there doesn't appear to be any clean separability in (3.8) in the general setting with determinantal varieties.

We describe next a heuristic to approximate (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 $\{(\hat{\mathcal{C}}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))\}_{i=1}^B$ denote the row/column space estimates from B subsamples $\{\mathcal{D}_i\}_{i=1}^B \subset \mathcal{D}$ of the data. We average the projection operators associated to these row/column spaces:

$$\mathcal{P}_{\text{avg}}^{\mathcal{C}} = \frac{1}{B} \sum_{i=1}^B \mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}_i)}, \quad \mathcal{P}_{\text{avg}}^{\mathcal{R}} = \frac{1}{B} \sum_{i=1}^B \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)} \quad (3.10)$$

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 $\mathbb{R}^{p_1 \times p_2}$. In contrast, the average operators $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$ are self-adjoint maps on the spaces $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively. Based on these separate column-space and row-space averages, we approximate (3.8) as follows:

$$T_{\text{approx}}(r) = T \left(\underset{\mathcal{C} \subset \mathbb{R}^{p_2} \text{ subspace of dimension } r}{\operatorname{argmax}} \sigma_{\min}(\mathcal{P}_{\mathcal{C}} \mathcal{P}_{\text{avg}}^{\mathcal{C}} \mathcal{P}_{\mathcal{C}}), \underset{\mathcal{R} \subset \mathbb{R}^{p_1} \text{ subspace of dimension } r}{\operatorname{argmax}} \sigma_{\min}(\mathcal{P}_{\mathcal{R}} \mathcal{P}_{\text{avg}}^{\mathcal{R}} \mathcal{P}_{\mathcal{R}}) \right). \quad (3.11)$$

The advantage of this latter 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 (resp. row-space) approximation of dimension r is equal to the span of the eigenvectors corresponding to the r largest eigenvalues of $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ (resp. $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$). We have that $\sigma_{\min}(\mathcal{P}_{T_{\text{approx}}(r)} \mathcal{P}_{\text{avg}} \mathcal{P}_{T_{\text{approx}}(r)}) \leq \sigma_{\min}(\mathcal{P}_{T_{\text{OPT}}(r)} \mathcal{P}_{\text{avg}} \mathcal{P}_{T_{\text{OPT}}(r)})$ 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 the case, 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 $\mathcal{M}_{\mathcal{C}}, \mathcal{M}_{\mathcal{R}}$ do not have a transverse intersection, 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 the Section 4. Further, our final estimate of a tangent space still involves the solution of (3.9) using the approximation (3.11) instead of (3.8). Consequently, we continue to retain our guarantees from Section 3.2 on false discovery control. The full procedure is presented in Algorithm 1.

Algorithm 1 Subspace Stability Selection Algorithm

- 1: **Input:** A set of observations $\mathcal{D}$; a collection of subsamples $\{\mathcal{D}_i\}_{i=1}^B \subset \mathcal{D}$; a row/column space (equivalently, tangent space) estimation procedure $(\hat{\mathcal{C}}, \hat{\mathcal{R}})$; a parameter $\alpha \in (0, 1)$.
 - 2: **Obtain Tangent Space Estimates:** For each bag $\{\mathcal{D}_i, i = 1, 2, \dots, B\}$, obtain row/column space estimates $\{(\hat{\mathcal{C}}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))\}_{i=1}^B$ and set $\hat{T}(\mathcal{D}_i) = T(\hat{\mathcal{C}}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))$.
 - 3: **Compute Average Projection Operators:** Compute the average tangent space projection operator $\mathcal{P}_{\text{avg}}$ according to (3.2) and the average row/column space projection operators $\mathcal{P}_{\text{avg}}^{\mathcal{R}}, \mathcal{P}_{\text{avg}}^{\mathcal{C}}$ according to (3.10).
 - 4: **Compute Optimal Row/Column Space Approximations:** Compute ordered singular vectors $\{u_1, u_2, \dots, u_{p_1}\} \subset \mathbb{R}^{p_1}$ and $\{v_1, v_2, \dots, v_{p_2}\} \subset \mathbb{R}^{p_2}$ of $\mathcal{P}_{\text{avg}}^{\mathcal{C}}$ and $\mathcal{P}_{\text{avg}}^{\mathcal{R}}$, respectively. For each $r = 1, \dots, \min\{p_1, p_2\}$, set $\mathcal{C}^*(r) = \text{span}(u_1, \dots, u_r)$ and $\mathcal{R}^*(r) = \text{span}(v_1, \dots, v_r)$.
 - 5: **Tangent Space Selection via (3.9):** Let r_{S3} denote the largest r such that $T(\mathcal{C}^*(r), \mathcal{R}^*(r)) \in \mathcal{T}_\alpha$.
 - 6: **Output:** Tangent space $T_{\text{S3}} = T(\mathcal{C}^*(r_{\text{S3}}), \mathcal{R}^*(r_{\text{S3}}))$.
-

