A determinantal point process for column subset selection

Ayoub Belhadji, R. Bardenet, Pierre Chainais

A determinantal point process for column subset selection

Ayoub Belhadji[1]*, Rémi Bardenet[1], Pierre Chainais[1]
[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, 59651 Villeneuve d'Ascq, France

Abstract

Dimensionality reduction is a first step of many machine learning pipelines. Two popular approaches are principal component analysis, which projects onto a small number of well chosen but non-interpretable directions, and feature selection, which selects a small number of the original features. Feature selection can be abstracted as a numerical linear algebra problem called the column subset selection problem (CSSP). CSSP corresponds to selecting the best subset of columns of a matrix $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, where *best* is often meant in the sense of minimizing the approximation error, i.e., the norm of the residual after projection of $\boldsymbol{X}$ onto the space spanned by the selected columns. Such an optimization over subsets of $\{1,\dots,d\}$ is usually impractical. One workaround that has been vastly explored is to resort to polynomial-cost, random subset selection algorithms that favor small values of this approximation error. We propose such a randomized algorithm, based on sampling from a projection determinantal point process (DPP), a repulsive distribution over a fixed number k of indices $\{1,\dots,d\}$ that favors diversity among the selected columns. We give bounds on the ratio of the expected approximation error for this DPP over the optimal error of PCA. These bounds improve over the state-of-the-art bounds of *volume sampling* when some realistic structural assumptions are satisfied for $\boldsymbol{X}$. Numerical experiments suggest that our bounds are tight, and that our algorithms have comparable performance with the *double phase* algorithm, often considered to be the practical state-of-the-art. Column subset selection with DPPs thus inherits the best of both worlds: good empirical performance and tight error bounds.

Contents

*Corresponding author: ayoub.belhadji@centralelille.fr

1 Introduction

Datasets come in always larger dimensions, and dimension reduction is thus often one the first steps in any machine learning pipeline. Two of the most widespread strate-

gies are principal component analysis (PCA) and feature selection. PCA projects the data in directions of large variance, called principal components. While the initial features (the canonical coordinates) generally have a direct interpretation, principal components are linear combinations of these original variables, which makes them hard to interpret. On the contrary, using a selection of original features will preserve interpretability when it is desirable. Once the data are gathered in an $N \times d$ matrix, of which each row is an observation encoded by d features, feature selection boils down to selecting columns of $\boldsymbol{X}$. Independently of what comes after feature selection in the machine learning pipeline, a common performance criterion for feature selection is the approximation error in some norm, that is, the norm of the residual after projecting $\boldsymbol{X}$ onto the subspace spanned by the selected columns. Optimizing such a criterion over subsets of $\{1, \dots, d\}$ requires exhaustive enumeration of all possible subsets, which is prohibitive in high dimension. One alternative is to use a polynomial-cost, random subset selection strategy that favors small values of the criterion.

This rationale corresponds to a rich literature on randomized algorithms for column subset selection (Deshpande and Vempala, 2006; Drineas et al., 2008; Boutsidis et al., 2011). A prototypal example corresponds to sampling s columns of $\boldsymbol{X}$ i.i.d. from a multinomial distribution of parameter $\boldsymbol{p} \in \mathbb{R}^d$. This parameter $\boldsymbol{p}$ can be the squared norms of each column (Drineas et al., 2004), for instance, or the more subtle k-leverage scores (Drineas et al., 2008). While the former only takes $\mathcal{O}(dN^2)$ time to evaluate, it comes with loose guarantees; see Section 3.2. The k-leverage scores are more expensive to evaluate, since they call for a truncated SVD of order k, but they come with tight bounds on the ratio of their expected approximation error over that of PCA.

To minimize approximation error, the subspace spanned by the selected columns should be as large as possible. Simultaneously, the number of selected columns should be as small as possible, so that intuitively, diversity among the selected columns is desirable. The column subset selection problem (CSSP) then becomes a question of designing a discrete point process over the column indices $\{1, \dots, d\}$ that favors diversity in terms of directions covered by the corresponding columns of $\boldsymbol{X}$. Beyond the problem of designing such a point process, guarantees on the resulting approximation error are desirable. Since, given a target dimension $k \leq d$ after projection, PCA provides the best approximation in Frobenius or spectral norm, it is often used a reference: a good CSS algorithm preserves interpretability of the c selected features while guaranteeing an approximation error not much worse than that of rank-k PCA, all of this with c not much larger than k.

In this paper, we introduce and analyze a new randomized algorithm for selecting k diverse columns. Diversity is ensured using a determinantal point process (DPP). DPPs can be viewed as the kernel machine of point processes; they were introduced by Macchi (1975) in quantum optics, and their use widely spread after the 2000s in random matrix theory (Johansson, 2005), machine learning (Kulesza et al., 2012), spatial statistics (Lavancier et al., 2015), and Monte Carlo methods (Bardenet and Hardy, 2016), among others. In a sense, the DPP we propose is a nonindependent generalization of the multinomial sampling with k-leverage scores of (Boutsidis et al., 2009). It further naturally connects to volume sampling, the CSS algorithm that has the best error bounds (Deshpande et al., 2006). We give error bounds for DPP

sampling that exploit sparsity and decay properties of the k-leverage scores, and outperform volume sampling when these properties hold. Our claim is backed up by experiments on toy and real datasets.

The paper is organized as follows. Section 2 introduces our notation. Section 3 is a survey of column subset selection, up to the state of the art to which we later compare. In Section 4, we discuss determinantal point processes and their connection to volume sampling. Section 5 contains our main results, in the form of both classical bounds on the approximation error and risk bounds when CSS is a prelude to linear regression. In Section 6, we numerically compare CSS algorithms, using in particular a routine that samples random matrices with prescibed k-leverage scores.

2 Notation

We use $[n]$ to denote the set $\{1, \dots, n\}$, and $[n : m]$ for $\{n, \dots, m\}$. We use bold capitals $\boldsymbol{A}, \boldsymbol{X}, \dots$ to denote matrices . For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and subsets of indices $I \subset [m]$ and $J \subset [n]$, we denote by $\boldsymbol{A}_{I,J}$ the submatrix of $\boldsymbol{A}$ obtained by keeping only the rows indexed by I and the columns indexed by J. When we mean to take all rows or $\boldsymbol{A}$, we write $\boldsymbol{A}_{:,J}$, and similarly for all columns. We write $\mathrm{rk}(\boldsymbol{A})$ for the rank of $\boldsymbol{A}$, and $\sigma_i(\boldsymbol{A})$, $i = 1, \dots, \mathrm{rk}(\boldsymbol{A})$ for its singular values, ordered decreasingly. Sometimes, we will need the vectors $\Sigma(\boldsymbol{A})$ and $\Sigma(\boldsymbol{A})^2$ the vectors of $\mathbb{R}^d$ with respective entries $\sigma_i(\boldsymbol{A})$ and $\sigma_i^2(\boldsymbol{A})$, $i = 1, \dots, \mathrm{rk}(\boldsymbol{A})$. Similarly, when $\boldsymbol{A}$ can be diagonalized, $\Lambda(\boldsymbol{A})$ (and $\Lambda(\boldsymbol{A})^2$) are vectors with the decreasing eigenvalues (squared eigenvalues) of $\boldsymbol{A}$ as entries.

The spectral norm of $\boldsymbol{A}$ is $\|\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A})$, while the Frobenius norm of $\boldsymbol{A}$ is defined by

$$\|\boldsymbol{A}\|_{\mathrm{Fr}} = \sqrt{\sum_{i=1}^{\mathrm{rk}(\boldsymbol{A})} \sigma_i(\boldsymbol{A})^2}.$$

For $\ell \in \mathbb{N}$, we need to introduce the ℓ-th elementary symmetric polynomial on $L \in \mathbb{N}$ variables, that is

$$e_\ell(X_1, \dots, X_L) = \sum_{\substack{T \subset [L] \\ |T| = \ell}} \prod_{j \in T} X_j. \tag{1}$$

Finally, we follow Ben-Israel (1992) and denote spanned volumes by

$$\mathrm{Vol}_q(\boldsymbol{A}) = \sqrt{e_q\left(\sigma_1(\boldsymbol{A})^2, \dots, \sigma_{\mathrm{rk}(A)}(\boldsymbol{A})^2\right)}, \quad q = 1, \dots, \mathrm{rk}(\boldsymbol{A}).$$

Throughout the paper, $\boldsymbol{X}$ will always denote an $N \times d$ matrix that we think of as the original data matrix, of which we want to select $k \leq d$ columns. Unless otherwise specified, r is the rank of $\boldsymbol{X}$, and matrices $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ are reserved for the SVD of $\boldsymbol{X}$, that is,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{2}$$

$$= \left[\begin{array}{c|c} \boldsymbol{U}_k & \boldsymbol{U}_{r-k} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_k & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma}_{r-k} \end{array}\right] \left[\begin{array}{c} \boldsymbol{V}_k^T \\ \hline \boldsymbol{V}_{r-k}^T \end{array}\right], \tag{3}$$

where $\boldsymbol{U} \in \mathbb{R}^{N \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ are orthogonal, and $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ is diagonal. The diagonal entries of Σ are denoted by $\sigma_i = \sigma_i(\boldsymbol{X})$, $i = 1, \dots, r$, and we assume they

are in decreasing order. We will also need the blocks given in (3), where we separate blocks of size k corresponding to the largest k singular values. To simplify notation, we abusively write $\boldsymbol{U}_k$ for $\boldsymbol{U}_{:,[k]}$ and $\boldsymbol{V}_k$ for $\boldsymbol{V}_{:,[k]}$ in (3), among others. Though they will be introduced and discussed at length in Section 3.3, we also recall here that we note $\ell_i^k = \|\boldsymbol{V}_{[k],i}\|_2^2$ the so-called *k-leverage score* of the i-th column of $\boldsymbol{X}$.

We need some notation for the selection of columns. Let $S \subset [d]$ be such that $|S| = k$, and let $\boldsymbol{S} \in \{0,1\}^{d\times k}$ be the corresponding sampling matrix: $\boldsymbol{S}$ is defined by $\forall \boldsymbol{M} \in \mathbb{R}^{N\times d}, \boldsymbol{M}\boldsymbol{S} = \boldsymbol{M}_{:,S}$. In the context of column selection, it is often referred to $\boldsymbol{X}\boldsymbol{S} = \boldsymbol{X}_{:,S}$ as $\boldsymbol{C}$. We set for convenience $\boldsymbol{Y}_{:,S}^\intercal = (\boldsymbol{Y}_{:,S})^\intercal$.

The result of column subset selection will usually be compared to the result of PCA. We denote by $\Pi_k \boldsymbol{X}$ the best rank-k approximation to $\boldsymbol{X}$. The sense of *best* can be understood either in Frobenius or spectral norm, as both give the same result. On the other side, for a given subset $S \subset [d]$ of size $|S| = s$ and $\nu \in \{2, \mathrm{Fr}\}$, let

$$\Pi_{S,k}^\nu \boldsymbol{X} = \arg\min_A \|\boldsymbol{X} - A\|_\nu$$

where the minimum is taken over all matrices $\boldsymbol{A} = \boldsymbol{X}_{:,S}\boldsymbol{B}$ such that $\boldsymbol{B} \in \mathbb{R}^{s\times d}$ and $\operatorname{rk} \boldsymbol{B} \leq k$; in words, the minimum is taken over matrices of rank at most k that lie in the column space of $\boldsymbol{C} = \boldsymbol{X}_{:,S}$. When $|S| = k$, we simply write $\Pi_S^\nu \boldsymbol{X} = \Pi_{S,k}^\nu \boldsymbol{X}$. In practice, the Frobenius projection can be computed as $\Pi_S^{\mathrm{Fr}} \boldsymbol{X} = \boldsymbol{C}\boldsymbol{C}^+\boldsymbol{X}$, yet there is no simple expression for $\Pi_S^2 \boldsymbol{X}$. However, $\Pi_S^{\mathrm{Fr}} \boldsymbol{X}$ can be used as an approximation of $\Pi_S^2 \boldsymbol{X}$ since

$$\|\boldsymbol{X} - \Pi_S^2 \boldsymbol{X}\|_2 \leq \|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}} \boldsymbol{X}\|_2 \leq \sqrt{2}\|\boldsymbol{X} - \Pi_S^2 \boldsymbol{X}\|_2, \tag{4}$$

see (Boutsidis et al., 2011, Lemma 2.3).

3 Related Work

In this section, we review the main results about column subset selection.

3.1 Rank revealing QR decompositions

The first k-CSSP algorithm can be traced back to the article of Golub (1965) on pivoted QR factorization. This work introduced the concept of Rank Revealing QR factorization (RRQR). The original motivation was to calculate a well-conditioned QR factorization of a matrix $\boldsymbol{X}$ that reveals its numerical rank.

Definition 1 *Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ and $k \in \mathbb{N}$ $(k \leq d)$. A RRQR factorization of $\boldsymbol{X}$ is a 3-tuple $(\boldsymbol{\Pi}, \boldsymbol{Q}, \boldsymbol{R})$ with $\boldsymbol{\Pi} \in \mathbb{R}^{d\times d}$ a permutation matrix, $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$ an orthogonal matrix, and $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ a triangular matrix, such that $\boldsymbol{X}\boldsymbol{\Pi} = \boldsymbol{Q}\boldsymbol{R}$,*

$$\frac{\sigma_k(\boldsymbol{X})}{p_1(k,d)} \leq \sigma_{min}(\boldsymbol{R}_{[k],[k]}) \leq \sigma_k(\boldsymbol{X}), \tag{5}$$

and

$$\sigma_{k+1}(\boldsymbol{X}) \leq \sigma_{max}(\boldsymbol{R}_{[k+1:d],[k+1:d]}) \leq p_2(k,d)\sigma_{k+1}(\boldsymbol{X}), \tag{6}$$

where $p_1(k,d)$ and $p_2(k,d)$ are controlled.

In practice, a RRQR factorization algorithm interchanges pairs of columns and updates or builds a QR decomposition on the fly. The link between RRQR factorization and k-CSSP was first discussed by Boutsidis, Mahoney, and Drineas (2009). The structure of a RRQR factorization indeed gives a deterministic selection of a subset of k columns of $\boldsymbol{X}$. More precisely, if we take $\boldsymbol{C}$ to be the first k columns of $\boldsymbol{X}\boldsymbol{\Pi}$, $\boldsymbol{C}$ is a subset of columns of $\boldsymbol{X}$ and $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_2 = \|\boldsymbol{R}_{[k+1:r],[k+1:r]}\|_2$. By (6), any RRQR algorithm thus provides provable guarantees in spectral norm for k-CSSP.

Following (Golub, 1965), many papers gave algorithms that improved on $p_1(k,d)$ and $p_2(k,d)$ in Definition 1. Table 1 sums up the guarantees of the original algorithm of (Golub, 1965) and the state-of-the-art algorithms of Gu and Eisenstat (1996). Note the dependency of $p_2(k,d)$ on the dimension d through the term $\sqrt{d-k}$; this term is common for guarantees in spectral norm for k-CSSP. We refer to (Boutsidis et al., 2009) for an exhaustive survey on RRQR factorization.

3.2 Length square importance sampling and additive bounds

Drineas, Frieze, Kannan, Vempala, and Vinay (2004) proposed a randomized CSS algorithm based on i.i.d. sampling s indices $S = \{i_1, \dots, i_s\}$ from a multinomial distribution of parameter $\boldsymbol{p}$, where

$$p_j = \frac{\|\boldsymbol{X}_{:,j}\|_2^2}{\|\boldsymbol{X}\|_{\mathrm{Fr}}^2}, j \in [d]. \tag{7}$$

Let $\boldsymbol{C} = \boldsymbol{X}_{:,S}$ be the corresponding submatrix. First, we note that some columns of $\boldsymbol{X}$ may appear more than once in $\boldsymbol{C}$. Second, (Drineas et al., 2004, Theorem 3) states that

$$\mathbb{P}\left(\|\boldsymbol{X} - \Pi_{S,k}^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}^2 + 2(1 + \sqrt{8\log(\frac{2}{\delta})})\sqrt{\frac{k}{s}}\|\boldsymbol{X}\|_{\mathrm{Fr}}^2\right) \geq 1 - \delta. \tag{8}$$

Equation (8) is a high-probability additive upper bound for $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2$. The drawback of such bounds is that they can be very loose if the first k singular values of $\boldsymbol{X}$ are large compared to σ_{k+1}. For this reason, multiplicative approximation bounds have been considered.

3.3 k-leverage scores sampling and multiplicative bounds

Drineas, Mahoney, and Muthukrishnan (2008) proposed an algorithm with provable multiplicative upper bound using multinomial sampling, but this time according to k-leverage scores.

Algorithm	$p_2(k,d)$	Complexity	References
Pivoted QR	$2^k\sqrt{d-k}$	$\mathcal{O}(dNk)$	(Golub and Van Loan, 1996)
Strong RRQR (Alg. 3)	$\sqrt{(d-k)k+1}$	not polynomial	(Gu and Eisenstat, 1996)
Strong RRQR (Alg. 4)	$\sqrt{f^2(d-k)k+1}$	$\mathcal{O}(dNk\log_f(d))$	(Gu and Eisenstat, 1996)

Table 1: Examples of some RRQR algorithms and their theoretical performances.

Definition 2 (k-leverage scores) *Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal \in \mathbb{R}^{N\times d}$ be the SVD of $\boldsymbol{X}$. We note $\boldsymbol{V}_k = \boldsymbol{V}_{:,[k]}$ the first k columns of $\boldsymbol{V}$. For $j \in [d]$, the k-leverage score of the j-th column of $\boldsymbol{X}$ is defined by*

$$\ell_j^k = \sum_{i=1}^{k} V_{i,j}^2. \tag{9}$$

In particular, it holds

$$\sum_{j\in[d]} \ell_j^k = \sum_{j\in[d]} \|(\boldsymbol{V}_k^\intercal)_{:,j}\|_2^2 = \mathrm{Tr}(\boldsymbol{V}_k\boldsymbol{V}_k^\intercal) = k, \tag{10}$$

since $\boldsymbol{V}_k$ is an orthogonal matrix. Therefore, one can consider the multinomial distribution on $[d]$ with parameters

$$p_j = \frac{\ell_j^k}{k} \ , j \in [d]. \tag{11}$$

This multinomial is called the *k-leverage scores distribution*.