Computational Cost of Algorithm 1 – We do not account for the cost of obtaining the row/column space estimates $\{(\hat{\mathcal{C}}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))\}_{i=1}^B$ on each subsample in Step 2, and focus exclusively on the cost of combining these estimates via Steps 3 – 5. In Step 3, the computational complexity of computing the average projection maps $\mathcal{P}_{\text{avg}}^{\mathcal{R}}, \mathcal{P}_{\text{avg}}^{\mathcal{C}}$ requires $\mathcal{O}(B \max\{p_1, p_2\}^2)$ operations and computing the average tangent space projection map $\mathcal{P}_{\text{avg}}$ requires $\mathcal{O}(B p_1^2 p_2^2)$ 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 $\mathcal{O}(\max\{B p_1^2, B p_2^2, B p_1^2 p_2^2, p_1^3, p_2^3, p_1^3 p_2^3\})$.

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 $B p_1^2 p_2^2$ and $p_1^3 p_2^3$ in the final expression is due to 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 only consists of computations involving projection maps onto row and column spaces (which belong to $\mathbb{R}^{p_2}$ and $\mathbb{R}^{p_1}$ respectively).

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 only computing the average row/column space projection maps $\mathcal{P}_{\text{avg}}^{\mathcal{R}}, \mathcal{P}_{\text{avg}}^{\mathcal{C}}$. 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 $\Gamma_{r'_{\text{S3}}, r'_{\text{S3}}} \geq \alpha$ and $\Gamma_{r'_{\text{S3}}, r'_{\text{S3}}} \geq \alpha$. One can check that the cost associated to this modified procedure is $\mathcal{O}(\max\{B p_1^2, B p_2^2, p_1^3, p_2^3\})$.

This modified method has the property that the row and column spaces are individually well-aligned with the averages produced from the subsamples. However, it is a priori unclear whether the resulting tangent space belongs to some set of stable tangent spaces. The following proposition addresses this question:

Proposition 5 (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, $T_{\text{S3-modified}} \in \mathcal{T}_{1-4(1-\alpha)}$.*

Proof. Let T be a tangent space produced by the modified algorithm with associated column and row spaces $(\mathcal{C}, \mathcal{R})$.

We proceed by obtaining an upper bound on $\|\mathcal{P}_T(\mathcal{I} - \mathcal{P}_{\text{avg}})\mathcal{P}_T\|_2$, which gives a lower bound on $\sigma_{\min}(\mathcal{P}_T\mathcal{P}_{\text{avg}}\mathcal{P}_T)$:

$$\begin{aligned}
\|\mathcal{P}_T(\mathcal{I} - \mathcal{P}_{\text{avg}})\mathcal{P}_T\|_2 &= \max_{M \in T, \|M\|_F=1} \frac{1}{B} \text{trace} \left(\sum_{i=1}^B M' \mathcal{P}_{\hat{T}(\mathcal{D}_i)^\perp} (M) \right) \\
&\stackrel{(a)}{=} \max_{M \in T, \|M\|_F=1} \frac{1}{B} \sum_{i=1}^B \|\mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}_i)^\perp} M \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp}\|_F^2 \\
&\stackrel{(b)}{\leq} \max_{M \in T, \|M\|_F=1} \frac{2}{B} \sum_{i=1}^B \|\mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}_i)^\perp} \mathcal{P}_{\mathcal{C}} M \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp}\|_F^2 + \frac{2}{B} \sum_{i=1}^B \|\mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}_i)^\perp} \mathcal{P}_{\mathcal{C}^\perp} M \mathcal{P}_{\mathcal{R}} \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp}\|_F^2 \\
&\stackrel{(c)}{\leq} \max_{M \in T, \|M\|_F=1} \frac{2}{B} \sum_{i=1}^B \|\mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}_i)^\perp} \mathcal{P}_{\mathcal{C}} M\|_F^2 + \frac{2}{B} \sum_{i=1}^B \|\mathcal{P}_{\mathcal{R}} \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp} M'\|_F^2 \\
&= \max_{M \in T, \|M\|_F=1} 2 \text{trace}(\mathcal{P}_{\mathcal{C}}(\mathcal{I} - \mathcal{P}_{\text{avg}})\mathcal{P}_{\mathcal{C}} M M') + 2 \text{trace}(\mathcal{P}_{\mathcal{R}}(\mathcal{I} - \mathcal{P}_{\text{avg}})\mathcal{P}_{\mathcal{R}} M' M) \\
&\leq 2 \|\mathcal{P}_{\mathcal{C}}(\mathcal{I} - \mathcal{P}_{\text{avg}})\mathcal{P}_{\mathcal{C}}\|_2 + 2 \|\mathcal{P}_{\mathcal{R}}(\mathcal{I} - \mathcal{P}_{\text{avg}})\mathcal{P}_{\mathcal{R}}\|_2 \leq 4(1 - \alpha).
\end{aligned}$$