Theorem 3 (Drineas et al., 2008, Theorem 3) *If the number s of sampled columns satisfies*

$$s \geq \frac{4000k^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right), \tag{12}$$

then, under the k-leverage scores distribution,

$$\mathbb{P}\left(\|\boldsymbol{X} - \Pi_{S,k}^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq (1+\epsilon)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}^2\right) \geq 1-\delta. \tag{13}$$

Drineas et al. (2008) also considered replacing multinomial with Bernoulli sampling, still using the k-leverage scores. The expected number of columns needed for (13) to hold is then lowered to $\mathcal{O}(\frac{k\log k}{\epsilon^2})$. A natural question is then to understand how low the number of columns can be, while still guaranteeing a multiplicative bound like (13). A partial answer has been given by Deshpande and Vempala (2006).

Proposition 4 (Deshpande and Vempala, 2006, Proposition 4) *Given $\epsilon > 0$, $k, d \in \mathbb{N}$ such that $d\epsilon \geq 2k$, there exists a matrix $\boldsymbol{X}^\epsilon \in \mathbb{R}^{kd\times k(d+1)}$ such that for any $S \subset [d]$,*

$$\|\boldsymbol{X}^\epsilon - \Pi_{S,k}^{\mathrm{Fr}}\boldsymbol{X}^\epsilon\|_{\mathrm{Fr}}^2 \geq (1+\epsilon)\|\boldsymbol{X}^\epsilon - \boldsymbol{X}_k^\epsilon\|_{\mathrm{Fr}}^2. \tag{14}$$

This suggests that the lower bound for the number of columns is k/ϵ, at least in the worst case sense of Proposition 4. Interestingly, the k-leverage scores distribution of the matrix $\boldsymbol{X}^\epsilon$ in the proof of Proposition 4 is uniform, so that k-leverage score sampling boils down to simple uniform sampling.

To match the lower bound of Deshpande and Vempala (2006), Boutsidis, Drineas, and Magdon-Ismail (2011) proposed a greedy algorithm to select columns. This algorithm is inspired by the sparsification of orthogonal matrices proposed in (Batson et al., 2009). The full description of this family of algorithms is beyond the scope of this article. We only recall one of the results of the article.

Theorem 5 (Boutsidis et al., 2011, Theorem 1.5) *There exists a randomized algorithm $\mathcal{A}$ that select at most $c = \frac{2k}{\epsilon}(1 + o(1))$ columns of $\boldsymbol{X}$ such that*

$$\mathbb{E}_{\mathcal{A}} \|\boldsymbol{X} - \Pi^{\mathrm{Fr}}_{S,k}\boldsymbol{X}\|^2_{\mathrm{Fr}} \leq (1+\epsilon)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|^2_{\mathrm{Fr}}. \tag{15}$$

Finally, a deterministic algorithm based on k-leverage score sampling was proposed by Papailiopoulos, Kyrillidis, and Boutsidis (2014). The algorithm selects the $c(\theta)$ columns of $\boldsymbol{X}$ with the largest k-leverage scores, where

$$c(\theta) \in \arg\min_{u} \left(\sum_{i=1}^{u} \ell_i^k > \theta \right), \tag{16}$$

and θ is a free parameter that controls the approximation error. To guarantee that there exists a matrix of rank k in the subspace spanned by the selected columns, Papailiopoulos et al. (2014) assume that

$$0 \leq k - \theta < 1. \tag{17}$$

Loosely speaking, this condition is satisfied for a low value of $c(\theta)$ if the k-leverage scores (after ordering) are decreasing rapidly enough. The authors give empirical evidence that this condition is satisfied by a large proportion of real datasets.

Theorem 6 (Papailiopoulos et al., 2014, Theorem 2) *Let $\epsilon = k - \theta \in [0, 1)$, letting S index the columns with the $c(\theta)$ largest k-leverage scores,*

$$\|\boldsymbol{X} - \Pi^{\nu}_{S,k}\boldsymbol{X}\|_{\nu} \leq \frac{1}{1-\epsilon}\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\nu}, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{18}$$

In particular, if $\epsilon \in [0, \frac{1}{2}]$,

$$\|\boldsymbol{X} - \Pi^{\nu}_{S,k}\boldsymbol{X}\|_{\nu} \leq (1+2\epsilon)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\nu}, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{19}$$

Furthermore, they proved that if the k-leverage scores decay like a power law, the number of columns needed to obtain a multiplicative bound can actually be smaller than $\frac{k}{\epsilon}$.

Theorem 7 (Papailiopoulos et al., 2014, Theorem 3) *Assume, for $\eta > 0$,*

$$\ell_i^k = \frac{\ell_1^k}{i^{\eta+1}}. \tag{20}$$

Let $\epsilon = k - \theta \in [0, 1)$, then

$$c(\theta) = \max\left\{ \left(\frac{4k}{\epsilon}\right)^{\frac{1}{\eta+1}} - 1, \left(\frac{4k}{\eta\epsilon}\right)^{\frac{1}{\eta}}, k \right\}. \tag{21}$$

This complements the fact that the worst case example in Proposition 4 had uniform k-leverage scores. Loosely speaking, matrices with fast decaying k-leverage scores can be efficiently subsampled.

3.4 The geometric interpretation of the k-leverage scores

The k-leverage scores can be given a geometric interpretation, the generalization of which serves as a first motivation for our work.

For $i \in [d]$, let $\boldsymbol{e}_i$ be the i-th canonical basis vector of $\mathbb{R}^d$. Let further θ_i be the angle between $\boldsymbol{e}_i$ and the subspace $\mathcal{P}_k = \mathrm{Span}(\boldsymbol{V}_k)$, and denote by $\Pi_{\mathcal{P}_k}\boldsymbol{e}_i$ the orthogonal projection of $\boldsymbol{e}_i$ onto the subspace $\mathcal{P}_k$. Then

$$\cos^2(\theta_i) := \frac{(\boldsymbol{e}_i, \Pi_{\mathcal{P}_k}\boldsymbol{e}_i)^2}{\|\Pi_{\mathcal{P}_k}\boldsymbol{e}_i\|^2} = (\boldsymbol{e}_i, \Pi_{\mathcal{P}_k}(\boldsymbol{e}_i)) = (\boldsymbol{e}_i, \sum_{j=1}^{k} V_{i,j}\boldsymbol{V}_{:,j}) = \sum_{j=1}^{k} V_{i,j}^2 = \ell_i^k. \quad (22)$$

A large k-leverage score ℓ_i^k thus indicates that $\boldsymbol{e}_i$ is almost aligned with $\mathcal{P}_k$. Selecting columns with large k-leverage scores as in (Drineas et al., 2008) can thus be interpreted as replacing the principal eigenspace $\mathcal{P}_k$ by a subspace that must contain k of the original coordinate axes. Intuitively, a closer subspace to the original $\mathcal{P}_k$ would be obtained by selecting columns *jointly* rather than independently, considering the angle with $\mathcal{P}_k$ of the subspace spanned by these columns. More precisely, consider $S \subset [d], |S| = k$, and denote $\mathcal{P}_S = \mathrm{Span}(\boldsymbol{e}_j, j \in S)$. A natural definition of the cosine between $\mathcal{P}_k$ and $\mathcal{P}_S$ is in terms of the so-called *principal angles* (Golub and Van Loan, 1996, Section 6.4.4); see Appendix C. In particular, Proposition 27 in Appendix C yields

$$\cos^2(\mathcal{P}_k, \mathcal{P}_S) = \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2. \quad (23)$$

This paper is about sampling k columns proportionally to (23).

In Appendix A, we contribute a different interpretation of k-leverage scores and volumes, which relates them to the length-square distribution of Section 3.2.

3.5 Negative correlation: volume sampling and the double phase algorithm

In this section, we survey algorithms that randomly sample exactly k columns from $\boldsymbol{X}$, unlike the multinomial sampling schemes of Sections 3.2 and 3.3, which typically require more than k columns.

Deshpande, Rademacher, Vempala, and Wang (2006) obtained a multiplicative bound on the expected approximation error, with only k columns, using so-called *volume sampling*.

Theorem 8 (Deshpande et al., 2006) *Let S be a random subset of $[d]$, chosen with probability*

$$\mathbb{P}_{\mathrm{VS}}(S) = Z\,\mathrm{Det}(\boldsymbol{X}_{:,S}^\intercal \boldsymbol{X}_{:,S})\mathbb{1}_{\{|S|=k\}}, \quad (24)$$

where $Z = \sum_{|S|=k} \mathrm{Det}(\boldsymbol{X}_{:,S}^\intercal \boldsymbol{X}_{:,S})$. Then

$$\mathbb{E}_{\mathrm{VS}}\,\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq (k+1)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}^2 \quad (25)$$

and

$$\mathbb{E}_{\mathrm{VS}}\,\|\boldsymbol{X} - \Pi_S^2\boldsymbol{X}\|_2^2 \leq (d-k)(k+1)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}^2. \quad (26)$$

Note that the bound for the spectral norm was proven in (Deshpande et al., 2006) for the Frobenius projection, that is, they bound $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_2$. The bound (26) easily follows from (4). Later, sampling according to (24) was shown to be doable in polynomial time (Deshpande and Rademacher, 2010). Using a worst case example, Deshpande et al. (2006) proved that the $k+1$ factor in (25) cannot be improved.

Proposition 9 (Deshpande et al., 2006) *Let $\epsilon > 0$. There exists a $(k+1) \times (k+1)$ matrix $\boldsymbol{X}^\epsilon$ such that for every subset S of k columns of $\boldsymbol{X}^\epsilon$,*

$$\|\boldsymbol{X}^\epsilon - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}^\epsilon\|_{\mathrm{Fr}}^2 > (1-\epsilon)(k+1)\|\boldsymbol{X}^\epsilon - \Pi_k\boldsymbol{X}^\epsilon\|_{\mathrm{Fr}}^2. \tag{27}$$

We note that there has been recent interest in a similar but different distribution called *dual volume sampling* (Avron and Boutsidis, 2013; Li et al., 2017a; Dereziński and Warmuth, 2018), sometimes also termed volume sampling. The main application of dual VS is row subset selection of a matrix $\boldsymbol{X}$ for linear regression on label budget constraints.

(Boutsidis et al., 2009) proposed a k-CSSP algorithm, called *double phase*, that combines ideas from multinomial sampling and RRQR factorization. The motivating idea is that the theoretical performance of RRQR factorizations depends on the dimension through a factor $\sqrt{d-k}$; see Table 1. To improve on that, the authors propose to first reduce the dimension d to c by preselecting a large number of columns $c > k$ using multinomial sampling from the k-leverage scores distribution, as in Section 3.3. Then only, they perform a RRQR factorization of the reduced matrix $\boldsymbol{V}_k^\intercal \boldsymbol{S}_1 \boldsymbol{D}_1 \in \mathbb{R}^{k\times c}$, where $\boldsymbol{S}_1 \in \mathbb{R}^{d\times c}$ is the sampling matrix of the multinomial phase and $\boldsymbol{D}_1 \in \mathbb{R}^{c\times c}$ is a scaling matrix.

Theorem 10 (Boutsidis et al., 2009) *Let S be the output of the double phase algorithm with $c = \Theta(k \log k)$. Then*

$$\mathbb{P}_{\mathrm{DPh}}\left(\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}} \leq \Theta(k\log^{\frac{1}{2}} k)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}\right) \geq 0.8\,. \tag{28}$$

$$\mathbb{P}_{\mathrm{DPh}}\left(\|\boldsymbol{X} - \Pi_S^2\boldsymbol{X}\|_2 \leq \Theta(k\log^{\frac{1}{2}} k)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_2 + \Theta(k^{\frac{3}{4}}\log^{\frac{1}{4}} k)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}\right) \geq 0.8. \tag{29}$$

Note that the spectral norm bound was proven for a slightly different distribution in the randomized phase. Furthermore this bound was proved in (Deshpande et al., 2006) for $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_2$ but using (4) the bound (29) follows. The constants $\Theta(k\log^{\frac{1}{2}} k)$ and $\Theta(k^{\frac{3}{4}}\log^{\frac{1}{4}} k)$ in the bounds (28) and (29) depends on c the number of pre-selected columns in the randomized step. In practice, the choice of the parameter c of the randomized pre-selection phase has an influence on the quality of the approximation. We refer to (Boutsidis et al., 2009) for details.

3.6 Excess risk in sketched linear regression

So far, we have focused on approximation bounds in spectral or Frobenius norm for $\boldsymbol{X} - \Pi_{S,k}^{\nu}\boldsymbol{X}$. This is a reasonable measure of error as long as it is not known what

the practitioner wants to do with the submatrix $\boldsymbol{X}_{:,S}$. In this section, we assume that the ultimate goal is to perform linear regression of some $\mathbf{y} \in \mathbb{R}^{\mathbf{N}}$ onto $\boldsymbol{X}$. Other measures of performance then become of interest, such as the excess risk incurred by regressing onto $\boldsymbol{X}_{:,S}$ rather than $\boldsymbol{X}$. We use here the framework of Slawski (2018), further assuming well-specification for simplicity.

For every $i \in [N]$, assume $y_i = \boldsymbol{X}_{i,:}\boldsymbol{w}^* + \xi_i$, where the noises ξ_i are i.i.d. with mean 0 and variance v. For a given estimator $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{X}, \boldsymbol{y})$, its excess risk is defined as

$$\mathcal{E}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{X}\boldsymbol{w}\|_2^2}{N}\right]. \tag{30}$$

In particular, it is easy to show that the ordinary least squares (OLS) estimator $\hat{\boldsymbol{w}} = \boldsymbol{X}^+\boldsymbol{y}$ has excess risk

$$\mathcal{E}(\hat{\boldsymbol{w}}) = v \times \frac{\mathrm{rk}(\boldsymbol{X})}{N}. \tag{31}$$

Selecting k columns indexed by S in $\boldsymbol{X}$ prior to performing linear regression yields $\boldsymbol{w}_S = (\boldsymbol{X}\boldsymbol{S})^+\boldsymbol{y} \in \mathbb{R}^k$. We are interested in the excess risk of the corresponding sparse vector

$$\hat{\boldsymbol{w}}_S := \boldsymbol{S}\boldsymbol{w}_S = \boldsymbol{S}(\boldsymbol{X}\boldsymbol{S})^+\boldsymbol{y} \in \mathbb{R}^d$$

which has all coordinates zero, except those indexed by S.

Proposition 11 (Theorem 9, Mor-Yosef and Avron, 2018) *Let $S \subset [d]$, such that $|S| = k$. Let $(\theta_i(S))_{i\in[k]}$ be the principal angles between* Span $\boldsymbol{S}$ *and* Span $\boldsymbol{V}_k$*, see Appendix C. Then*

$$\mathcal{E}(\hat{\boldsymbol{w}}_S) \leq \frac{1}{N}\left(1 + \max_{i\in[k]} \tan^2\theta_i(S)\right)\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2 + \frac{vk}{N}. \tag{32}$$

Compared to the excess risk (31) of the OLS estimator, the second term of the right-hand side of (32) replaces $\mathrm{rk}\boldsymbol{X}$ by k. But the price is the first term of the right-hand side of (32), which we loosely term *bias*. To interpret this bias term, we first look at the excess risk of the principal component regressor (PCR)

$$\boldsymbol{w}_k^* \in \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathrm{Span}\,\boldsymbol{V}_k} \mathbb{E}_\xi\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2/N\right]. \tag{33}$$

Proposition 12 (Corollary 11, Mor-Yosef and Avron, 2018)

$$\mathcal{E}(\boldsymbol{w}_k^*) \leq \frac{\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{34}$$

The right-hand side of (34) is almost that of (32), except that the bias term in the CSS risk (32) is larger by a factor that measures how well the subspace spanned by S is aligned with the principal eigenspace $\boldsymbol{V}_k$. This makes intuitive sense: the performance of CSS will match PCR if selecting columns yields almost the same eigenspace.

The excess risk (32) is yet another motivation to investigate DPPs for column subset selection. We shall see in Section 5.2 that the expectation of (32) under a well-chosen DPP for S has a particularly simple bias term.

4 Determinantal Point Processes

In this section, we introduce discrete determinantal point processes (DPPs) and the related k-DPPs, of which volume sampling is an example. DPPs were introduced by Macchi (1975) as probabilistic models for beams of fermions in quantum optics. Since then, DPPs have been thoroughly studied in random matrix theory (Johansson, 2005), and have more recently been adopted in machine learning (Kulesza et al., 2012), spatial statistics Lavancier et al. (2015), and Monte Carlo methods (Bardenet and Hardy, 2016).

4.1 Definitions

For all the definitions in this section, we refer the reader to (Kulesza et al., 2012). Recall that $[d] = \{1, \dots, d\}$.

Definition 13 (DPP) *Let $\boldsymbol{K} \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix. A random subset $Y \subset [d]$ is drawn from a DPP of marginal kernel $\boldsymbol{K}$ if and only if*

$$\forall S \subset [d], \quad \mathbb{P}(S \subset Y) = \mathrm{Det}(\boldsymbol{K}_S), \tag{35}$$

where $\boldsymbol{K}_S = [\boldsymbol{K}_{i,j}]_{i,j \in S}$. We take as a convention $\mathrm{Det}(\boldsymbol{K}_\emptyset) = 1$.

For a given matrix $\boldsymbol{K}$, it is not obvious that (35) consistently defines a point process. One sufficient condition is that $\boldsymbol{K}$ is symmetric and its spectrum is in $[0, 1]$; see (Macchi, 1975) and (Soshnikov, 2000)[Theorem 3]. In particular, when the spectrum of $\boldsymbol{K}$ is included in $\{0, 1\}$, we call $\boldsymbol{K}$ a projection kernel and the corresponding DPP a *projection* DPP[1]. Letting r be the number of unit eigenvalues of its kernel, samples from a projection DPP have fixed cardinality r with probability 1 (Hough et al., 2005, Lemma 17).

For symmetric kernels $\boldsymbol{K}$, a DPP can be seen as a *repulsive* distribution, in the sense that for all $i, j \in [d]$,

$$\begin{aligned} \mathbb{P}(\{i, j\} \subset Y) &= \boldsymbol{K}_{i,i}\boldsymbol{K}_{j,j} - \boldsymbol{K}_{i,j}^2 && (36)\\ &= \mathbb{P}(\{i\} \subset Y)\,\mathbb{P}(\{j\} \subset Y) - \boldsymbol{K}_{i,j}^2 && (37)\\ &\leq \mathbb{P}(\{i\} \subset Y)\,\mathbb{P}(\{j\} \subset Y). && (38) \end{aligned}$$

Besides projection DPPs, there is another natural way of using a kernel matrix to define a random subset of $[d]$ with prespecified cardinality k.