Here (a) follows from the cyclicity of the trace functional and the idempotence of projection maps; (b) from the fact that $M \in T$ implies that $M = \mathcal{P}_{\mathcal{C}}M + \mathcal{P}_{\mathcal{C}^\perp}M\mathcal{P}_{\mathcal{R}}$ and the elementary inequality $(a + b)^2 \leq 2a^2 + 2b^2$; and (c) from the property $\|A\mathcal{P}\|_F \leq \|A\|_F$ for any projection matrix $\mathcal{P}$. $\square$

Proposition 5 guarantees that our modification of Algorithm 1 can be a valuable substitute in settings where the problem dimension is large as it continues to provide a false discovery control guarantee. We use this modified approach in some of our larger experiments in Section 4.

3.4 Further Illustrations

In the remainder of this section, we explore various facets of Algorithm 1 via illustrations on the synthetic matrix completion problem setup 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.

Illustration : α vs. r_{S3} – The threshold parameter α determines the eventual optimal rank r_{S3} , with larger values of α yielding a smaller r_{S3} . To better understand this relationship, we plot in Figure 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 $\text{SNR} = \{0.2, 0.5, 10, 50\}$. Each curve in the different plots corresponds to a particular value of r_{S3} , with the solid curves representing $r_{\text{S3}} = 1, \dots, 10$ and the dotted curves representing $r_{\text{S3}} = 11, \dots, 70$. 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 due to 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 indexed by 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 $r_{\text{S3}} = 4, \dots, 8$, 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^* . On the other hand, for low values of SNR, the clustering is less pronounced as the components of L^* with small singular values are overwhelmed by noise.

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 dataset. We choose the noise level so that SNR takes on one of the values in $\{1.5, 2, 2.5, 3\}$. On the one hand, we employ the 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 constitute the test set. We use cross-validation to identify an optimal choice λ^* of the regularization parameter. The estimate produced by (3.1) on the training set for this choice of λ^* is recorded as the output of the non-subsampled approach. On the other hand, the 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 run ten experiments and average to find an empirical approximation to the expected false discovery (2.3). Table 1 compares the expected false discovery of the non-subsampled approach to that of the subspace stability selection procedure for the different

Figure 2: Relationship between r_{s3} and α in Algorithm 1 for a large range of λ and $\text{SNR} = \{0.4, 0.8, 1.2, 50\}$.

problem settings. Evidently, subspace stability selection yields a much smaller amount of false discovery compared to not employing subsampling.

Method	SNR = 1.5	SNR = 2	SNR = 2.5	SNR = 3
No subsampling	1274.8	1423.4	1525.3	1576.2
Subspace stability selection	37.1	43.8	52.9	61.4

Table 1: False discovery of subspace stability selection vs a non-subsampled approach on the stylized matrix completion problem. The maximum possible amount of false discovery is $\dim(T^{\star\perp}) = (70 - 10)^2 = 3600$.

At this stage, it is natural to wonder whether the source of the improved false discovery control provided by subspace stability selection over not using subsampling is simply due to the non-subsampled approach providing 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 have smaller expected false discovery. Thus, is subsampling better primarily because it produces lower-rank estimates? To address this point in our stylized setup, we consider a population L^* with associated incoherence parameter equal to 0.8¹. We sweep over the regularization parameter λ , and we compare the following two estimates: first, the estimate $\hat{L}$ obtained via (3.1) and then truncated to its first three singular values, and subsampled estimates obtained via Algorithm 1 with r_{s3} set to three. The choice of three here is motivated by the

¹The 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 i 'th standard basis vector, and it plays a prominent role in various analyses of the low-rank matrix completion problem (*Candés and Recht, 2009*).

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 Figure 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 (3.1) applied to the entire dataset. Similar behavior is also

Figure 3: False discovery of subspace stability selection vs a non-subsampled approach with $\text{SNR} = 1.6, 0.8$. Here, we choose a rank-3 approximation of the non-subsampled approach and $r_{\text{S3}} = 3$ in Algorithm 1 of subspace stability selection. The maximum possible amount of false discovery is $\dim(T^{\star\perp}) = (70 - 10)^2 = 3600$. Here ‘N-S’ denotes no subsampling and ‘S3’ denotes subspace stability selection.

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 $r = 1, 2, \dots, 5$ and compare its false discovery to the estimate produced by subspace stability selection with $r_{\text{S3}} = r$ (shown in Table 2).

Method	rank = 1	rank = 2	rank = 3	rank = 4	rank = 5
No subsampling	29.07	74.19	144.2	208.006	268.04
Subspace stability selection	23.64	51.69	98.01	128.00	175.49

Table 2: False discovery of subspace stability selection vs a non-subsampled approach with $\text{SNR} = 0.8$ and rank of the estimate set to vary from 1 to 5. The maximum possible amount of false discovery is $\dim(T^{\star\perp}) = 3600$.

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 (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

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

which measures the degree to which T and $\tilde{T}$ are misaligned. If $T = \tilde{T}$, then $\mu(T, \tilde{T}) = 0$, and on the other hand, $T \subseteq \tilde{T}^\perp$ 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 approach with subspace stability selection. In our stylized setup, we choose the noise level so that $\text{SNR} = 4$ and we select $\lambda = 0.03$ (based on cross-validation). Letting T be the tangent space of the 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.

4 Experiments

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

1. Low-rank matrix completion: This is the problem considered in the stylized demonstrations of Section 3.1. One point of departure from that discussion in the present section is that in experiments with real data, the dimensions p_1, p_2 can get very large. As a result, employing the nuclear norm regularized estimator (3.1) on each subsample is impractical. Instead, we use on each subsample the following non-convex formulation that is commonly utilized in large-scale settings:

$$(\hat{U}, \hat{V}) = \underset{U \in \mathbb{R}^{p_1 \times k}, V \in \mathbb{R}^{p_2 \times k}}{\operatorname{argmin}} \quad \|(Y - UV')_S\|_F^2 + \lambda (\|U\|_F^2 + \|V\|_F^2), \quad (4.1)$$

where $\|U\|_F^2 + \|V\|_F^2$ serves as a surrogate for the nuclear norm penalty in (3.1), $\lambda > 0$ is a regularization parameter, and $S \subset \{1, \dots, p_1\} \times \{1, \dots, p_2\}$ 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 with larger values of λ leading to a smaller rank. Fixing U (resp. V) the above problem is convex in V (resp. U), and thus a natural approach that is employed in practice is alternating least-squares (ALS) (*Jain et al*, 2013; *Srebro et al*, 2005; *Zhou et al*, 2008).

2. Linear measurements: this problem is in some sense a generalization of matrix completion in which we are given noisy linear functionals of a low rank matrix $L^* \in \mathbb{R}^{p_1 \times p_2}$ of the form $Y_i \approx \langle \mathcal{A}_i, L^* \rangle$, $i = 1, \dots, n$ where each $\mathcal{A}_i \in \mathbb{R}^{p_1 \times p_2}$. In this case, the estimator that we employ on subsamples is the following convex program (*Recht et al*, 2010):

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2}}{\operatorname{argmin}} \quad \sum_{i \in S} (Y_i - \langle \mathcal{A}_i, L \rangle)^2 + \lambda \|L\|_*. \quad (4.2)$$

Here $\lambda > 0$ is a regularization parameter and $S \subset \{1, \dots, n\}$ represents the subset of measurements corresponding to a subsample.

3. Factor analysis: We observe samples $\{Y^{(i)}\}_{i=1}^n \subset \mathbb{R}^p$ of a random vector and we wish to identify a factor model that best explains these observations, i.e., a model in which the coordinates 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 low-rank matrix. Using the Woodbury Inversion Lemma, we have that the precision matrix can be decomposed as a diagonal matrix minus a low-rank matrix. The virtue of working with precision matrices is that the log-likelihood function is concave with respect to this parametrization. On each subsample, we use the following estimator (*Shapiro*, 1982):

$$(\hat{D}, \hat{L}) = \underset{L \in \mathbb{S}^p, D \in \mathbb{S}^p}{\operatorname{argmin}} \quad -\log \det(D - L) + \operatorname{trace} \left(\left(\frac{1}{|S|} \sum_{i \in S} Y^{(i)} Y^{(i)'} \right) (D - L) \right) + \lambda \operatorname{trace}(L). \quad (4.3)$$

subject to $D - L \succ 0, L \succeq 0, D$ is diagonal

Here $\lambda > 0$ is a regularization parameter, $S \subset \{1, \dots, n\}$ represents the subset of measurements corresponding to a subsample, and $\operatorname{trace}(\cdot)$ is the restriction of the nuclear norm to symmetric positive-semidefinite matrices.