Definition 14 (k-DPP) *Let $\boldsymbol{L} \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix. A random subset $Y \subset [d]$ is drawn from a k-DPP of kernel $\boldsymbol{L}$ if and only if*

$$\forall S \subset [d], \quad \mathbb{P}(Y = S) \propto \mathbb{1}_{\{|S|=k\}} \mathrm{Det}(\boldsymbol{L}_S) \tag{39}$$

where $\boldsymbol{L}_S = [\boldsymbol{L}_{i,j}]_{i,j \in S}$.

DPPs and k-DPPs are closely related but different objects. For starters, k-DPPs are always well defined, provided $\boldsymbol{L}$ has a nonzero minor of size k.

[1] All projection DPPs in this paper have symmetric kernels

4.2 Sampling from a DPP and a k-DPP

Let $\boldsymbol{K} \in \mathbb{R}^{d \times d}$ be a symmetric, positive semi-definite matrix, with eigenvalues in $[0, 1]$, so that $\boldsymbol{K}$ is the marginal kernel of a DPP on $[d]$. Let us diagonalize it as $\boldsymbol{K} = \boldsymbol{V}\mathrm{Diag}(\lambda_i)\boldsymbol{V}^\intercal$. Hough et al. (2005) established that sampling from the DPP with kernel $\boldsymbol{K}$ can be done by *(i)* sampling independent Bernoullis $B_i, i = 1, \ldots, d$, with respective parameters λ_i, *(ii)* forming the submatrix $\boldsymbol{V}_{:,B}$ of $\boldsymbol{V}$ corresponding to columns i such that that $B_i = 1$, and *(iii)* sampling from the projection DPP with kernel

$$\boldsymbol{K}_{\mathrm{proj}} = \boldsymbol{V}_{:,B}\boldsymbol{V}_{:,B}^\intercal.$$

The only nontrivial step is sampling from a projection DPP, for which we give pseudocode in Figure 1; see (Hough et al., 2005, Theorem 7) or (Kulesza et al., 2012, Theorem 2.3) for a proof. For a survey of variants of the algorithm, we also refer to (Tremblay et al., 2018) and the documentation of the DPPy toolbox[2] (Gautier et al., 2018). For our purposes, it is enough to remark that general DPPs are mixtures of projection DPPs of different ranks, and that the cardinality of a general DPP is a sum of independent Bernoulli random variables.

```
ProjectionDPP(K_proj = V V^T)
1   Y ⟵ ∅
2   W ⟵ V
3   while rk(W) > 0
4       Sample i from Ω with probability ∝ ‖W_{i,:}‖_2^2   ▷ Chain rule
5       Y ⟵ Y ∪ {i}
6       V ⟵ V_⊥ an orthonormal basis of Span(V ∩ e_i^⊥)
7   return Y
```

Figure 1: Pseudocode for sampling from a DPP of marginal kernel $\boldsymbol{K}$.

The next proposition establishes that k-DPPs also are mixtures of projection DPPs.

Proposition 15 *(Kulesza et al. (2012, Section 5.2.2)) Let Y be a random subset of $[d]$ sampled from a k-DPP with kernel $\boldsymbol{L}$. We further assume that $\boldsymbol{L}$ is symmetric, we denote its rank by r and its diagonalization by $\boldsymbol{L} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal$. Finally, let $k \leq r$. It holds*

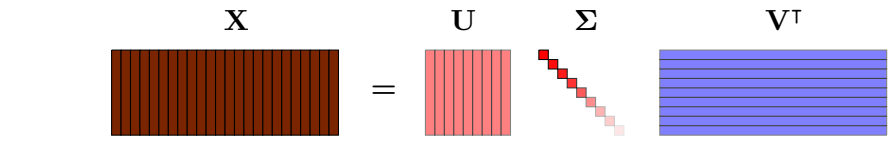
$$\mathbb{P}(Y = S) = \sum_{\substack{T \subset [r] \\ |T| = k}} \mu_T \left[\frac{1}{k!} \mathrm{Det}\left(\boldsymbol{V}_{T,S}\boldsymbol{V}_{T,S}^\intercal\right)\right] \tag{40}$$

where

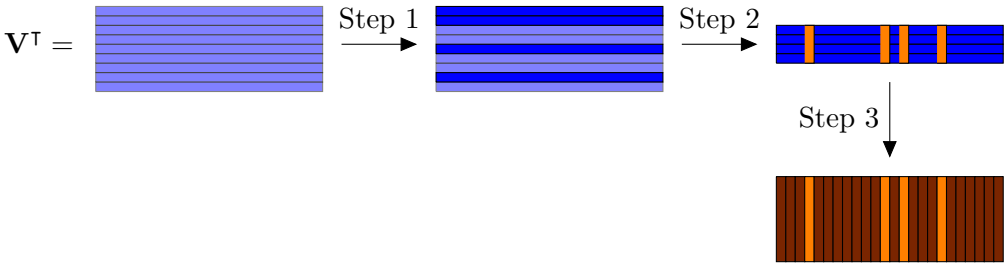
$$\mu_T = \frac{\prod_{i \in T} \lambda_i}{\sum_{\substack{U \subset [r] \\ |U| = k}} \prod_{i \in U} \lambda_i}. \tag{41}$$

[2] http://github.com/guilgautier/DPPy

Each mixture component in square brackets in (40) is a projection DPP with cardinality k. Sampling a k-DPP can thus be done by *(i)* sampling a multinomial distribution with parameters (41), and *(ii)* sampling from the corresponding projection DPP using the algorithm in Figure 1. The main difference between k-DPPs and DPPs is that all mixture components in (40) have the same cardinality k. In particular, projection DPPs are the only DPPs that are also k-DPPs.



(a) SVD of $\boldsymbol{X}$



(b) Sampling k columns according to VS and our DPP

Figure 2: A graphical depiction of the sampling algorithms for volume sampling (VS) and the DPP with marginal kernel $\boldsymbol{V}_k\boldsymbol{V}_k^\intercal$. (a) Both algorithms start with an SVD. (b) In Step 1, VS randomly selects k rows of $\boldsymbol{V}^\intercal$, while the DPP always picks the first k rows. Step 2 is the same for both algorithms: jointly sample k columns of the subsampled $\boldsymbol{V}^\intercal$, proportionally to their squared volume. Step 3 is simply the extraction of the corresponding columns of $\boldsymbol{X}$.

A fundamental example of k-DPP is volume sampling, as defined in Section 3.5. Its kernel is the Gram matrix of the data $\boldsymbol{L} = \boldsymbol{X}^\intercal\boldsymbol{X}$. In general, $\boldsymbol{L}$ is not an orthogonal projection, so that volume sampling is not a DPP.

4.3 Motivations for column subset selection using projection DPPs

volume sampling has been successfully used for column subset selection, see Section 3.5. Our motivation to investigate projection DPPs instead of volume sampling is twofold.

Following (40), volume sampling can be seen as a mixture of projection DPPs indexed by $T \subset [d], |T| = k$, with marginal kernels $\boldsymbol{K}_T = \boldsymbol{V}_{:,T}\boldsymbol{V}_{:,T}^\intercal$ and mixture weights $\mu_T \propto \prod_{i\in T}\sigma_i^2$. The component with the highest weight thus corresponds to the k largest singular values, that is, the projection DPP with marginal kernel $\boldsymbol{K} := \boldsymbol{V}_k\boldsymbol{V}_k^\intercal$. This paper is about column subset selection using precisely this DPP. Alternately, we could motivate the study of this DPP by remarking that its marginals $\mathbb{P}(i \subset Y)$ are the k-leverage scores introduced in Section 3.3. Since $\boldsymbol{K}$ is symmetric, this DPP can be seen as a repulsive generalization of leverage score sampling.

Finally, we recap the difference between volume sampling and the DPP with kernel $\boldsymbol{K}$ with a graphical depiction in Figure 2 of the two procedures to sample

from them that we introduced in Section 4.2. Figure 2 is another illustration of the decomposition of volume sampling as a mixture of projection DPPs.

5 Main Results

In this section, we prove bounds for $\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^{\nu}\boldsymbol{X}\|_{\nu}$ under the projection DPP of marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}$ presented in Section 4. Throughout, we compare our bounds to the state-of-the-art bounds of volume sampling obtained by Deshpande et al. (2006); see Theorem 8 and Section 3.5. For clarity, we defer the proofs of our results from this section to Appendix D.

5.1 Multiplicative bounds in spectral and Frobenius norm

Let S be a random subset of k columns of $\boldsymbol{X}$ chosen with probability:

$$\mathbb{P}_{\mathrm{DPP}}(S) = \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2. \tag{42}$$

First, without any further assumption, we have the following result.

Proposition 16 *Under the projection DPP of marginal kernel $\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}$, it holds*

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^{\nu}\boldsymbol{X}\|_{\nu}^2 \leq k(d+1-k)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\nu}^2, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{43}$$

For the spectral norm, the bound is practically the same as that of volume sampling (26). However, our bound for the Frobenius norm is worse than (25) by a factor $(d-k)$. In the rest of this section, we sharpen our bounds by taking into account the sparsity level of the k-leverage scores and the decay of singular values.

In terms of sparsity, we first replace the dimension d in (43) by the number $p \in [d]$ of nonzero k-leverage scores

$$p = \left|\{i \in [d], \boldsymbol{V}_{i,[k]} \neq \boldsymbol{0}\}\right|. \tag{44}$$

To quantify the decay of the singular values, we define the flatness parameter

$$\beta = \boldsymbol{\sigma}_{k+1}^2 \left(\frac{1}{d-k}\sum_{j\geq k+1} \boldsymbol{\sigma}_j^2\right)^{-1}. \tag{45}$$

In words, $\beta \in [1, d-k]$ measures the flatness of the spectrum of $\boldsymbol{X}$ above the cut-off at k. Indeed, (45) is the ratio of the largest term in a sum to that sum. The closer β is to 1, the more similar the terms in the sum in the denominator of (45). At the extreme, $\beta = d-k$ when $\sigma_{k+1}^2 > 0$ while $\sigma_j^2 = 0$, $\forall j \geq k+2$.

Proposition 17 *Under the projection DPP of marginal kernel $\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}$, it holds*

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^2\boldsymbol{X}\|_2^2 \leq k(p-k)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_2^2 \tag{46}$$

and

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \left(1 + \beta\frac{p-k}{d-k}k\right)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}^2. \tag{47}$$

The bound in (46) compares favorably with volume sampling (26) since the dimension d has been replaced by the sparsity level p. For β close to 1, the bound in (47) is better than the bound (25) of volume sampling since $(p-k)/(d-k) \leq 1$. Again, the sparser the k-leverage scores, the smaller the bounds.

Now, one could argue that, in practice, sparsity is never exact: it can well be that $p = d$ while there still are a lot of small k-leverage scores. We will demonstrate in Section 6 that the DPP still performs better than volume sampling in this setting, which Proposition 17 doesn't reflect. We introduce two ideas to further tighten the bounds of Proposition 17. First, we define an effective sparsity level in the vein of Papailiopoulos et al. (2014), see Section 3.3. Second, we condition the DPP on a favourable event with controlled probability.

Theorem 18 *Let π be a permutation of $[d]$ such that leverage scores are reordered*

$$\ell^k_{\pi_1} \geq \ell^k_{\pi_2} \geq \ldots \geq \ell^k_{\pi_d}. \tag{48}$$

For $\delta \in [d]$, let $T_\delta = [\pi_\delta, \ldots, \pi_d]$. Let $\theta > 1$ and

$$p_{\text{eff}}(\theta) = \min \left\{ q \in [d] \;\middle/\; \sum_{i \leq q} \ell^k_{\pi_i} \geq k - 1 + \frac{1}{\theta} \right\}. \tag{49}$$

Finally, let $\mathcal{A}_\theta$ be the event $\{S \cap T_{p_{\text{eff}}(\theta)} = \emptyset\}$. Then, on the one hand,

$$\mathbb{P}_{\text{DPP}}\left(\mathcal{A}_\theta\right) \geq \frac{1}{\theta}, \tag{50}$$

and, on the other hand,

$$\mathbb{E}_{\text{DPP}}\left[\|\boldsymbol{X} - \Pi_S^2 \boldsymbol{X}\|_2^2 \,\middle|\, \mathcal{A}_\theta\right] \leq (p_{\text{eff}}(\theta) - k + 1)(k - 1 + \theta)\, \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_2^2 \tag{51}$$

and

$$\mathbb{E}_{\text{DPP}}\left[\|\boldsymbol{X} - \Pi_S^{\text{Fr}} \boldsymbol{X}\|_{\text{Fr}}^2 \,\middle|\, \mathcal{A}_\theta\right] \leq \left(1 + \beta \frac{(p_{\text{eff}}(\theta) + 1 - k)}{d - k}(k - 1 + \theta)\right) \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_{\text{Fr}}^2. \tag{52}$$

In Theorem 18, the effective sparsity level $p_{\text{eff}}(\theta)$ replaces the sparsity level p of Proposition 17. The key is to condition on S not containing any index of column with too small a k-leverage score, that is, the event $\mathcal{A}_\theta$. In practice, this is achieved by rejection sampling: we repeatedly and independently sample $S \sim \text{DPP}(\boldsymbol{K})$ until $S \cap T_{p_{\text{eff}}}(\theta) = \emptyset$.

The caveat of any rejection sampling procedure is a potentially large number of samples required before acceptance. But in the present case, Equation (50) guarantees that the expectation of that number of samples is less than θ. The free parameter θ thus interestingly controls both the "energy" threshold in (49), and the complexity of the rejection sampling. The approximation bounds suggest picking θ close to 1, which implies a compromise with the value of $p_{\text{eff}}(\theta)$ that should not be too large either. We have empirically observed that the performance of the DPP is relatively insensitive to the choice of θ.

5.2 Bounds for the excess risk in sketched linear regression

In Section 3.6, we surveyed bounds on the excess risk of ordinary least squares estimators that relied on a subsample of the columns of $\boldsymbol{X}$. Importantly, the generic bound (32) of Mor-Yosef and Avron (2018) has a bias term that depends on the maximum squared tangent of the principal angles between $\mathrm{Span}(\boldsymbol{S})$ and $\mathrm{Span}(\boldsymbol{V}_k)$. When $|S| = k$, this quantity is hard to control without making strong assumptions on the matrix $\boldsymbol{V}_k$. But it turns out that, in expectation under the same DPP as in Section 5.1, this bias term drastically simplifies.

Proposition 19 *We use the notation of Section 3.6. Under the projection DPP with marginal kernel $\boldsymbol{V}_k\boldsymbol{V}_k^\intercal$, it holds*

$$\mathbb{E}_{\mathrm{DPP}}\left[\mathcal{E}(\boldsymbol{w}_S)\right] \leq \big(1 + k(p-k)\big)\frac{\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{53}$$

The sparsity level p appears again in the bound (53): The sparser the k-leverage scores distribution, the smaller the bias term. The bound (53) only features an additional $(1 + k(p-k))$ factor in the bias term, compared to the bound obtained by Mor-Yosef and Avron (2018) for PCR, see Proposition 12. Loosely speaking, this factor is to be seen as the price we accept to pay in order to get more interpretable features than principal components in the linear regression problem. Finally, a natural question is to investigate the choice of k to minimize the bound in (53), but this is out of the scope of this paper.

As in Theorem 18, for practical purposes, it can be desirable to bypass the need for the exact sparsity level p in Proposition 19. We give a bound that replaces p with the effective sparsity level $p_{\mathrm{eff}}(\theta)$ introduced in (49).

Theorem 20 *Using the notation of Section 3.6 for linear regression, and of Theorem 18 for leverage scores and their indices, it holds*

$$\mathbb{E}_{\mathrm{DPP}}\left[\mathcal{E}(\hat{\boldsymbol{w}}_S)\,\middle|\,\mathcal{A}_\theta\right] \leq \left[1 + \big(k - 1 + \theta\big)\big(p_{\mathrm{eff}}(\theta) - k + 1\big)\right]\frac{\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{54}$$

In practice, the same rejection sampling routine as in Theorem 18 can be used to sample conditionally on $\mathcal{A}_\theta$. Finally, to the best of our knowledge, bounding the excess risk in linear regression has not been investigated under volume sampling.

In summary, we have obtained two sets of results. We have proven a set of multiplicative bounds in spectral and Frobenius norm for $\mathbb{E}_{\mathrm{DPP}}\|\boldsymbol{X} - \Pi_S^\nu\boldsymbol{X}\|_\nu$, $\nu \in \{2, \mathrm{Fr}\}$, under the projection DPP of marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k\boldsymbol{V}_k^\intercal$, see Propositions 16 & 17 and Theorem 18. As far as the linear regression problem is concerned, we have proven bounds for the excess risk in sketched linear regression, see Proposition 19 and Theorem 20.

6 Numerical experiments

In this section, we empirically compare our algorithm to the state of the art in column subset selection. In Section 6.1, the projection DPP with kernel $\boldsymbol{K} = \boldsymbol{V}_k\boldsymbol{V}_k^\intercal$ and

volume sampling are compared on toy datasets. In Section 6.2, several column subset selection algorithms are compared to the projection DPP on four real datasets from genomics and text processing. In particular, the numerical simulations demonstrate the favorable influence of the sparsity of the k-leverage scores on the performance of our algorithm both on toy datasets and real datasets. Finally, we packaged all CSS algorithms in this section in a Python toolbox[3].

6.1 Toy datasets

This section is devoted to comparing the expected approximation error $\mathbb{E}\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2$ for the projection DPP and volume sampling. We focus on the Frobenius norm to avoid effects to due different choices of the projection Π_S^{Fr}, see (4).

In order to be able to evaluate the expected errors *exactly*, we generate matrices of low dimension ($d = 20$) so that the subsets of $[d]$ can be exhaustively enumerated. Furthermore, to investigate the role of leverage scores and singular values on the performance of CSS algorithms, we need to generate datasets $\boldsymbol{X}$ with prescribed spectra and k-leverage scores.