4.1 Synthetic Simulations

To corroborate the false discovery bound of Theorem 4, we consider the factor modeling setup. We draw i.i.d samples of a Gaussian random vector in $\mathbb{R}^p$ that is distributed as $\mathcal{N}(0, \mathcal{I} + L^*)$ with $L^* \in \mathbb{R}^{p \times p}$ being symmetric and positive-semidefinite; we set $p = 70$ and the rank of L^* to either 6 or 10, with row/column spaces selected uniformly at random from the Steifel manifold and the nonzero singular values equal to 0.3. We use $n_{\text{train}} = 105p$ samples drawn from this model as training data for the estimator (4.3), which is solved using the LogDetPPA solver (*Toh et al*, 2006). The regularization parameter λ is selected based on holdout validation on a separate

$n_{\text{test}} = 45p$ observations. With this choice of regularization parameter, we apply subspace stability selection with $\alpha \in [0.85, 0.95]$ and $B = 100$ complementary bags, and we obtain an empirical approximation of the expected false discovery over 50 trials. Since the population model is known, the quantities inside Theorem 4 are readily obtainable. Figures 4(a) and Figure 4(b) compare the computed false discovery and the bound from Theorem 4 for $\text{rank}(L^*) = 6$ and $\text{rank}(L^*) = 10$. As expected, the result of Theorem 4 bounds the false discovery of subspace stability selection.

Figure 4: False discovery of subspace stability selection procedure as a function of α for a synthetic factor model with (a) rank = 6 and (b) rank = 10. The blue curve is the false discovery obtained by subspace stability selection, the red curve is the bound from Theorem 4, and the dotted line is $\dim(T^{\star\perp})$.

Next, we focus on the matrix completion setting. We consider a population low-rank matrix $L^* \in \mathbb{R}^{p \times p}$ with $p = 70$, the rank of L^* in the set $\{1, 2, 3, 4\}$, the nonzero singular values set to 1, and the row and column spaces sampled uniformly from the Steifel manifold. A fraction $1/10$ of the total entries are chosen uniformly at random as the observation set Ω so that $|\Omega| = p^2/10$. Furthermore, these observations are corrupted by i.i.d Gaussian noise with mean zero and variance chosen so that the SNR is in the set $\{1, 2, 3, 4, 5\}$. This generates a total of 20 problem instances (i.e. 4 different ranks and 5 different SNR levels). For each problem instance, we solve for a low rank matrix using the nuclear norm estimator (3.1) with λ chosen via holdout validation with a validation set of size $p^2/20$. With this choice of λ , we evaluate the expected false discovery and the power computed empirically over 10 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. Figure 5(a) demonstrates the performance of the non-subsampled approach and subspace stability selection for all the problem settings. Evidently, for most problem instances, subspace stability selection yields a solution with a significantly smaller amount of false discovery without much loss in power. We repeat the same experiment with general linear measurements, where the linear functionals specifying the measurements consist of i.i.d. standard Gaussian entries. These measurements are corrupted with Gaussian noise of mean zero and variance chosen so that SNR lies in the set $\{1, 2, 3, 4, 5\}$. We obtain $n_{\text{train}} = p^2/10$ measurements as input to the estimator (4.2), with λ selected based on holdout validation on a $n_{\text{test}} = p^2/20$ validation set. Figure 5(b) compares the performance of the non-subsampled approach and subspace stability selection. Once again, we see a vast reduction in false discoveries without much loss in power in most problem settings.

4.2 Experimental Results on Real Datasets

4.2.1 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 solve this problem based either on the convex relaxation (3.1) or the non-convex approach (4.1) via ALS. We describe experimental results on two popular datasets in collaborative filtering: 1) the Amazon Book-Crossing dataset (obtained from <http://www2.informatik.uni-freiburg.de/~cziegler/BX/>) of which we

Figure 5: False discovery vs power with (a) matrix completion and (b) linear measurements over 20 different problem instances (varying rank and noise level). Blue circles corresponds to the performance of the non-subsampled approach and red crosses correspond to subspace stability selection with $\alpha = 0.7$. The lines connect dots corresponding to the same problem instance. Both the false discovery and the power are normalized by dividing the expressions (2.3) and (2.4) by $\dim(T^{\star\perp})$ and $\dim(T^{\star})$, respectively.

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 2) the Amazon Video Games dataset (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 dataset as follows: we set aside 85% of the observations as a training set, 10% of the observations as a holdout 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 the formulation (2.6) given estimates of tangent spaces. We set $\alpha = 0.7$ and $B = 100$ complementary bags. Figure 6 illustrates the mean squared error of ALS and subspace stability selection on the holdout set for these two datasets for a range of values of the regularization parameter λ . For both datasets, we observe that subspace stability selection yields models with better MSE on the holdout set over the entire range of regularization parameters. On the Book-Crossings dataset, we further note that at the cross-validated λ , the rank of the estimate obtained from the non-subsampled approach is 80 (i.e., the maximum allowable rank) with the first three singular values equal to 4329, 135.4, 63.1. The MSE of this model on the evaluation set is equal to 0.83. On the other hand, 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 dataset, the rank of the estimate obtained from the non-subsampled approach is 39 with the first five singular values equal to 1913.5, 49.4, 43.6, 28.4, 27.4, with an MSE of 0.87 on the evaluation set. On the other hand, 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 datasets 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 .

4.2.2 Hyperspectral unmixing

Here we give an illustration with real hyperspectral imaging data in which the underlying population parameters are known based on 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 j 'th image pixel to the i 'th frequency band. The spectral unmixing problem aims to find $W \in \mathbb{R}^{p_1 \times k}$ (called the endmember 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 endmembers (Manolakis, 2003). Of particular interest is the k -dimensional column-space of W , which corresponds to the space spanned by the k endmembers that are present in the image. We discuss two natural hyperspectral unmixing

Figure 6: Collaborative filtering: MSE on holdout set of non-subsampled approach (denoted ‘N-S’ and colored in blue) and subspace stability selection (denoted ‘S3’ and colored in red). Dotted black line represents the cross-validated choice of λ with the non-subsampled approach.

problems that arise commonly in practice. We focus on the Urban dataset (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 $2 \times 2m^2$ area with 210 wavelengths ranging from $400nm$ to $2500nm$. Following previous analyses of this dataset, we remove 48 noisy channels to obtain 162 wavelengths and select a 30×25 patch (equal to 750 pixels) shown in Figure 7(a). In the selected patch, there are a total of 3 endmembers (shown in Figure 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

Figure 7: Urban hyperspectral image (left) and spectra of three materials present in the image (right). The data and the population spectra are obtained from http://www.escience.cn/people/feiyunZHU/Dataset_GT.html.

accurately reconstruct a hyperspectral image 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 especially popular in hyperspectral unmixing (*Sun and Kelly, 2009*). To simulate such a hyperspectral unmixing problem, we randomly subsample 10% of the hyperspectral data in the patch as training data. We further select another 10% of the remaining data as a holdout 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. Due to the scale of this problem being 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 endmember detection, the quantities of interest for evaluating performance are $\overline{FD} =$

$\mathbb{E} \left[\text{trace} \left(\mathcal{P}_{\text{col-space}(W^*)^\perp} \mathcal{P}_{\text{col-space}(\widehat{W})} \right) \right]$ and $\overline{\text{PW}} = \mathbb{E} \left[\text{trace} \left(\mathcal{P}_{\text{col-space}(W^*)} \mathcal{P}_{\text{col-space}(\widehat{W})} \right) \right]$. 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 endmembers in Figure 7(b)), and $\widehat{W}$ is the estimated matrix. At the cross-validated choice of $\lambda = 0.79$ with no subsampling, ALS produces a rank-20 estimate with $\overline{\text{FD}} = 0.1 \dim(\text{col-space}(W^{*\perp}))$ and $\overline{\text{PW}} = 0.98 \dim(\text{col-space}(W^*))$. In contrast, for the same $\lambda = 0.79$, subspace stability selection produces a rank-3 estimate with $\overline{\text{FD}} = 0.0005 \dim(\text{col-space}(W^{*\perp}))$ and $\overline{\text{PW}} = 0.96 \dim(\text{col-space}(W^*))$. Furthermore, even if λ is set large enough (for example, $\lambda = 12.1$) so that the non-subsampled ALS estimate has rank equal to 3, this solution yields $\overline{\text{FD}} = 0.003 \dim(\text{col-space}(W^*)^\perp)$, 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 approximated by a low-rank matrix. Thus, another natural approach for endmember 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, hence the row and column spaces are the same) serve as estimates of the subspace spanned by the endmembers. 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 $\text{SNR} = 0.78$. We then set aside 80% of the data as training data and the remaining 20% as a holdout validation set. Employing the estimator (4.3) without subsampling and with λ chosen via cross-validation yields false discovery $\text{FD} = 0.04 \dim(T^{*\perp})$ and power $\text{PW} = 0.46 \dim(T^*)$. (Here T^* represents the population tangent space.) On the other hand, subspace stability selection with $\alpha = 0.7$ and $B = 100$ complementary bags yields a tangent space estimate with a false discovery and power $\text{FD} = 0.015 \dim(T^{*\perp})$ and $\text{PW} = 0.68 \dim(T^*)$, respectively. 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 proposed framework 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 datasets in a range of low-rank estimation tasks.