6.1.1 Generating toy datasets

Recall that the SVD of $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ reads $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal$, where $\boldsymbol{\Sigma}$ is a diagonal matrix and $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices. To sample a matrix $\boldsymbol{X}$, $\boldsymbol{U}$ is first drawn from the Haar measure of $\mathcal{O}_N(\mathbb{R})$, then $\boldsymbol{\Sigma}$ is chosen among a few deterministic diagonal matrices that illustrate various spectral properties. Sampling the matrix $\boldsymbol{V}$ is trickier. The first k columns of $\boldsymbol{V}$ are structured as follows: the number of non vanishing rows of $\boldsymbol{V}_k$ is equal to p and the norms of the nonvanishing rows are prescribed by a vector $\boldsymbol{\ell}$. By considering the matrix $\boldsymbol{K} = \boldsymbol{V}_k\boldsymbol{V}_k^\intercal$, generating $\boldsymbol{V}$ boils down to the simulation of an Hermitian matrix with prescribed diagonal and spectrum (in this particular case the spectrum is included in $\{0, 1\}$). For this reason, we propose an algorithm that takes as input a leverage scores profile $\boldsymbol{\ell}$ and a spectrum $\boldsymbol{\sigma}^2$, and outputs a corresponding random orthogonal matrix $\boldsymbol{X}$; see Appendix E. This algorithm is a randomization[4] of the algorithm proposed by Fickus, Mixon, Poteet, and Strawn (2011b). Finally, the matrix $\boldsymbol{V}_k \in \mathbb{R}^{d\times k}$ is completed by applying the Gram-Schmidt procedure to $d - k$ additional i.i.d. unit Gaussian vectors, resulting in a matrix $\boldsymbol{V} \in \mathbb{R}^{d\times d}$. Figure 3 summarizes the algorithm proposed to generate matrices $\boldsymbol{X}$ with a k-leverage scores profile $\boldsymbol{\ell}$ and a sparsity level p.

6.1.2 volume sampling vs projection DPP

This section sums up the results of numerical simulations on toy datasets. The number of observations is fixed to $N = 100$, the dimension to $d = 20$, and the number of selected columns to $k \in \{3, 5\}$. Singular values of are chosen from the following profiles: a spectrum with a cutoff called the projection spectrum,

$$\boldsymbol{\Sigma}_{k=3,\mathrm{proj}} = 100\sum_{i=1}^{3}\boldsymbol{e}_i\boldsymbol{e}_i^\intercal + 0.1\sum_{i=4}^{20}\boldsymbol{e}_i\boldsymbol{e}_i^\intercal,$$

[3] http://github.com/AyoubBelhadji/CSSPy
[4] http://github.com/AyoubBelhadji/FrameBuilder

MatrixGenerator$(\boldsymbol{\ell}, \boldsymbol{\Sigma})$

1 Sample $\boldsymbol{U}$ from the Haar measure $\mathbb{O}_N(\mathbb{R})$.
2 Pick $\boldsymbol{\Sigma}$ a diagonal matrix.
3 Pick $p \in [k+1:d]$.
4 Pick a k-leverage-scores profile $\boldsymbol{\ell} \in \mathbb{R}_+^d$ with a sparsity level p.
5 Generate a matrix $\boldsymbol{V}_k$ with the k-leverage-scores profile $\boldsymbol{\ell}$.
6 Extend the matrix $\boldsymbol{V}_k$ to an orthogonal matrix $\boldsymbol{V}$.
7 **return** $\boldsymbol{X} \longleftarrow \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal$

Figure 3: The pseudocode of the algorithm generating a matrix $\boldsymbol{X}$ with prescribed profile of k-leverage scores.

$$\boldsymbol{\Sigma}_{k=5,\text{proj}} = 100 \sum_{i=1}^{5} \boldsymbol{e}_i \boldsymbol{e}_i^\intercal + 0.1 \sum_{i=6}^{20} \boldsymbol{e}_i \boldsymbol{e}_i^\intercal.$$

and a smooth spectrum

$$\boldsymbol{\Sigma}_{k=3,\text{smooth}} = 100 \boldsymbol{e}_1 \boldsymbol{e}_1^\intercal + 10 \boldsymbol{e}_2 \boldsymbol{e}_2^\intercal + \boldsymbol{e}_3 \boldsymbol{e}_3^\intercal + 0.1 \sum_{i=4}^{20} \boldsymbol{e}_i \boldsymbol{e}_i^\intercal,$$

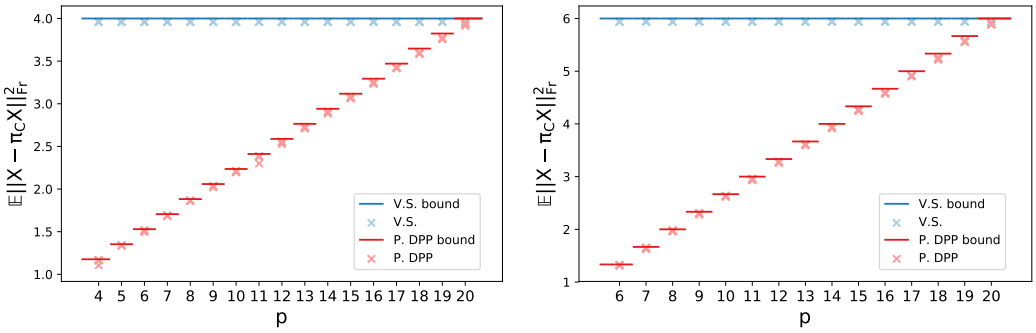
$$\boldsymbol{\Sigma}_{k=5,\text{smooth}} = 10000 \boldsymbol{e}_1 \boldsymbol{e}_1^\intercal + 1000 \boldsymbol{e}_2 \boldsymbol{e}_2^\intercal + 100 \boldsymbol{e}_3 \boldsymbol{e}_3^\intercal + 10 \boldsymbol{e}_4 \boldsymbol{e}_4^\intercal + \boldsymbol{e}_5 \boldsymbol{e}_5^\intercal + 0.1 \sum_{i=6}^{20} \boldsymbol{e}_i \boldsymbol{e}_i^\intercal.$$

Note that all profiles satisfy $\beta = 1$; see (45). In each experiment, for each spectrum, we sample 200 independent leverage scores profiles that satisfy the sparsity constraints from a Dirichlet distribution with concentration parameter 1 and equal means. For each leverage scores profile, we sample a matrix $\boldsymbol{X}$ from the algorithm in Appendix E.

Figure 4 compares, on the one hand, the theoretical bounds in Theorem 8 for volume sampling and Proposition 17 for the projection DPP, to the numerical evaluation of the expected error for sampled toy datasets on the other hand. The x-axis indicates various sparsity levels p. The unit on the y-axis is the error of PCA. There are 400 crosses on each subplot: each of the 200 matrices appears once for both algorithms. The 200 matrices are spread evenly across the values of p.

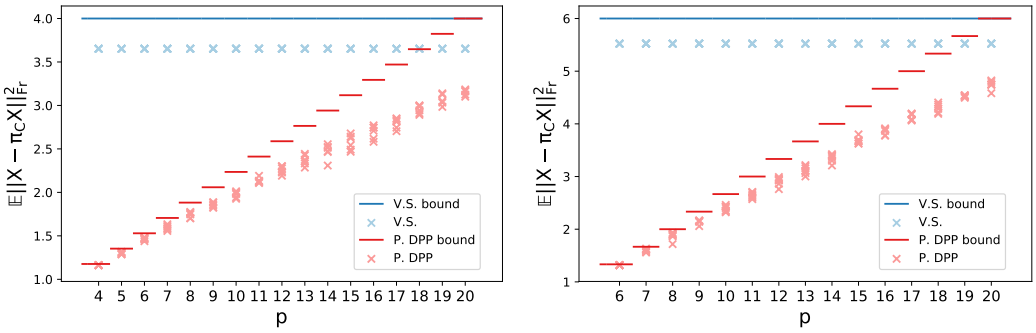
The VS bounds in $(k+1)$ are independent of p. They appear to be tight for projection spectra, and looser for smooth spectra. For the projection DPP, the bound $(k+1)\frac{p-k}{d-k}$ is linear in p, and can be much lower than the bound of VS. The numerical evaluations of the error also suggest that this DPP bound is tight for a projection spectrum, and looser in the smooth case. In both cases, the bound is representative of the actual behavior of the algorithm.

The bottom row of Figure 4 displays the same results for identity spectra, again for $k = 3$ and $k = 5$. This setting is extremely nonsparse and represents an arbitrarily bad scenario where even PCA would not make much practical sense. Both VS and DPP sampling perform constantly badly, and all crosses superimpose at $y = 1$, which indicates the PCA error. In this particular case, our linear bound in



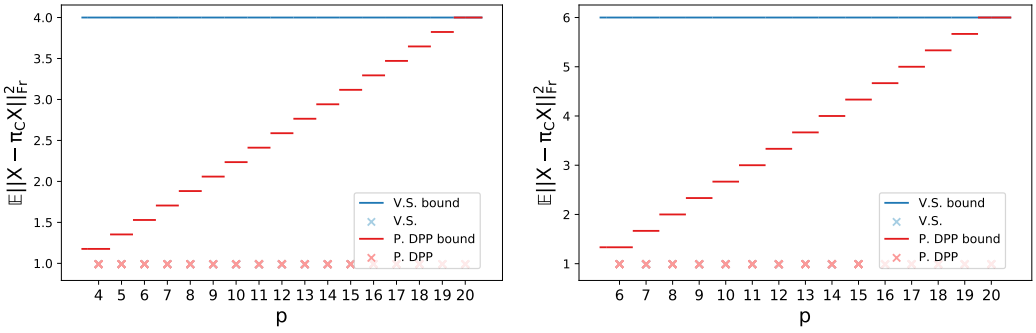
(a) $\Sigma_{3,\text{proj}}$, $k = 3$

(d) $\Sigma_{5,\text{proj}}$, $k = 5$



(b) $\Sigma_{3,\text{smooth}}$, $k = 3$

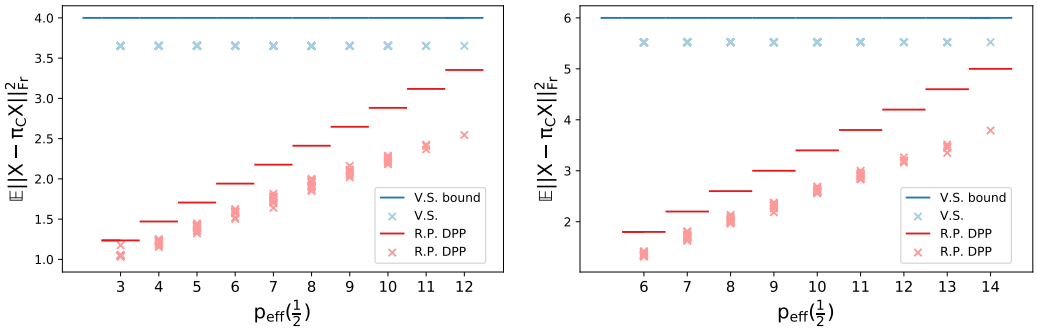
(e) $\Sigma_{5,\text{smooth}}$, $k = 5$



(d) I_{20}, $k = 3$

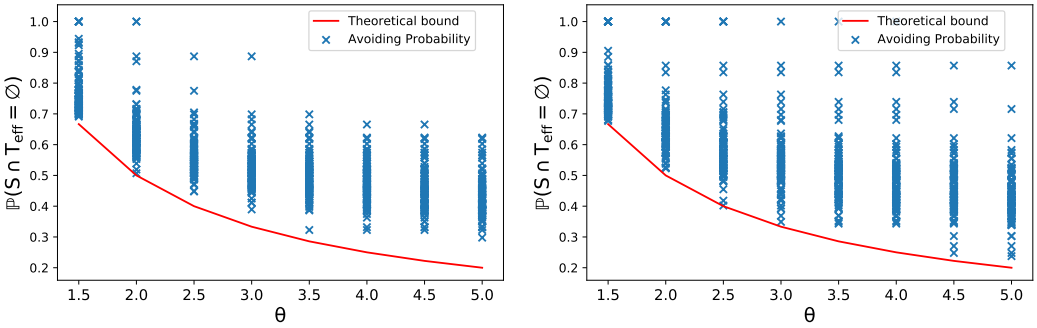
(f) I_{20}, $k = 5$

Figure 4: Realizations and bounds for $\mathbb{E}\|\boldsymbol{X} - \Pi_S^{\text{Fr}}\boldsymbol{X}\|_{\text{Fr}}^2$ as a function of the sparsity level p.



(a) $\Sigma_{3,\text{smooth}}$ (b) $\Sigma_{5,\text{smooth}}$

Figure 5: Realizations and bounds for $\mathbb{E}\|\boldsymbol{X} - \Pi_S^{\text{Fr}}\boldsymbol{X}\|_{\text{Fr}}^2$ as a function of the effective sparsity level $p_{\text{eff}}(\frac{1}{2})$.



(a) $\Sigma_{3,\text{smooth}}$ (b) $\Sigma_{5,\text{smooth}}$

Figure 6: Realizations and bounds for the avoiding probability $\mathbb{P}(S \cap T_{p_{\text{eff}}(\theta)} = \emptyset)$ in Theorem 18 as a function of θ.

p is not representative of the actual behavior of the error. This observation can be explained for volume sampling using Theorem 25, which states that the expected squared error under VS is Schur-concave, and is thus minimized for flat spectra. We have no similar result for the projection DPP.

Figure 5 provides a similar comparison for the two smooth spectra $\boldsymbol{\Sigma}_{3,\text{smooth}}$ and $\boldsymbol{\Sigma}_{5,\text{smooth}}$, but this time using the effective sparsity level $p_{\text{eff}}(\theta)$ introduced in Theorem 18. We use $\theta = 1/2$; qualitatively, we have observed the results to be robust to the choice of θ. The 200 sampled matrices are now unevenly spread across the x-axis, since we do not control $p_{\text{eff}}(\theta)$. Note finally that the DPP here is conditioned on the event $\{S \cap T_{p_{\text{eff}}(\theta)} = \emptyset\}$, and sampled using an additional rejection sampling routine as detailed below Theorem 18.

For the DPP, the bound is again linear on the effective sparsity level $p_{\text{eff}}(\frac{1}{2})$, and can again be much lower than the VS bound. The behavior of both VS and the projection DPP are similar to the exact sparsity setting of Figure 4: the DPP has uniformly better bounds and actual errors, and the bound reflects the actual behavior, if relatively loosely when $p_{\text{eff}}(1/2)$ is large.

Figure 6 compares the theoretical bound in Theorem 18 for the avoiding probability $\mathbb{P}(S \cap T_{p_{\text{eff}}(\theta)} = \emptyset)$ with 200 realizations, as a function of θ. More precisely, we drew 200 matrices $\boldsymbol{X}$, and then for each $\boldsymbol{X}$, we computed exactly – by enumeration – the value $\mathbb{P}(S \cap T_{p_{\text{eff}}(\theta)} = \emptyset)$ for all values of θ. The only randomness is thus in the sampling of $\boldsymbol{X}$, not the evaluation of the probability. The results suggest again that the bound is relatively tight.

6.2 Real datasets

This section compares the empirical performances of several column subset selection algorithms on the datasets in Table 2.

Dataset	Application domain	$N \times d$	References
Colon	genomics	62×2000	(Alon et al., 1999)
Leukemia	genomics	72×7129	(Golub et al., 1999)
Basehock	text processing	1993×4862	(Li et al., 2017b)
Relathe	text processing	1427×4322	(Li et al., 2017b)

Table 2: Datasets used in the experimental section.

These datasets are illustrative of two extreme situations regarding the sparsity of the k-leverage scores. For instance, the dataset Basehock has a very sparse profile of k-leverage scores, while the dataset Colon has a quasi-uniform distribution of k-leverage scores, see Figures 7 (a) & (b).

We consider the following algorithms presented in Section 3: 1) the projection DPP with marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$, 2) volume sampling, using the implementation proposed by Kulesza and Taskar (2011), 3) deterministically picking the largest k-leverage scores, 4) pivoted QR as in (Golub, 1965), although the only known bounds for this algorithm are for the spectral norm, and 5) double phase, with c manually tuned to optimize the performance, usually around $c \approx 10k$.

The rest of Figure 7 sums up the empirical results of the previously described algorithms on the Colon and Basehock datasets. Figures 7 (c) & (d) illustrate the results of the five algorithms in the following setting. An ensemble of 50 subsets are sampled from each algorithm. We give the corresponding boxplots for the Frobenius errors, on Colon and Basehock respectively. We observe that the increase in performance using projection DPP compared to volume sampling is more important for the Basehock dataset than for the Colon dataset: this improvement can be explained by the sparsity of the k-leverage scores as predicted by our approximation bounds. Deterministic methods (largest leverage scores and pivoted QR) perform well compared with other algorithms on the Basehock dataset; in contrast, they display very bad performances on the Colon dataset. The double phase algorithm has the best results on both datasets. However its theoretical guarantees cannot predict such an improvement, as noted in Section 3. The performance of the projection DPP is comparable to those Double Phase and makes it a close second, with a slightly larger gap on the Colon dataset. We emphasize that our approximation bounds are sharp compared to numerical observations.

Figures 7 (e) & (f) show results obtained using a classical boosting technique for randomized algorithms. We repeat 20 times: sample 50 subsets and take the

best subset selection. Displayed boxplots are for these 20 best results. The same comments apply as without boosting.

Figure 8 calls for similar comments, comparing this time the datasets Relathe (with concentrated profile of k-leverage scores) and Leukemia (with almost uniform profile of k-leverage scores).

6.3 Discussion

The performance of our algorithm has been compared to state-of-the-art column subset selection algorithms. We emphasize that the theoretical performances of the proposed approach take into account the sparsity of the k-leverage scores as in Proposition 17 or their fast decrease as in Proposition 18, and that the bounds are in good agreement with the actual behavior of the algorithm. In contrast, state-of-the-art algorithms like volume sampling have looser bounds and worse performances, or like double phase display great performance but have overly pessimistic theoretical bounds.

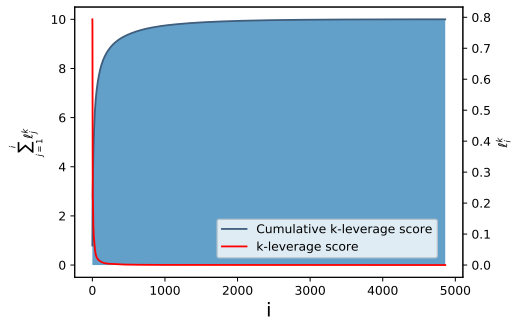
7 Conclusion

We have proposed, analyzed, and empirically investigated a new randomized column subset selection (CSS) algorithm. The crux of our algorithm is a discrete determinantal point process (DPP) that selects a diverse set of k columns of a matrix $\boldsymbol{X}$. This DPP is tailored to CSS through its parametrization by the marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$, where $\boldsymbol{V}_k$ are the first k right singular vectors of the matrix $\boldsymbol{X}$. This specific kernel is related to volume sampling, the state-of-the-art for CSS guarantees in Frobenius and spectral norm.