There are several interesting directions for further investigation that arise from our work. First, Algorithm 1 from Section 3.3 outputs an estimate that does provide false discovery control, but it is unclear whether this is the most powerful procedure possible. In particular, it is of interest to obtain an optimal solution to the problem (3.8), or to prove that Algorithm 1 computes a near-optimal solution. Next, 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 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 associated with the population tangent space and the estimated tangent space (when translated to the geometric viewpoint of our paper). Finally, the geometric framework developed in this paper for assessing false discovery is potentially relevant beyond the specific setting of low-rank estimation. For example, our setup 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 presented here may be useful in addressing many other structured estimation problems.

References

- BHATIA, R. (2013) Matrix analysis. *Springer Science & Business Media*.
- BENJAMINI, Y. AND Y. HOCKBERG(1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing . *Journal of Royal Statistical Society Series B*, 57(1), 289–300.
- CANDÉS, E. AND B. RECHT (2009) Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9:717.
- CHANDRASEKARAN, V., P. A. PARILLO AND A. S. WILLSKY (2012) Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4), 1935–1967.
- CHOI, Y., J. TAYLOR AND R. TIBSHIRANI (2017) Selecting the number of principal components: estimation of the true rank of a noisy matrix. *Annals of Statistics*, 45(6), 2590–2617.
- FA, R. AND R. LAMARE (2011) *Reduced-Rank STAP Algorithms using Joint Iterative Optimization of Filters*. IEEE Transactions on Aerospace and Electronic Systems, 47(3), 1668 –1684.
- FAZEL, M. (2002) *Matrix rank minimization with applications*. PhD thesis, Dept. Elec. Engr., Stanford University.
- GELMAN, A. AND E. LOKEN(2014) *The statistical crisis in science*. American Scientist, 102: 460–465.
- GOLDBERG, D., D. NICHOLS, B. OKI AND D. TERRY (1992) Using collaborative filtering to weave an information tapestry. *Communication ACM*, 35(12), 61–70.
- HARRIS, J. (1995) Algebraic Geometry: A First Course. *Springer-Verlog*, Berlin.
- HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441.
- JAIN, P., P. NETRAPALLI, S. SANGHAVI (2013) Low-rank matrix completion using alternating minimization . Proceedings of the forty-fifth annual ACM symposium on Theory of computing, 665–674.
- Ji, H., C. LIU, Z. SHEN AND Y. XU (2010) Robust video denoising using low rank matrix completion. *Computer Vision and Pattern Recognition*, 15, 1791–1798.
- KOLDA, T. AND B. BADER (2009) Tensor decompositions and applications. *SIAM review*, 51, 455–500.
- LIU, Z. AND L. VANDENBERGHE(2009) Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3), 1235–1256.
- LIU, Z. AND X. LIN (2018) A Geometric Perspective on the Power of Principal Component Association Tests in Multiple Phenotype Studies. *Journal of the American Statistical Association* , 1–36.
- MANOLAKIS, D. (2003) Detection algorithms for hyperspectral imaging applications: a signal processing perspective. *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, 378–384.
- MEINSHAUSEN, N. AND P. BÜHLMANN (2010) Stability Selection. *Journal of Royal Statistical Methodology (Series B)*, 72(4), 417–473.
- PATI, Y. AND T. KAILATH (1994) Phase-shifting masks for microlithography: Automated design and mask requirements. *Journal of the Optical Society of America A*, 11(9).
- RECHT, B., M. FAZEL AND P. PARRILO (2010) Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *Siam Review*, 52, 471–501.
- SHAH, R. AND J. SAMWORTH (2013) Variable selection and error control: another look at stability selection. *Journal of Royal Statistical Society (Series B)*, 75(1), 55–80.
- SHAPIRO (1982) Weighted minimum trace factor analysis. *Psychometrika*, 47, 243–264.
- SONG, J. AND S. SHIN (2018) Stability approach to selecting the number of principal components. *Computational Statistics*, 33(4), 1923–1938.

- SPEARMAN, C. (1904) ‘General intelligence’, objectively determined and measured. *American Journal of Psychology*, 15, 201–292.
- SREBRO, N., J. RENNIE AND T. JAAKKOLA (2005) Maximum Margin Matrix Factorizations. *Advances in Neural Information Processing Systems*, 17.
- SREBRO, N. AND A. SHRAIBMAN (2005) Rank, trace-norm and max-norm. *Proceedings of the 18th Annual Conference on Learning Theory*, 545–560.
- SUN, T. AND K. KELLY (2009) Compressive sensing hyperspectral imager. *Computational and Optical Sciences Imagery*.
- TOH, K. C, M. J. TODD AND R. H. TUTUNCU (2006) SDPT3 - A MATLAB software package for semidefinite-quadratic-linear programming. Available from <http://www.math.nus.edu.sg/~mattohkc/sdpt3.html>. Accessed April 2016.
- ZHOU, Y., D. WILKINSON, R. SCHREIBER, AND R. PAN (2008) Large-scale Parallel Collaborative Filtering for the Netflix Prize. *Proceedings of the 4th international conference on Algorithmic Aspects in Information and Management*, 337–348.

Supplemental Material: Subspace Stability Selection for Variable Selection

As described in Remark 3 of Theorem 4, the result of *Meinshausen and Bühlmann* (2010) is obtained by considering a specific basis (standard basis coordinate) that simultaneously diagonalizes the subsample and population projection matrices. In this section, we illustrate how this observation leads to a refined bound. We first suppose that $w_j \in T$. Then:

$$\begin{aligned}
\text{trace}(\mathcal{P}_T \mathcal{P}_{\text{span}(w_j)}) &\stackrel{(a)}{\leq} \min_{i=1,2,\dots,B/2} \min_{k \in \{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star} + \text{trace}(\mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^{\perp}} \mathcal{P}_{\text{span}(w_j)}) \\
&\stackrel{(b)}{\leq} \min_{i=1,2,\dots,B/2} \min_{k \in \{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star} + \text{trace}(\mathcal{P}_{T \cap \text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^{\perp}}) \\
&\stackrel{(c)}{\leq} \min_{i=1,2,\dots,B/2} \frac{\min_{k \in \{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star}}{\min_{k \in \{0,1\}} \text{trace}(\mathcal{P}_{T \cap \text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})})} \\
&\stackrel{(d)}{=} \min_{i=1,2,\dots,B/2} \frac{\prod_{k \in \{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star}}{\prod_{k \in \{0,1\}} \text{trace}(\mathcal{P}_{T \cap \text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})})} \\
&\stackrel{(e)}{\leq} \frac{\frac{2}{B} \sum_{i=1}^{B/2} \prod_{k \in \{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star}}{\frac{2}{B} \sum_{i=1}^{B/2} \prod_{k \in \{0,1\}} \text{trace}(\mathcal{P}_{T \cap \text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})})} \\
&\stackrel{(f)}{\leq} \frac{\frac{2}{B} \sum_{i=1}^{B/2} \prod_{k \in \{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star}}{\frac{2}{B} \sum_{i=1}^{B/2} \sum_{k \in \{0,1\}} \text{trace}(\mathcal{P}_{T \cap \text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})}) - 1} \\
&\stackrel{(g)}{\leq} \frac{\frac{2}{B} \sum_{i=1}^{B/2} \prod_{k \in \{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star}}{2\alpha - 1} \tag{5.1}
\end{aligned}$$

Here (a) follows from the decomposition (3.7) and the fact that the choice of subsample in (3.7) is arbitrary, (b) follows from simultaneous diagonalizability of projection matrices of the sparse variety, (c) is from noting that $\left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})} \mathcal{P}_{\text{span}(w_j)} \right\|_{\star} \in \{0,1\}$ and $\text{trace}(\mathcal{P}_{T \cap \text{span}(w_j)} \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^{\perp}}) \in \{0,1\}$ due to simultaneous diagonalizability of these projection matrices and that $\min\{a+b, c+d\} \leq \frac{\min\{a,b\}}{\min\{c,d\}}$ for $a, b, c, d \in \{0,1\}$, (d) uses the fact that $\min\{a,b\} = ab$ for $a, b \in \{0,1\}$, (e) is due to the inequality $\min\{\frac{a}{b}, \frac{c}{d}\} \leq \frac{a+c}{b+d}$ for $a, b, c, d \geq 0$, (f) is from $ab \geq a+b-1$ for $a, b \in (0,1)$, and (g) uses the fact that $\sigma_{\min}(\mathcal{P}_T \mathcal{P}_{\text{span}(w_j)} \mathcal{P}_T) \geq \alpha$. We assumed that $w_j \in T$. If this was not the case, then $\text{trace}(\mathcal{P}_T \mathcal{P}_{\text{span}(w_j)}) = 0$, and so the bound in (5.1) is still valid. Noting that $\text{trace}(\mathcal{P}_T \mathcal{P}_{T^{\perp}}) = \sum_{j=1}^{\dim(T^{\perp})} \text{trace}(\mathcal{P}_T \mathcal{P}_{\text{span}(w_j)})$, and taking an expectation gives the desired result. $\square$