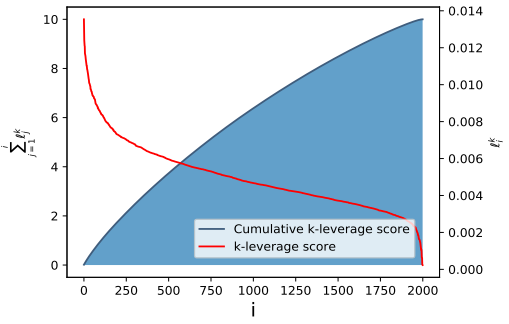
We have identified generic conditions on the matrix $\boldsymbol{X}$ under which our algorithm has bounds that improve on volume sampling. In particular, our bounds highlight the importance of the sparsity and the decay of the k-leverage scores on the approximation performance of our algorithm. This resonates with the compressed sensing literature. We have further numerically illustrated this relation to the sparsity and decay of the k-leverage scores using toy and real datasets. In these experiments, our algorithm performs comparably to the so-called double phase algorithm, which is the empirical state-of-the-art for CSS despite more conservative theoretical guarantees than volume sampling. Thus, our DPP sampling inherits both favorable bounds and increased empirical performance under sparsity or fast decay of the k-leverage scores.

In terms of computational cost, our algorithms scale with the cost of finding the k first right singular vectors, which is currently the main bottleneck. In the line of (Drineas et al., 2012) and (Boutsidis et al., 2011), where the authors estimates the k-leverage scores using random projections, we plan to investigate the impact of random projections to estimate the full matrix $\boldsymbol{K}$ on the approximation guarantees of our algorithms.

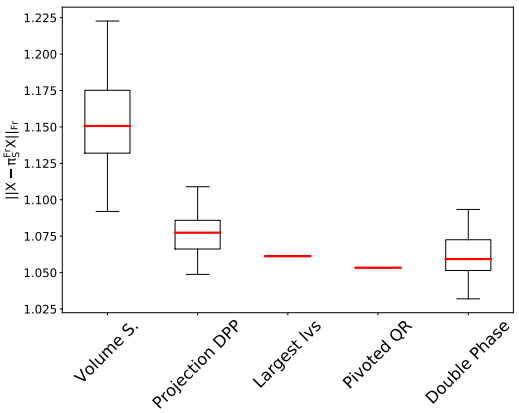
Although often studied as an independent task, in practice CSS is often a prelude to a learning algorithm. We have considered linear regression and we have given a bound on the excess risk of a regression performed on the selected columns only. In particular, sparsity and decay of the k-leverage scores are again involved: the more



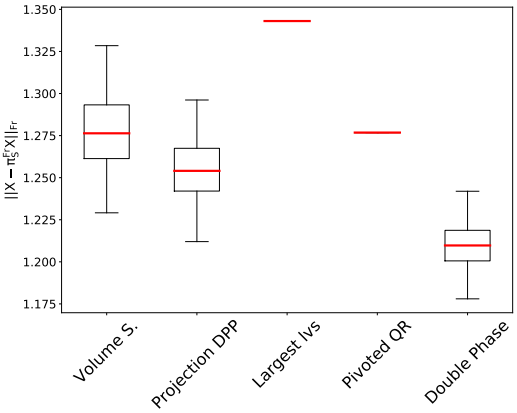
(a) k-leverage scores profile for the dataset Basehock (k=10).



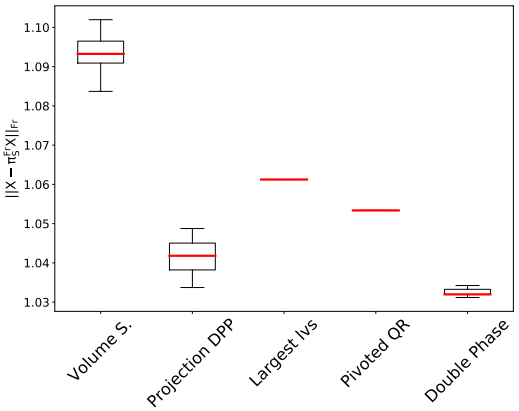
(b) k-leverage scores profile for the dataset Colon (k=10).



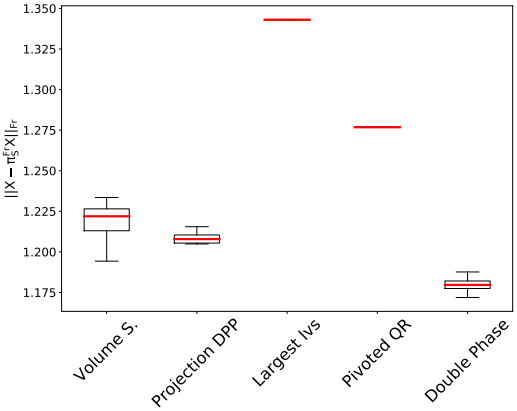
(c) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Basehock (k=10).



(d) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Colon (k=10).

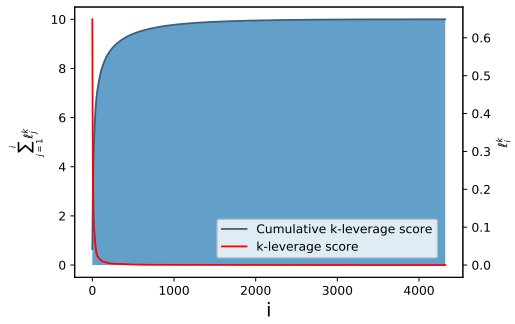


(e) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Basehock (k=10).

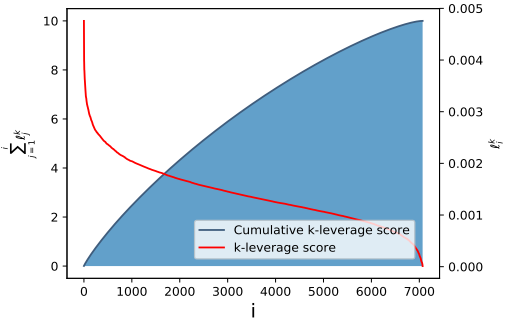


(f) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Colon (k=10).

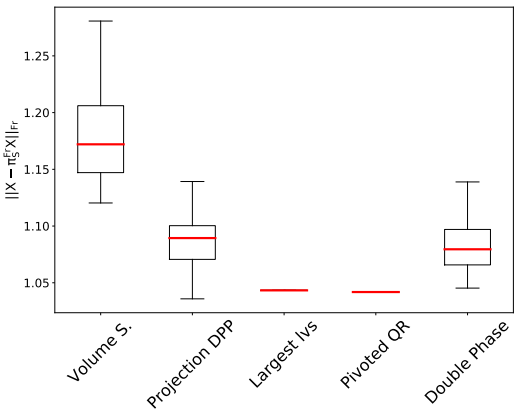
Figure 7: Comparison of several column subset selection algorithms for two datasets: Basehock and Colon.



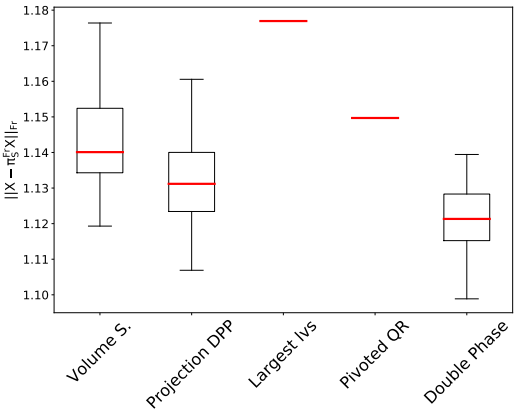
(a) k-leverage scores profile for the dataset Relathe (k=10).



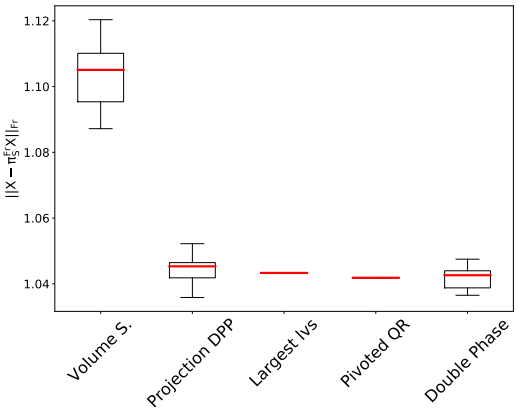
(b) k-leverage scores profile for the dataset Leukemia (k=10).



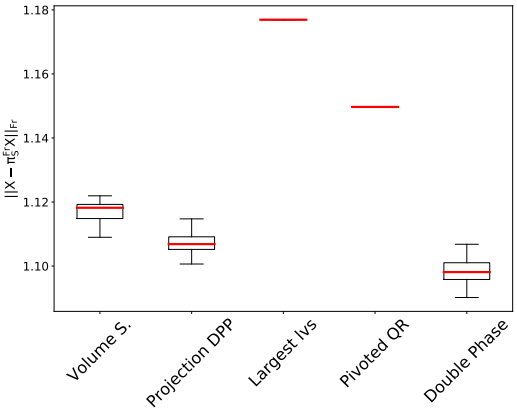
(c) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Relathe (k=10).



(d) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the five algorithms on the dataset Leukemia (k=10).



(e) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Relathe (k=10).



(f) Boxplots of $\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}$ on a batch of 50 samples for the boosting of randomized algorithms on the dataset Leukemia (k=10).

Figure 8: Comparison of several column subset selection algorithms for two datasets: Relathe and Leukemia.

localized the k-leverage scores, the smaller the excess risk bounds. Such an analysis of the excess risk in regression highlights the interest of the proposed approach since it would be difficult to conduct for both volume sampling or the double phase algorithms. Future work in this direction includes investigating the importance of the sparsity of the k-leverage scores on the performance of other learning algorithms such as spectral clustering or support vector machines.

Finally, in our experimental section, we used an adhoc randomized algorithm inspired by (Fickus et al., 2011b) to sample toy datasets with a prescribed profile of k-leverage scores. An interesting question would be to characterize the distribution of the output of our algorithm. In particular, sampling from the uniform measure on the set of symmetric matrices with prescribed spectrum and leverage scores is still an open problem (Dhillon et al., 2005).

Acknowledgments

AB and RB acknowledge support from ANR grant BoB (ANR-16-CE23-0003), and all authors acknowledge support from ANR grant BNPSI (ANR-13-BS03-0006).

References

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

H. Avron and C. Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

R. Bardenet and A. Hardy. Monte Carlo with Determinantal Point Processes. *ArXiv e-prints*, May 2016.

Y. Baryshnikov. GUEs and queues. *Probability Theory and Related Fields*, 119(2): 256–274, 2001.

J. D. Batson, D. A. Spielman, and N. Srivastava. Twice-ramanujan sparsifiers. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 255–262, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-506-2. doi: 10.1145/1536414.1536451. URL http://doi.acm.org/10.1145/1536414.1536451.

Adi Ben-Israel. A volume associated with m x n matrices. *Linear Algebra and its Applications*, 167:87 – 111, 1992. ISSN 0024-3795. doi: http://dx.doi.org/10.1016/0024-3795(92)90340-G. URL http://www.sciencedirect.com/science/article/pii/002437959290340G.

Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, 1973.

C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Twentieth*

Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '09, pages 968–977, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=1496770.1496875.

C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near optimal column-based matrix reconstruction. In *Proceedings of the 2011 IEEE 52Nd Annual Symposium on Foundations of Computer Science*, FOCS '11, pages 305–314, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-0-7695-4571-4. doi: 10.1109/FOCS.2011.21. URL http://dx.doi.org/10.1109/FOCS.2011.21.

M. Dereziński and M. K. Warmuth. Reverse iterative volume sampling for linear regression. *arXiv preprint arXiv:1806.01969*, 2018.

A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 329–338, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4244-7. doi: 10.1109/FOCS.2010.38. URL http://dx.doi.org/10.1109/FOCS.2010.38.

A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *Proceedings of the 9th International Conference on Approximation Algorithms for Combinatorial Optimization Problems, and 10th International Conference on Randomization and Computation*, APPROX'06/RANDOM'06, pages 292–303, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-38044-2, 978-3-540-38044-3. doi: 10.1007/11830924_28. URL http://dx.doi.org/10.1007/11830924_28.

A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pages 1117–1126, Philadelphia, PA, USA, 2006. Society for Industrial and Applied Mathematics. ISBN 0-89871-605-5. URL http://dl.acm.org/citation.cfm?id=1109557.1109681.

I. Dhillon, R. Heath, M. Sustik, and J. Tropp. Generalized finite algorithms for constructing hermitian matrices with prescribed diagonal and spectrum. *SIAM Journal on Matrix Analysis and Applications*, 27(1):61–71, 2005. doi: 10.1137/S0895479803438183. URL https://doi.org/10.1137/S0895479803438183.

P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1-3):9–33, June 2004. ISSN 0885-6125. doi: 10.1023/B:MACH.0000033113.59016.96. URL https://doi.org/10.1023/B:MACH.0000033113.59016.96.

P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

M. Fickus, D. G. Mixon, and M. J. Poteet. Frame completions for optimally robust reconstruction. 2011a. doi: 10.1117/12.891813.

M. Fickus, D. G. Mixon, M. J. Poteet, and N. Strawn. Constructing all self-adjoint matrices with prescribed spectrum and diagonal, 2011b.

G. Gautier, R. Bardenet, and M. Valko. DPPy: Sampling determinantal point processes with Python. *arXiv preprint arXiv:1809.07258*, 2018.

G. H. Golub. Numerical methods for solving linear least squares problems. *Numer. Math.*, 7(3):206–216, June 1965. ISSN 0029-599X. doi: 10.1007/BF01436075. URL http://dx.doi.org/10.1007/BF01436075.

G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J. Sci. Comput.*, 17(4):848–869, July 1996. ISSN 1064-8275. doi: 10.1137/0917055. URL http://dx.doi.org/10.1137/0917055.

V. Guruswami and A. K. Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1207–1214. SIAM, 2012.

A. Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3):620–630, 1954.

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. 2005. doi: 10.1214/154957806000000078.

K. Johansson. Random matrices and determinantal processes. *ArXiv Mathematical Physics e-prints*, October 2005.

A. Kulesza and B. Taskar. k-DPPs: Fixed-size determinantal point processes. In Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 1193–1200. Omnipress, 2011. URL http://dblp.uni-trier.de/db/conf/icml/icml2011.html#KuleszaT11.

A. Kulesza, B. Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):853–877, 2015.

C. Li, S. Jegelka, and S. Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems*, pages 5038–5047, 2017a.

J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2017b.

P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, 16(1):861–911, 2015.

O Macchi. The coincidence approach to stochastic point processes. 7:83–122, 03 1975.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and its Applications*, volume 143. Springer, second edition, 2011. doi: 10.1007/978-0-387-68276-1.

L. Mor-Yosef and H. Avron. Sketching for principal component regression, 2018.

D. Papailiopoulos, A. Kyrillidis, and C. Boutsidis. Provable deterministic leverage score sampling. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 997–1006, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623698. URL http://doi.acm.org/10.1145/2623330.2623698.

G. Raskutti and M. W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

M. Slawski. On principal components regression, random projections, and column subsampling. *Electronic Journal of Statistics*, 12(2):3673–3712, 2018.

A. Soshnikov. Determinantal random point fields. *Russian Mathematical Surveys*, 55:923–975, October 2000. doi: 10.1070/RM2000v055n05ABEH000321.

N. Tremblay, S. Barthelmé, and P.-O. Amblard. Optimized algorithms to sample determinantal point processes. *arXiv preprint arXiv:1802.08471*, 2018.

A Another interpretation of the k-leverage scores

For $i \in [d]$, the SVD of $\boldsymbol{X}$ yields

$$\boldsymbol{X}_{:,i} = \sum_{\ell=1}^{r} V_{i,\ell} \boldsymbol{f}_\ell, \tag{55}$$

where $\boldsymbol{f}_\ell = \sigma_\ell \boldsymbol{U}_{:,\ell}$, $\ell \in [r]$, are orthogonal. Thus

$$\boldsymbol{X}_{:,i}^\mathsf{T} \boldsymbol{f}_j = V_{i,j} \|\boldsymbol{f}_j\|^2 = V_{i,j} \sigma_j^2. \tag{56}$$

Then

$$\frac{V_{i,j}}{\|\boldsymbol{X}_{:,i}\|} = \frac{\boldsymbol{X}_{:,i}^\mathsf{T} \boldsymbol{f}_j}{\sigma_j \|\boldsymbol{X}_{:,i}\| \|\boldsymbol{f}_j\|} =: \frac{\cos \eta_{i,j}}{\sigma_j}, \tag{57}$$

where $\eta_{i,j} \in [0, \pi/2]$ is the angle formed by $\boldsymbol{X}_{:,i}$ and $\boldsymbol{f}_j$. Finally, (56) also yields

$$\ell_i^k = \|\boldsymbol{X}_{:,i}\|^2 \sum_{j=1}^{k} \frac{\cos^2 \eta_{i,j}}{\sigma_j^2}. \tag{58}$$

Compared to the length-square distribution in Section 3.2, k-leverage scores thus favour columns that are aligned with the principal features. The weight $1/\sigma_j^2$ corrects the fact that features associated with large singular values are typically aligned with more columns. One could also imagine more arbitrary weights w_j/σ_j^2 in lieu of $1/\sigma_j^2$, or, equivalently, modified k-leverage scores

$$\ell_i^k(\boldsymbol{w}) = \sum_{j=1}^{k} w_j V_{i,j}^2.$$

However, the projection DPP with marginal kernel $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\mathsf{T}$ that we study in this paper is invariant to such reweightings. Indeed, for any $S \subset [d]$ of cardinality k,

$$\mathrm{Det}\left[\boldsymbol{V}_{S,[k]} \operatorname{Diag}(\boldsymbol{w}_{[k]}) \boldsymbol{V}_{[k],S}^\mathsf{T}\right] = \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \prod_{j \in [k]} w_j^2 \propto \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2. \tag{59}$$

Such a scaling is thus not a free parameter in $\boldsymbol{K}$.

B Majorization and Schur convexity

This section recalls some definitions and results from the theory of majorization and the notions of Schur-convexity and Schur-concavity. We refer to (Marshall et al., 2011) for further details. In this section, a subset $\mathcal{D} \subset \mathbb{R}^d$ is a symmetric domain if $\mathcal{D}$ is stable under coordinate permutations. Furthermore, a function f defined on a symmetric domain $\mathcal{D}$ is called symmetric if it is stable under coordinate permutations.

Definition 21 *Let $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_+^d$. $\boldsymbol{p}$ is said to majorize $\boldsymbol{q}$ according to Schur order and we note $\boldsymbol{q} \prec_S \boldsymbol{p}$ if*

$$\begin{cases} q_{i_1} \leq p_{j_1} \\ q_{i_1} + q_{i_2} \leq p_{j_1} + p_{j_2} \\ \dots \\ \sum_{k=1}^{d-1} q_{i_k} \leq \sum_{k=1}^{d-1} p_{j_k} \\ \sum_{k=1}^{d} q_{i_k} = \sum_{k=1}^{d} p_{j_k} \end{cases} \tag{60}$$

where $\boldsymbol{p}, \boldsymbol{q}$ are reordered so that $p_{i_d} \leq \dots \leq p_{i_1}$ and $q_{j_d} \leq \dots \leq q_{j_1}$.

The majorization order has an algebraic characterization using doubly stochastic matrices first proven by Hardy, Littlewood, and Polya in 1929.

Proposition 22 (Theorem B.2. Marshall et al., 2011) *The vector $\boldsymbol{p}$ majorizes the vector $\boldsymbol{q}$ if and only if there exists a $d \times d$ doubly stochastic matrix Π such that $\boldsymbol{q} = \boldsymbol{p}\boldsymbol{\Pi}$.*

Example 1 *Let $\boldsymbol{p} = (3, 0, 0)$ and $\boldsymbol{q} = (1, 1, 1)$. We check easily that $\boldsymbol{p}$ majorizes $\boldsymbol{q}$. Note that we can 'redistribute' $\boldsymbol{p}$ over $\boldsymbol{q}$ as follows: $\boldsymbol{q} = \frac{1}{3}\boldsymbol{J}\boldsymbol{p}$, where $\boldsymbol{J}$ is a 3×3 matrix of ones. The matrix $\boldsymbol{\Pi} = \frac{1}{3}\boldsymbol{J}$ is a doubly stochastic matrix.*

Schur order compares two vectors using multiple inequalities. To avoid such cumbersome calculations, a scalar metric of inequality in a vector is desired. This is possible using the notion of Schur-convex/concave function.

Definition 23 *Let f be a function on a symmetric domain $\mathcal{D} \subset \mathbb{R}_+^d$.*
f is said to be Schur convex if

$$\forall \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_+^d, \boldsymbol{q} \prec_S \boldsymbol{p} \implies f(\boldsymbol{q}) \leq f(\boldsymbol{p}). \tag{61}$$

f is said to be Schur concave if

$$\forall \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_+^d, \boldsymbol{q} \prec_S \boldsymbol{p} \implies f(\boldsymbol{q}) \geq f(\boldsymbol{p}). \tag{62}$$

Proposition 24 (Theorem A.3, Marshall et al., 2011) *Let f be a symmetric function defined on $\mathbb{R}_+^d$, let $\mathcal{D}$ be a permutation-symmetric domain in $\mathbb{R}_+^d$ and suppose that*

$$\forall x_i, x_j \in \mathbb{R}_+, (x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) > 0 \tag{63}$$

then

$$\forall \boldsymbol{p}, \boldsymbol{q} \in \mathcal{D}, \boldsymbol{q} \prec_S \boldsymbol{p} \implies f(\boldsymbol{q}) \leq f(\boldsymbol{p}), \tag{64}$$

and f is Schur convex.

We get a similar result for Schur concavity by switching the orders in the previous proposition.

Theorem 25 (Theorem 3.1, Guruswami and Sinop, 2012) *Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$, and let $\boldsymbol{\sigma} \in \mathbb{R}^d$ the vector containing the squares of the singular values of $\boldsymbol{X}$. The function*

$$\boldsymbol{\sigma} \mapsto \mathbb{E}_{\mathrm{VS}} \|\boldsymbol{X} - \Pi_S \boldsymbol{X}\|_{\mathrm{Fr}}^2 = (k+1)\frac{e_k(\boldsymbol{\sigma})}{e_{k-1}(\boldsymbol{\sigma})} \tag{65}$$

is Schur-concave.

C Principal angles and the Cosine Sine decomposition

C.1 Principal angles

This section surveys the notion of principal angles between subspaces, see (Golub and Van Loan, 1996, Section 6.4.3) for details.

Definition 26 *Let $\mathcal{P}, \mathcal{Q}$ be two subspaces in $\mathbb{R}^d$. Let $p = \dim \mathcal{P}$ and $q = \dim \mathcal{Q}$ and assume that $q \leq p$. To define the vector of principal angles $\boldsymbol{\theta} \in [0, \pi/2]^q$ between $\mathcal{P}$ and $\mathcal{Q}$, let*

$$\cos(\theta_1) = \max \left\{ \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}; \quad \boldsymbol{x} \in \mathcal{P}, \boldsymbol{y} \in \mathcal{Q} \right\} \tag{66}$$

be the cosine of the smallest angle between a vector of $\mathcal{P}$ and a vector of $\mathcal{Q}$, and let $(\boldsymbol{x}_1, \boldsymbol{y}_1) \in \mathcal{P} \times \mathcal{Q}$ be a pair of vectors realizing the maximum. For $i \in [2, q]$, define successively

$$\cos(\theta_i) = \max \left\{ \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}; \quad \boldsymbol{x} \in \mathcal{P}, \boldsymbol{y} \in \mathcal{Q}; \boldsymbol{x} \perp \boldsymbol{x}_j, \boldsymbol{y} \perp \boldsymbol{y}_j \, , \forall j \in [1 : i-1] \right\} \tag{67}$$

and denote $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{P} \times \mathcal{Q}$ such that $\cos(\theta_i) = \boldsymbol{x}_i^\intercal \boldsymbol{y}_i$.

Note that although the so-called principal vectors $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i \in [q]}$ are not uniquely defined by (66) and (67), the principal angles $\boldsymbol{\theta}$ are uniquely defined, see (Björck and Golub, 1973). The following result confirms this, while also providing a way to compute $\boldsymbol{\theta}$.

Proposition 27 (Björck and Golub, 1973, Ben-Israel, 1992) *Let $\mathcal{P}$ and $\mathcal{Q}$ and $\boldsymbol{\theta}$ be as in Definition 26. Let $\boldsymbol{P} \in \mathbb{R}^{d \times p}$, $\boldsymbol{Q} \in \mathbb{R}^{d \times q}$ be two orthogonal matrices, whose columns are orthonormal bases of $\mathcal{P}$ and $\mathcal{Q}$, respectively. Then*

$$\forall i \in [q], \quad \cos(\theta_i) = \sigma_i(\boldsymbol{Q}^\intercal \boldsymbol{P}). \tag{68}$$

In particular

$$\mathrm{Vol}_q^2(\boldsymbol{Q}^\intercal \boldsymbol{P}) = \prod_{i \in [q]} \cos^2(\theta_i). \tag{69}$$

An important case for our work arises when $q = k$, $\boldsymbol{Q} = \boldsymbol{V} \in \mathbb{R}^{d \times k}$, and $\boldsymbol{P} = \boldsymbol{S} \in \mathbb{R}^{d \times k}$ is a sampling matrix. The left-hand side of (69) then equals $\mathrm{Det}(\boldsymbol{V}_{:,S})^2$.

C.2 The Cosine Sine decomposition

The Cosine Sine (CS) decomposition is useful for the study of the relative position of two subspaces. It generalizes the notion of cosine, sine and tangent to subspaces.

Proposition 28 (Golub and Van Loan, 1996) *Let $q \leq d/2$ and $\boldsymbol{Q} = \left[\begin{array}{c} \boldsymbol{Q}_1 \\ \hline \boldsymbol{Q}_2 \end{array}\right]$ be a $d \times q$ orthogonal matrix, where $\boldsymbol{Q}_1 \in \mathbb{R}^{q \times q}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{(d-q) \times q}$. Assume that $\boldsymbol{Q}_1$ is non singular, then there exist orthogonal matrices $\boldsymbol{Y} \in \mathbb{R}^{d \times q}$ and*

$$\boldsymbol{W} = \left[\begin{array}{c|c} \boldsymbol{W}_1 & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{W}_2 \end{array}\right] \in \mathbb{R}^{d \times d}, \tag{70}$$

and a matrix

$$\mathbf{\Sigma} \;=\; \left[\begin{array}{c} \mathcal{C} \\ \hline \mathcal{S} \\ \hline \mathbf{0} \end{array}\right] \in \mathbb{R}^{d\times q}, \tag{71}$$

such that

$$\boldsymbol{Q} = \boldsymbol{W}\mathbf{\Sigma}\boldsymbol{Y}^T, \tag{72}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{q\times q}$ *and* $\boldsymbol{W}_2 \in \mathbb{R}^{d-q\times d-q}$, *and* $\mathcal{C},\mathcal{S} \in \mathbb{R}^{q\times q}$ *are diagonal matrices satisfying the identity* $\mathcal{C}^2+\mathcal{S}^2 = \mathbb{I}_q$. *In particular, each block* $\boldsymbol{Q}_i$ *factorizes as*

$$\begin{aligned} \boldsymbol{Q}_1 =& \boldsymbol{W}_1\,\mathcal{C}\,\boldsymbol{Y}^T \\ \boldsymbol{Q}_2 =& \boldsymbol{W}_2 \left[\begin{array}{c} \mathcal{S} \\ \hline \mathbf{0} \end{array}\right] \boldsymbol{Y}^T. \end{aligned} \tag{73}$$

The CS decomposition is defined for every orthogonal matrix. An important case is when $\boldsymbol{Q}$ is the product of an orthogonal matrix $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ and a sampling matrix $\boldsymbol{S} \in \mathbb{R}^{d\times k}$, that is $\boldsymbol{Q} = \boldsymbol{V}^\intercal\boldsymbol{S}$.

Corollary 29 *Let* $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ *be an orthogonal matrix and* $\boldsymbol{S} \in \mathbb{R}^{d\times k}$ *be a sampling matrix. Let*

$$\boldsymbol{Q} = \boldsymbol{V}^\intercal\boldsymbol{S} = \left[\begin{array}{c} \boldsymbol{V}_k^\intercal\boldsymbol{S} \\ \hline \boldsymbol{V}_{d-k}^\intercal\boldsymbol{S} \end{array}\right] \tag{74}$$

be a $d\times k$ *orthogonal matrix, with* $\mathrm{Det}(\boldsymbol{V}_k^\intercal\boldsymbol{S})^2 > 0$. *Let further* $\boldsymbol{Z}_S = \boldsymbol{V}_{d-k}^\intercal\boldsymbol{S}(\boldsymbol{V}_k^\intercal\boldsymbol{S})^{-1}$. *Then*

$$\mathrm{Tr}(\boldsymbol{Z}_S\boldsymbol{Z}_S^\intercal) \le \sum_{i\in[k]} \tan^2(\theta_i(S)). \tag{75}$$

Proof In the case $k \le d/2$, Proposition 28 applied to the matrix $\boldsymbol{Q} = \boldsymbol{V}^\intercal\boldsymbol{S}$ with $\boldsymbol{Q}_1 = \boldsymbol{V}_k^\intercal\boldsymbol{S}$ and $\boldsymbol{Q}_2 = \boldsymbol{V}_{d-k}^\intercal\boldsymbol{S}$ yields

$$\boldsymbol{Q}_1 = \boldsymbol{W}_1\,\mathcal{C}\,\boldsymbol{Y}^T \tag{76}$$

$$\boldsymbol{Q}_2 = \boldsymbol{W}_2 \left[\begin{array}{c} \mathcal{S} \\ \hline \mathbf{0} \end{array}\right] \boldsymbol{Y}^T. \tag{77}$$

Thus, the diagonal matrix $\mathcal{C}$ contains the singular values of the matrix $\boldsymbol{V}_k^\intercal\boldsymbol{S}$ that are cosines of the principal angles $(\theta_i(S))_{i\in[k]}$ between $\mathrm{Span}(\boldsymbol{V}_k)$ and $\mathrm{Span}(\boldsymbol{S})$ thanks to Proposition 27.
The identity $\mathcal{C}^2+\mathcal{S}^2 = \mathbb{I}_k$ and the fact that $\theta_i(S) \in [0, \frac{\pi}{2}]$ imply that the (diagonal) elements of $\mathcal{S}$ are equal to the sines of the principal angles between $\mathrm{Span}(\boldsymbol{V}_k)$ and $\mathrm{Span}(\boldsymbol{S})$. Let $\mathcal{T} = \mathcal{S}\,\mathcal{C}^{-1}$. $\mathcal{T} \in \mathbb{R}^{k\times k}$ is a diagonal matrix containing the tangents of the principal angles $(\theta_i(S))_{i\in[k]}$. Using (76), we get

$$\boldsymbol{Z}_S = \boldsymbol{V}_{d-k}^\intercal\boldsymbol{S}(\boldsymbol{V}_k^\intercal\boldsymbol{S})^{-1} = \boldsymbol{W}_2 \left[\begin{array}{c} \mathcal{S} \\ \hline \mathbf{0} \end{array}\right] \boldsymbol{Y}^\intercal\boldsymbol{Y}\,\mathcal{C}^{-1}\,\boldsymbol{W}_1^\intercal = \boldsymbol{W}_2 \left[\begin{array}{c} \mathcal{S} \\ \hline \mathbf{0} \end{array}\right] \mathcal{C}^{-1}\,\boldsymbol{W}_1^\intercal = \boldsymbol{W}_2 \left[\begin{array}{c} \mathcal{S}\,\mathcal{C}^{-1} \\ \hline \mathbf{0} \end{array}\right] \boldsymbol{W}_1^\intercal. \tag{78}$$

Then,

$$\mathrm{Tr}(\boldsymbol{Z}_S\boldsymbol{Z}_S^\intercal) = \mathrm{Tr}(\boldsymbol{W}_2 \left[\begin{array}{c|c} \mathcal{T}^2 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \boldsymbol{W}_2^\intercal) = \sum_{i\in[k]} \tan^2(\theta_i(S)). \tag{79}$$

■

D Proofs

D.1 Technical lemmas

We start with two useful emmas borrowed from the literature.

Lemma 30 (Lemma 3.1, Boutsidis et al., 2011) *Let $S \subset [d]$, then*

$$\|\boldsymbol{X} - \Pi_{S,k}^{\nu}\boldsymbol{X}\|_{\nu}^2 \leq \|\boldsymbol{E}(\boldsymbol{I} - \boldsymbol{P}_S)\|_{\nu}^2, \quad \nu \in \{2, \mathrm{Fr}\}, \tag{80}$$

where $\boldsymbol{E} = \boldsymbol{X} - \Pi_k\boldsymbol{X}$ and $\boldsymbol{P}_S = \boldsymbol{S}(\boldsymbol{V}_k^\intercal\boldsymbol{S})^{-1}\boldsymbol{V}_k^\intercal$. Furthermore,

$$\|\boldsymbol{X} - \Pi_{S,k}^{\nu}\boldsymbol{X}\|_{\nu}^2 \leq \frac{1}{\sigma_k^2(\boldsymbol{V}_{S,[k]})}\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\nu}^2, \quad \nu \in \{2, \mathrm{Fr}\}. \tag{81}$$

The following lemma was first proven by Deshpande et al., 2006, and later rephrased.

Lemma 31 (Lemma 11, Deshpande and Rademacher, 2010) *Let $\boldsymbol{V} \in \mathbb{R}^{k\times d}$, $r = \mathrm{rk}(\boldsymbol{V})$ and $\ell \in [1:r]$. Then*

$$\sum_{S\subset[d],|S|=\ell} e_\ell(\Sigma(\boldsymbol{V}_{:,S})^2) = e_\ell(\Sigma(\boldsymbol{V})^2) \tag{82}$$

where e_ℓ is the ℓ-th elementary symmetric polynomial on r variables, see Section 2.

Elementary symmetric polynomials play an important role in the proof of Proposition 18, in particular their interplay with the Schur order; see Appendix B for definitions.

Lemma 32 *Let $\phi, \psi : \mathbb{R}_+^d \to \mathbb{R}_+$ be defined by*

$$\phi : \boldsymbol{\sigma} \mapsto \frac{e_{k-1}(\boldsymbol{\sigma})}{e_k(\boldsymbol{\sigma})} \tag{83}$$

and

$$\psi : \boldsymbol{\sigma} \mapsto e_k(\boldsymbol{\sigma}). \tag{84}$$

Then both functions are symmetric, ϕ is Schur-convex, and ψ is Schur-concave.

Proof [of Lemma 32] Let $i, j \in [r], i \neq j$. Let $\sigma_i, \sigma_j \in \mathbb{R}_+$, it holds

$$\begin{aligned}(\sigma_i - \sigma_j)(\partial_i\phi(\boldsymbol{\sigma}) - \partial_j\phi(\boldsymbol{\sigma})) &= (\sigma_i - \sigma_j)(-\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}) \\ &= \frac{(\sigma_i - \sigma_j)^2(\sigma_i + \sigma_j)}{\sigma_i^2\sigma_j^2} \geq 0,\end{aligned}$$

so that ϕ is Schur-convex by Proposition 24. Similarly,

$$\begin{aligned}(\sigma_i - \sigma_j)(\partial_i\psi(\boldsymbol{\sigma}) - \partial_j\psi(\boldsymbol{\sigma})) &= (\sigma_i - \sigma_j)(\prod_{\ell\neq i}\sigma_\ell - \prod_{\ell\neq j}\sigma_\ell) \\ &= -(\sigma_i - \sigma_j)^2\prod_{\ell\neq i,j}\sigma_\ell \geq 0,\end{aligned}$$

so that ψ is Schur-concave by Proposition 24. ■

Elementary symmetric polynomials also interact nicely with "marginalizing" sums.

Lemma 33 *Let $\boldsymbol{V}$ be a real $k \times d$ matrix and let $r = \mathrm{rk}(\boldsymbol{V})$. Denote by p the number of non zero columns of $\boldsymbol{V}$. Then for all $k \leq r+1$,*

$$\sum_{\substack{S\subset[d],|S|=k\\ \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2>0}} \sum_{\substack{T\subset[S]\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (p-k+1)e_{k-1}(\Sigma(\boldsymbol{V})^2). \tag{85}$$

A fortiori,

$$\sum_{\substack{S\subset[d],|S|=k\\ \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2>0}} \sum_{\substack{T\subset[S]\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (d-k+1)e_{k-1}(\Sigma(\boldsymbol{V})^2). \tag{86}$$

Proof [of Lemma 33] For $T \subset [d]$, $|T| = k-1$,

$$\begin{aligned}\Omega_1(T) &= \{S \subset [d] : |S| = k, T \subset S,\ \forall i \in S,\ \boldsymbol{V}_{:,i} \neq \boldsymbol{0}\}\\ \Omega_2(T) &= \left\{S \subset [d] : |S| = k, T \subset S, \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2 > 0\right\}.\end{aligned}$$

Note that $\Omega_2(T) \subset \Omega_1(T)$ so that

$$\begin{aligned}\sum_{\substack{S\subset[d],|S|=k\\ \mathrm{Vol}_k(\boldsymbol{V}_{:,S})^2>0}} \sum_{\substack{T\subset S\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) &= \sum_{\substack{T\subset[d]\\ |T|=k-1}} \sum_{S\in\Omega_2(T)} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2)\\ &\leq \sum_{\substack{T\subset[d]\\ |T|=k-1}} \sum_{S\in\Omega_1(T)} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2).\end{aligned}$$

The set $\Omega_1(T)$ has at most $(p-k+1)$ elements so that

$$\sum_{\substack{T\subset[d]\\ |T|=k-1}} \sum_{S\in\Omega_1(T)} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (p-k+1) \sum_{\substack{T\subset[d]\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2). \tag{87}$$

Lemma 31 for $\ell = k-1$ further yields

$$(p-k+1) \sum_{\substack{T\subset[d]\\ |T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{:,T})^2) \leq (p-k+1)\, e_{k-1}(\Sigma(\boldsymbol{V})^2). \tag{88}$$

■

D.2 Proof of Proposition 16

First, Lemma 30 yields

$$\begin{aligned}\sum_{S\subset[d],|S|=k} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X} - \Pi_S^\nu \boldsymbol{X}\|_\nu^2 &\leq \sum_{S\subset[d],|S|=k} \frac{1}{\sigma_k^2(\boldsymbol{V}_{S,[k]})} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \, \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_\nu^2\\ &= \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_\nu^2 \sum_{S\subset[d],|S|=k} \prod_{\ell=1}^{k-1} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}),\end{aligned} \tag{89}$$

where the last equality follows from

$$\mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 = \prod_{\ell=1}^{k} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}). \tag{90}$$

By definition of the polynomial e_{k-1}, it further holds

$$\prod_{\ell=1}^{k-1} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}) \leq e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2), \tag{91}$$

so that (89) leads to

$$\sum_{S\subset[d],|S|=k} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X} - \Pi_S^\nu \boldsymbol{X}\|_\nu^2 \leq \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_\nu^2 \sum_{S\subset[d],|S|=k} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2). \tag{92}$$

Now, Lemma 31 applied to the matrix $\boldsymbol{V}_{S,[k]}^\intercal$ gives

$$e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) = \sum_{T\subset S,|T|=k-1} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2), \tag{93}$$

Therefore, Lemma 33 yields

$$\sum_{S\subset[d],|S|=k} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) \leq (d-k+1) \sum_{T\subset[d],|T|=k-1} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2). \tag{94}$$

Using Lemma 31 and the fact that $\boldsymbol{V}_k$ is orthogonal, we finally write

$$\sum_{T\subset[d],|T|=k-1} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2) = e_{k-1}(\Sigma(\boldsymbol{V}_k)^2) = k. \tag{95}$$

Plugging (95) into (94), and then into (92) concludes the proof of Proposition 16.

D.3 Proof of Proposition 17

We first prove the Frobenius norm bound, which requires more work. The spectral bound is easier and uses a subset of the arguments for the Frobenius norm.

D.3.1 Frobenius norm bound

Recall that $\boldsymbol{E} = \boldsymbol{X} - \Pi_k \boldsymbol{X}$. We start with Lemma 30:

$$\begin{aligned} \|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 &\leq \|\boldsymbol{E}(\boldsymbol{I} - \boldsymbol{P}_S)\|_{\mathrm{Fr}}^2 \\ &\leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \mathrm{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal) - 2\,\mathrm{Tr}(\boldsymbol{P}_S^\intercal \boldsymbol{E}^\intercal \boldsymbol{E}). \end{aligned} \tag{96}$$

Since $\boldsymbol{E}^\intercal \boldsymbol{E} = \boldsymbol{V}_{r-k} \boldsymbol{\Sigma}_{r-k}^2 \boldsymbol{V}_{r-k}^\intercal$ and $\boldsymbol{P}_S = \boldsymbol{S}(\boldsymbol{V}_k^\intercal \boldsymbol{S})^{-1} \boldsymbol{V}_k^\intercal$,

$$\begin{aligned} \mathrm{Tr}(\boldsymbol{P}_S^\intercal \boldsymbol{E}^\intercal \boldsymbol{E}) &= \mathrm{Tr}\left(\boldsymbol{V}_k \left((\boldsymbol{V}_k^\intercal \boldsymbol{S})^\intercal \right)^{-1} \boldsymbol{S}^\intercal \boldsymbol{V}_{r-k} \boldsymbol{\Sigma}_{r-k} \boldsymbol{V}_{r-k}^\intercal \right) \\ &= \mathrm{Tr}\left(\boldsymbol{V}_{r-k}^\intercal \boldsymbol{V}_k \left((\boldsymbol{V}_k^\intercal \boldsymbol{S})^\intercal \right)^{-1} \boldsymbol{S}^\intercal \boldsymbol{V}_{r-k} \boldsymbol{\Sigma}_{r-k} \right) \\ &= 0, \end{aligned} \tag{97}$$

where the last equality follows from $\boldsymbol{V}_{r-k}^\intercal \boldsymbol{V}_k = \boldsymbol{0}$. Therefore, (96) becomes

$$\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal). \tag{98}$$

Taking expectations,

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \sum_{S \subset [d], |S|=k} \operatorname{Det}(\boldsymbol{V}_{S,[k]})^2 \operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal). \tag{99}$$

Proposition 27 expresses $\operatorname{Det}(\boldsymbol{V}_{S,[k]})^2$ as a function of the principal angles $(\theta_i(S))$ between $\operatorname{Span}(\boldsymbol{V}_k)$ and $\operatorname{Span}(\boldsymbol{S})$, namely

$$\operatorname{Det}(\boldsymbol{V}_{S,[k]})^2 = \prod_{i \in [k]} \cos^2(\theta_i(S)). \tag{100}$$

The remainder of the proof is in two steps. First, we bound the second factor in the sum in the right-hand side of (99) with a similar geometric expression. This allows trigonometric manipulations. Second, we work our way back to elementary symmetric polynomials of spectra, and we conclude after some simple algebra.

First, for $S \subset [d], |S| = k$, let

$$\boldsymbol{Z}_S = \boldsymbol{V}_{d-k}^\intercal \boldsymbol{S} (\boldsymbol{V}_k^\intercal \boldsymbol{S})^{-1} = \boldsymbol{V}_{d-k}^\intercal \boldsymbol{P}_S \boldsymbol{V}_k.$$

It allows us to write

$$\operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal) = \operatorname{Tr}(\boldsymbol{V}_{d-k} \boldsymbol{\Sigma}_{d-k}^2 \boldsymbol{V}_{d-k}^\intercal \boldsymbol{P}_S \boldsymbol{P}_S^\intercal) = \operatorname{Tr}(\boldsymbol{\Sigma}_{r-k}^2 \boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal). \tag{101}$$

However, for real symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the same size, a simple diagonalization argument yields

$$\operatorname{Tr}(\boldsymbol{A}\boldsymbol{B}) \leq \|\boldsymbol{A}\|_2 \operatorname{Tr}(\boldsymbol{B}), \tag{102}$$

so that

$$\operatorname{Tr}(\boldsymbol{E}^\intercal \boldsymbol{E} \boldsymbol{P}_S \boldsymbol{P}_S^\intercal) = \operatorname{Tr}(\boldsymbol{\Sigma}_{r-k}^2 \boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal) \leq \boldsymbol{\sigma}_{k+1}^2 \operatorname{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal). \tag{103}$$

In Appendix C, we characterize $\operatorname{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal)$ using principal angles, see (75). This reads

$$\operatorname{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal) = \sum_{j \in [k]} \tan^2(\theta_j(S)). \tag{104}$$

Combining (99), (103), (100), and (104), we obtain the following intermediate bound

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \boldsymbol{\sigma}_{k+1}^2 \sum_{S \subset [d], |S|=k} \left[\prod_{i \in [k]} \cos^2(\theta_i(S))\right] \left[\sum_{j \in [k]} \tan^2(\theta_j(S))\right]. \tag{105}$$

Distributing the sum and using trigonometric identities, the general term of the sum in (105) becomes

$$\begin{aligned}\left[\prod_{i \in [k]} \cos^2(\theta_i(S))\right] \left[\sum_{j \in [k]} \tan^2(\theta_j(S))\right] &= \sum_{i \in [k]} (1 - \cos^2(\theta_i(S))) \prod_{j \in [k], j \neq i} \cos^2(\theta_j(S)) \\ &= \sum_{i \in [k]} \prod_{j \in [k], j \neq i} \cos^2(\theta_j(S)) - \sum_{i \in [k]} \prod_{j \in [k]} \cos^2(\theta_j(S)).\end{aligned} \tag{106}$$

The $(\cos(\theta_j(S)))_{j\in[k]}$ are the singular values of the matrix $\boldsymbol{V}_{S,[k]}$ so that

$$\sum_{i\in[k]}\prod_{j\in[k],j\neq i}\cos^2(\theta_j(S)) = e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2), \tag{107}$$

and

$$\prod_{j\in[k]}\cos^2(\theta_j(S)) = e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2). \tag{108}$$

Back to (106), one gets

$$\begin{aligned}\left[\prod_{i\in[k]}\cos^2(\theta_i(S))\right]\left[\sum_{j\in[k]}\tan^2(\theta_j(S))\right] &= e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - \sum_{i\in[k]} e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2) \\ &= e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - ke_k(\Sigma(\boldsymbol{V}_{S,[k]})^2). \end{aligned}\tag{109}$$

Thus, plugging (109) back into the intermediate bound (105), it comes

$$\begin{aligned}&\mathbb{E}_{\mathrm{DPP}}\,\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \\ &\qquad\leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \boldsymbol{\sigma}_{k+1}^2\left[\sum_{\substack{S\subset[d]\\|S|=k}} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) - k\sum_{\substack{S\subset[d]\\|S|=k}} e_k(\Sigma(\boldsymbol{V}_{S,[k]})^2)\right].\end{aligned}\tag{110}$$

Using Lemma 31 twice, it comes

$$\begin{aligned}&\mathbb{E}_{\mathrm{DPP}}\,\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \\ &\qquad\leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \boldsymbol{\sigma}_{k+1}^2\left[\sum_{\substack{S\subset[d]\\|S|=k}}\sum_{\substack{T\subset S\\|T|=k-1}} e_{k-1}(\Sigma(\boldsymbol{V}_{T,[k]})^2) - ke_k(\Sigma(\boldsymbol{V}_{:,[k]})^2)\right].\end{aligned}\tag{111}$$

Lemmas 33 and the identities $e_{k-1}(\Sigma(\boldsymbol{V}_{:,[k]})^2) = k$ and $e_k(\Sigma(\boldsymbol{V}_{:,[k]})^2) = 1$ allow us to conclude

$$\begin{aligned}\mathbb{E}_{\mathrm{DPP}}\,\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 &\leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \boldsymbol{\sigma}_{k+1}^2\left[(p-k+1)e_{k-1}(\Sigma(\boldsymbol{V}_{:,[k]})^2) - k\right] &(112)\\ &= \|\boldsymbol{E}\|_{\mathrm{Fr}}^2 + \boldsymbol{\sigma}_{k+1}^2(p-k)k. &(113)\end{aligned}$$

By definition of β (45), we have proven (47), i.e.,

$$\mathbb{E}_{\mathrm{DPP}}\,\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leq \|\boldsymbol{E}\|_{\mathrm{Fr}}^2\left(1 + \beta\frac{p-k}{d-k}k\right).$$

D.3.2 Spectral norm bound

The bound in spectral norm is easier to derive. We start from Lemma 30:

$$\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^2 \boldsymbol{X}\|_2^2 = \sum_{S\subset[d],|S|=k} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X} - \Pi_S \boldsymbol{X}\|_2^2 \tag{114}$$

$$\leq \|\boldsymbol{E}\|_2^2 \sum_{\substack{S\subset[d],|S|=k \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2>0}} \prod_{\ell=1}^{k-1} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}) \tag{115}$$

By definition of e_{k-1}, it comes

$$\begin{aligned}
\mathbb{E}_{\mathrm{DPP}} \|\boldsymbol{X} - \Pi_S^2 \boldsymbol{X}\|_2^2 &\leq \|\boldsymbol{E}\|_2^2 \sum_{\substack{S\subset[d],|S|=k \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2>0}} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) \\
&\leq (p-k+1)\, e_{k-1}(\Sigma(\boldsymbol{V}_{:,[k]})^2)\, \|\boldsymbol{E}\|_2^2 \\
&= (p-k+1)\, k\, \|\boldsymbol{E}\|_2^2,
\end{aligned}$$

where we again used the double sum trick of (111) and Lemma 33.

D.4 Proof of Theorem 18

We start with a lemma on evaluations of elementary symmetric polynomials on specific sequences.

Lemma 34 *Let $\boldsymbol{\lambda} \in [0,1]^k$ such that*

$$\begin{cases} \lambda_1 \geq \cdots \geq \lambda_k, \\ \Lambda = \sum_{i=1}^k \lambda_i \geq k - 1 + \frac{1}{\theta}. \end{cases} \tag{116}$$

Then, with the functions ϕ, ψ introduced in Lemma 32,

$$\begin{cases} \psi(\boldsymbol{\lambda}) & \geq \dfrac{1}{\theta}, \\ \phi(\boldsymbol{\lambda}) & \leq k - 1 + \theta. \end{cases} \tag{117}$$

Proof Let $\hat{\boldsymbol{\lambda}} = (1, ..., 1, \Lambda - k + 1) \in \mathbb{R}^k$. Then

$$\begin{cases} \lambda_1 \leq \hat{\lambda}_1 \\ \lambda_1 + \lambda_2 \leq \hat{\lambda}_1 + \hat{\lambda}_2 \\ \cdots \\ \sum_{i=1}^{k-1} \lambda_i \leq \sum_{i=1}^{k-1} \hat{\lambda}_i \\ \sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} \hat{\lambda}_i \end{cases} \tag{118}$$

so that, according to Definition 21,

$$\boldsymbol{\lambda} \prec_S \hat{\boldsymbol{\lambda}}. \tag{119}$$

Lemma 32 ensures the Schur-convexity of ϕ and the Schur-concavity of ψ, so that

$$\phi(\boldsymbol{\lambda}) \leq \phi(\hat{\boldsymbol{\lambda}}) = k - 1 + \frac{1}{\Lambda - k + 1} \leq k - 1 + \theta,$$

and

$$\psi(\boldsymbol{\lambda}) \geq \psi(\hat{\boldsymbol{\lambda}}) = \Lambda - k + 1 \geq \frac{1}{\theta}.$$

■

D.4.1 Frobenius norm bound

Let $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\intercal$, and π be a permutation of $[d]$ that reorders the leverage scores decreasingly,

$$\ell_{\pi_1}^k \geq \ell_{\pi_2}^k \geq ... \geq \ell_{\pi_d}^k. \tag{120}$$

By construction, $T_{p_{\text{eff}}} = [\pi_{p_{\text{eff}}}, ..., \pi_d]$ thus collects the indices of the smallest leverage scores. Finally, denoting by $\boldsymbol{\Pi} = (\delta_{i,\pi_j})_{(i,j)\in[d]\times[d]}$ the matricial representation of permutation π, we let

$$\boldsymbol{K}^\pi = \boldsymbol{\Pi} \boldsymbol{K} \boldsymbol{\Pi}^\intercal = ((\boldsymbol{K}_{\pi_i,\pi_j}))_{1\leq i,j\leq d}.$$

The goal of the proof is to bound

$$\mathbb{E}_{\text{DPP}}\left[\|\boldsymbol{X} - \Pi_S^{\text{Fr}} \boldsymbol{X}\|_{\text{Fr}}^2 | S \cap T_{p_{\text{eff}}} = \emptyset\right] = \frac{\sum \text{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X} - \Pi_S^{\text{Fr}} \boldsymbol{X}\|_{\text{Fr}}^2}{\sum \text{Det}(\boldsymbol{V}_{S,[k]})^2}, \tag{121}$$

where both sums run over subsets $S \subset [d]$ such that $|S| = k$ and $S \cap T_{p_{\text{eff}}(\theta)} = \emptyset$. For simplicity, let us write

$$Z_{k,p_{\text{eff}}(\theta)} = \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} \text{Det}(\boldsymbol{V}_{S,[k]})^2, \tag{122}$$

$$Y_{k,p_{\text{eff}}(\theta)} = \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} \text{Det}(\boldsymbol{V}_{S,[k]})^2 \,\text{Tr}(\boldsymbol{Z}_S \boldsymbol{Z}_S^\intercal). \tag{123}$$

Following steps (99) to (103) of the previous proof, one obtains

$$\mathbb{E}_{\text{DPP}}\left[\|\boldsymbol{X} - \Pi_S^{\text{Fr}} \boldsymbol{X}\|_{\text{Fr}}^2 \mid S \cap T_{p_{\text{eff}}} = \emptyset\right] \leq \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_{\text{Fr}}^2 + \boldsymbol{\sigma}_{k+1}^2 \frac{Y_{k,p_{\text{eff}}(\theta)}}{Z_{k,p_{\text{eff}}(\theta)}}. \tag{124}$$

By definition (45) of the flatness parameter β,

$$\boldsymbol{\sigma}_{k+1}^2 = \beta \frac{1}{d-k} \sum_{j\geq k+1} \boldsymbol{\sigma}_j^2 = \beta \frac{1}{d-k} \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_{\text{Fr}}^2. \tag{125}$$

Then, it remains to upper bound the ratio $Y_{k,p_{\text{eff}}(\theta)}/Z_{k,p_{\text{eff}}(\theta)}$ in (124), which is the important part of the proof. We first evaluate $Z_{k,p_{\text{eff}}(\theta)}$ and then bound $Y_{k,p_{\text{eff}}(\theta)}$.

The matrix $\boldsymbol{\Pi}\boldsymbol{V}_k \in \mathbb{R}^{d\times k}$ has its rows ordered by decreasing leverage scores. Let $\tilde{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)} \in \mathbb{R}^{p_{\mathrm{eff}}(\theta)\times k}$ be the submatrix corresponding to the first $p_{\mathrm{eff}}(\theta)$ rows of $\boldsymbol{\Pi}\boldsymbol{V}_k$. Let also

$$\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)} = \begin{pmatrix} \tilde{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)} \\ \mathbf{0}_{d-p_{\mathrm{eff}}(\theta),k} \end{pmatrix}$$

be padded with zeros. Then

$$\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)} = \left[\begin{array}{c|c} \tilde{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)}\tilde{\boldsymbol{V}}^{\intercal}_{\pi,p_{\mathrm{eff}}(\theta)} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] = \hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)}(\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)})^{\intercal} \in \mathbb{R}^{d\times d}. \tag{126}$$

The nonzero block of $\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)}$ is a submatrix of $\boldsymbol{K}^{\pi}$, and $\operatorname{rk}\boldsymbol{K}^{\pi} = \operatorname{rk}\boldsymbol{K} = k$. Hence $\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)}$ has at most k nonzero eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0 = \lambda_{k+1} = \cdots = \lambda_d. \tag{127}$$

Therefore,

$$e_k(\Lambda(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})) = \sum_{\substack{T\subset[d]\\|T|=k}} \prod_{j\in T} \lambda_j = \prod_{i\in[k]} \lambda_i. \tag{128}$$

Note moreover that

$$\forall \ell \in [k],\ \ e_\ell(\Sigma(\hat{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)})^2) = e_\ell(\Lambda(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})). \tag{129}$$

By construction,

$$Z_{k,p_{\mathrm{eff}}(\theta)} = \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \operatorname{Det}(\boldsymbol{V}_{S,[k]})^2 = \sum_{S\subset[d],|S|=k} \operatorname{Det}\left[\left(\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)}\right)_{S,:}\right]^2 \tag{130}$$

Then, Lemma 31 yields

$$Z_{k,p_{\mathrm{eff}}(\theta)} = e_k(\Sigma(\hat{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)})^2) = e_k(\Lambda(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})) = \prod_{i\in[k]} \lambda_i. \tag{131}$$

Now we bound $Y_{k,p_{\mathrm{eff}}(\theta)}$. We use again principal angles and trigonometric identities. Using (104) and (109) above, it holds

$$\begin{aligned}
Y_{k,p_{\mathrm{eff}}(\theta)} &= \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \operatorname{Det}(\boldsymbol{V}_{S,[k]})^2 \operatorname{Tr}(\boldsymbol{Z}_S\boldsymbol{Z}_S^{\intercal}) \\
&= \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \prod_{i\in[k]} \cos^2(\theta_i(S)) \sum_{j\in[k]} \tan^2(\theta_j(S)) \\
&= \sum_{\substack{S\subset[d],|S|=k\\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} e_{k-1}\left(\Sigma(\boldsymbol{V}_{S,[k]})^2\right) - k\, e_k\left(\Sigma(\boldsymbol{V}_{S,[k]}\right)^2 \qquad (132) \\
&= \sum_{S\subset[d],|S|=k} e_{k-1}\left(\Sigma\left(\left[\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)}\right]_{S,:}\right)^2\right) - k\, e_k\left(\Sigma\left(\left[\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)}\right]_{S,:}\right)^2\right) \qquad (133)
\end{aligned}$$

By Lemma 33 applied to the matrix $\hat{\boldsymbol{V}}_{\pi,p_{\mathrm{eff}}(\theta)}$ combined to (130), we get

$$\begin{aligned} Y_{k,p_{\mathrm{eff}}(\theta)} &\leq (p_{\mathrm{eff}}(\theta)-k+1)e_{k-1}(\Sigma(\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)})^2) - k\, e_k(\Sigma(\hat{\boldsymbol{V}}^{\pi}_{p_{\mathrm{eff}}(\theta)})^2) \\ &\leq (p_{\mathrm{eff}}(\theta)-k+1)e_{k-1}(\Lambda(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})) - k\, e_k(\Lambda(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})) \\ &\leq \Big((p_{\mathrm{eff}}(\theta)-k+1)\phi(\tilde{\boldsymbol{\lambda}}) - k\Big) Z_{k,p_{\mathrm{eff}}(\theta)}. \end{aligned} \tag{134}$$

where $\tilde{\boldsymbol{\lambda}} = (1,\dots,1,\mathrm{Tr}(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})-k+1) \in \mathbb{R}^k$, see Lemma 34. Now, as in the proof of Lemma 34,

$$\phi(\tilde{\boldsymbol{\lambda}}) = k-1+\frac{1}{\mathrm{Tr}(\boldsymbol{K}^{\pi}_{p_{\mathrm{eff}}(\theta)})-k+1} \leq k-1+\theta$$

by (49). Thus (134) yields

$$\frac{Y_{k,p_{\mathrm{eff}}(\theta)}}{Z_{k,p_{\mathrm{eff}}(\theta)}} \leq (p_{\mathrm{eff}}(\theta)-k+1)(k-1+\theta) - k \leq (p_{\mathrm{eff}}(\theta)-k+1)(k-1+\theta). \tag{135}$$

Finally, plugging (135) and (125) in (124) concludes the proof of (52).

D.4.2 Spectral norm bound

We proceed as for the Frobenius norm, using the notation of Section D.3.1. Lemma 30, Equations (132) and (135) yield

$$\begin{aligned} &\mathbb{E}_{\mathrm{DPP}}\left[\|\boldsymbol{X}-\Pi_S^2\boldsymbol{X}\|_2^2 \mid S\cap T_{p_{\mathrm{eff}}}=\emptyset\right] \\ &\qquad = Z^{-1}_{k,p_{\mathrm{eff}}(\theta)} \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset}} \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2 \|\boldsymbol{X}-\Pi_S^2\boldsymbol{X}\|_2^2, \\ &\qquad \leq Z^{-1}_{k,p_{\mathrm{eff}}(\theta)} \|\boldsymbol{X}-\Pi_k\boldsymbol{X}\|_2^2 \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset, \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2>0}} \prod_{\ell=1}^{k-1} \sigma_\ell^2(\boldsymbol{V}_{S,[k]}) \\ &\qquad \leq Z^{-1}_{k,p_{\mathrm{eff}}(\theta)} \|\boldsymbol{X}-\Pi_k\boldsymbol{X}\|_2^2 \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\mathrm{eff}}(\theta)}=\emptyset \\ \mathrm{Det}(\boldsymbol{V}_{S,[k]})^2>0}} e_{k-1}(\Sigma(\boldsymbol{V}_{S,[k]})^2) \\ &\qquad \leq \frac{Y_{k,p_{\mathrm{eff}}(\theta)}}{Z_{k,p_{\mathrm{eff}}(\theta)}} \|\boldsymbol{X}-\Pi_k\boldsymbol{X}\|_2^2 \\ &\qquad \leq (p_{\mathrm{eff}}(\theta)-k+1)(k-1+\theta)\|\boldsymbol{X}-\Pi_k\boldsymbol{X}\|_2^2, \end{aligned}$$

which is the claimed spectral bound.

D.4.3 Bounding the probability of rejection

Still with the notation of Section D.3.1, (130) yields

$$\begin{aligned}\mathbb{P}(S \cap T_{p_{\text{eff}}(\theta)} = \emptyset) &= \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} \text{Det}(\boldsymbol{V}_{S,[k]})^2 \\ &= e_k(\boldsymbol{K}^{\pi}_{p_{\text{eff}}(\theta)}) \qquad (136) \\ &= \prod_{i\in[k]} \lambda_i \\ &= \psi(\hat{\boldsymbol{\lambda}}). \qquad (137)\end{aligned}$$

Lemma 34 concludes the proof since

$$\psi(\hat{\boldsymbol{\lambda}}) \geq \frac{1}{\theta}. \tag{138}$$

D.5 Proof of Proposition 20

First, Proposition 11 gives

$$\mathcal{E}(\boldsymbol{w}_S) \leq \frac{(1 + \max_{i\in[k]} \tan^2\theta_i(S))\|\boldsymbol{w}^*\|^2\sigma^2_{k+1}}{N} + \frac{k}{N}\nu. \tag{139}$$

Now (75) further gives

$$\max_{i\in[k]} \tan^2\theta_i(S) \leq \sum_{i\in[k]} \tan^2\theta_i(S) = \text{Tr}(\boldsymbol{Z}_S\boldsymbol{Z}_S^\intercal). \tag{140}$$

The proof now follows the same lines as for the approximation bounds. First, following the lines of Section D.3, , we straightforwardly bound

$$\mathbb{E}_{\text{DPP}} \sum_{i\in[k]} \tan^2(\theta_i(S)) = \sum_{S\subset[d],|S|=k} \prod_{i\in[k]} \cos^2(\theta_i(S)) \sum_{j\in[k]} \tan^2(\theta_j(S)) \tag{141}$$

and obtain (53). In a similar vein, the same lines as in Section D.4 allow bounding

$$\mathbb{E}_{\text{DPP}}\left[\sum_{i\in[k]} \tan^2(\theta_i(S))|S\cap T_{p_{\text{eff}}} = \emptyset\right] = \sum_{\substack{S\subset[d],|S|=k \\ S\cap T_{p_{\text{eff}}(\theta)}=\emptyset}} \prod_{i\in[k]} \cos^2(\theta_i(S)) \sum_{j\in[k]} \tan^2(\theta_j(S). \tag{142}$$

and yield (54).

E Generating orthogonal matrices with prescribed leverage scores

In this section, we describe an algorithm that samples a random orthonormal matrix with a prescribed profile of k-leverage scores. This algorithm was used to generate the matrices $\boldsymbol{F} = \boldsymbol{V}_k^\intercal \in \mathbb{R}^{k\times d}$ for the toy datasets of Section 6. The orthogonality

constraint can be expressed as a condition on the spectrum of the matrix $\boldsymbol{K} = \boldsymbol{V}_k \boldsymbol{V}_k^\intercal$, namely $\mathrm{Sp}(\boldsymbol{K}) \subset \{0, 1\}$. On the other hand, the constraint on the k-leverage scores can be expressed as a condition on the diagonal of $\boldsymbol{K}$. Thus, the problem of generating an orthogonal matrix with a given profile of k-leverage scores boils down to enforcing conditions on the spectrum and the diagonal of a symmetric matrix $\boldsymbol{K}$.

E.1 Definitions and statement of the problem

We denote by $(\boldsymbol{f}_i)_{i\in[d]}$ the columns of the matrix $\boldsymbol{F}$. For $n \in \mathbb{N}$, we write $\mathbb{1}_n$ the vector containing ones living in $\mathbb{R}^n$, and $\mathbb{0}_n$ the vector containing zeros living in $\mathbb{R}^n$. We say that the vector $\boldsymbol{u} \in \mathbb{R}^n$ interlaces on $\boldsymbol{v} \in \mathbb{R}^n$ and we denote

$$\boldsymbol{u} \sqsubseteq \boldsymbol{v}$$

if $u_n \leq v_n$ and $\forall i \in [1 : n-1]$, $v_{i+1} \leq u_i \leq v_i$.

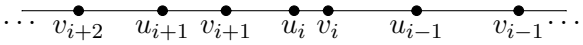


Figure 9: Illustration of the interlacing of $\boldsymbol{u}$ on $\boldsymbol{v}$.

Definition 35 *Let $k, d \in \mathbb{N}$, with $k \leq d$. Let $\boldsymbol{F} \in \mathbb{R}^{k\times d}$ be a full rank matrix*[5]*. Within this section, we denote $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$ the squares of the nonvanishing singular values of the matrix $\boldsymbol{F}$, and $\boldsymbol{\ell} = (\ell_1 = \|\boldsymbol{f}_1\|^2, \ell_2 = \|\boldsymbol{f}_2\|^2, \ldots, \ell_d = \|\boldsymbol{f}_d\|^2)$ are the squared norms of the columns of $\boldsymbol{F}$, which we assume to be ordered decreasingly:*

$$\ell_1 \geq \ell_2 \geq \cdots \geq \ell_d.$$

When $\boldsymbol{F}$ is orthogonal, we can think of $\boldsymbol{\ell}$ as a vector of leverage scores.

We are interested in the problem of constructing an orthogonal matrix given its leverage scores.

Problem 1 *Let $k, d \in \mathbb{N}$, with $k \leq d$, and let $\boldsymbol{\ell} \in \mathbb{R}_+^d$ such that $\sum_{i=1}^{d} \ell_i = k$. Build a matrix $\boldsymbol{F} \in \mathbb{R}^{k\times d}$ such that*

$$\mathrm{Sp}(\boldsymbol{F}^\intercal \boldsymbol{F}) = [\mathbb{1}_k, \mathbb{0}_{d-k}], \tag{143}$$

and

$$\mathrm{Diag}(\boldsymbol{F}^\intercal \boldsymbol{F}) = \boldsymbol{\ell}. \tag{144}$$

We actually consider here the generalization of Problem 2 to an arbitrary spectrum.

Problem 2 *Let $k, d \in \mathbb{N}$, with $k \leq d$, and let $\boldsymbol{\ell} \in \mathbb{R}_+^d$ such that $\sum_{i=1}^{d} \ell_i = \sum_{i=1}^{k} \sigma_i^2$. Build a matrix $\boldsymbol{F} \in \mathbb{R}^{k\times d}$ such that*

$$\mathrm{Sp}(\boldsymbol{F}^\intercal \boldsymbol{F}) = [\boldsymbol{\sigma}^2, \mathbb{0}_{d-k}] =: \hat{\boldsymbol{\sigma}}^2 \tag{145}$$

and

$$\mathrm{Diag}(\boldsymbol{F}^\intercal \boldsymbol{F}) = \boldsymbol{\ell}. \tag{146}$$

[5] A *frame*, using the definitions of (Fickus et al., 2011a) and (Fickus et al., 2011b).

Denote by

$$\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})} = \{\boldsymbol{M} \in \mathbb{R}^{d\times d} \text{ symmetric } / \ \mathrm{Diag}(\boldsymbol{M}) = \boldsymbol{\ell}, \ \ \mathrm{Sp}(\boldsymbol{M}) = \hat{\boldsymbol{\sigma}}^2\}. \tag{147}$$

The non-emptiness of $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$ is determined by a majorization condition between $\boldsymbol{\ell}$ and $\hat{\boldsymbol{\sigma}}$, see Appendix B for definitions. More precisely, we have the following theorem.

Theorem 36 (Schur-Horn) *Let $k, d \in \mathbb{N}$, with $k \leq d$, and let $\boldsymbol{\ell} \in \mathbb{R}^d_+$. We have*

$$\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})} \neq \emptyset \Leftrightarrow \boldsymbol{\ell} \prec_S \hat{\boldsymbol{\sigma}}. \tag{148}$$

The proof by Horn (1954) of the reciprocal in Theorem 36 is non constructive. In the next section, we survey algorithms that output an element of $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$.

E.2 Related work

Several articles (Raskutti and Mahoney, 2016, Ma et al., 2015) in the randomized linear algebra community propose the use of non Gaussian random matrices to generate matrices with a fast decreasing profile of leverage scores (so-called *heavy hitters*) without controlling the exact profile of the leverage scores.

Dhillon et al. (2005) showed how to generate matrices from $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$ using Givens rotations; see the algorithm in Figure 10. The idea of the algorithm is to start with a frame with the exact spectrum and repeatedly apply orthogonal matrices (Lines 4 and 6 of Figure 10) that preserve the spectrum while changing the leverage scores of only two columns, setting one of their leverage scores to the desired value. The orthogonal matrices are the so-called *Givens rotations*.

Definition 37 *Let $\theta \in [0, 2\pi[$ and $i, j \in [d]$. The Givens rotation $\boldsymbol{G}_{i,j}(\theta) \in \mathbb{R}^{d\times d}$ is defined by*

$$\boldsymbol{G}_{i,j}(\theta) = \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos(\theta) & & & & -\sin(\theta) & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & \sin(\theta) & & & & \cos(\theta) & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix}. \tag{149}$$

Figure 11 shows the output of the algorithm in Figure 10, for the input $(\boldsymbol{\ell}, \boldsymbol{\sigma}) = (\boldsymbol{\ell}, \mathbb{1})$ for three different values of $\boldsymbol{\ell}$. The main drawbacks of this algorithm are first that it is deterministic, so that it outputs a unique matrix $\boldsymbol{F}$ for a given input $(\boldsymbol{\ell}, \boldsymbol{\sigma})$, and second that the output is a highly structured matrix, as observed on Figure 11.

We propose an algorithm that outputs random, more "generic" matrices belonging to $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$. This algorithm is based on a parametrization of $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$ using the

GivensAlgorithm$(\boldsymbol{\ell}, \boldsymbol{\sigma})$

$\boldsymbol{F} \longleftarrow \left[\ \mathrm{Diag}(\boldsymbol{\sigma}) \ \middle|\ \boldsymbol{0}\ \right] \in \mathbb{R}^{k\times d}$
while $\exists i,j,k \in [d],\ i<k<j : \|\boldsymbol{f}_i\|^2 < \ell_i, \|\boldsymbol{f}_k\|^2 = \ell_k, \|\boldsymbol{f}_j\|^2 > \ell_j$
  if $\ell_i - \|\boldsymbol{f}_i\|^2 \le \|\boldsymbol{f}_j\|^2 - \ell_j$
    $\boldsymbol{F} \leftarrow \boldsymbol{G}_{i,j}(\theta)\boldsymbol{F}$, where $\|(\boldsymbol{G}_{i,j}(\theta)\boldsymbol{F})_i\|^2 = \ell_i$.
  else
    $\boldsymbol{F} \leftarrow \boldsymbol{G}_{i,j}(\theta)\boldsymbol{F}$, where $\|(\boldsymbol{G}_{i,j}(\theta)\boldsymbol{F})_j\|^2 = \ell_j$,
return $\boldsymbol{F} \in \mathbb{R}^{k\times d}$.

Figure 10: The pseudocode of the algorithm proposed by Dhillon et al. (2005) for generating a matrix given its leverage scores and spectrum by successively applying Givens rotations.

collection of spectra of all minors of $\boldsymbol{F} \in \mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$. This parametrization was introduced by Fickus et al. (2011b), and we recall it in Section E.3. For now, let us simply look at Figure 12, which displays a few outputs of our algorithm for the same input as in Figure 11a. We now obtain different matrices for the same input $(\boldsymbol{\ell}, \boldsymbol{\sigma})$, and these matrices are less structured than the output of Algorithm 10, as required.

E.3 The restricted Gelfand-Tsetlin polytope

Definition 38 *Recall that $(\boldsymbol{f}_i)_{i\in[d]}$ are the columns of the matrix $\boldsymbol{F} \in \mathbb{R}^{k\times d}$. For $r \in [d]$, we further define*

$$\boldsymbol{F}_r = \boldsymbol{F}_{:,[r]} \in \mathbb{R}^{k\times r}, \tag{150}$$

$$\boldsymbol{C}_r = \sum_{i\in[r]} \boldsymbol{f}_i \boldsymbol{f}_i^\mathsf{T} \in \mathbb{R}^{k\times k}, \tag{151}$$

$$\boldsymbol{G}_r = \boldsymbol{F}_r^\mathsf{T} \boldsymbol{F}_r \in \mathbb{R}^{r\times r}. \tag{152}$$

Furthermore, we note for $r \in [d]$,

$$(\lambda_{r,i})_{i\in[k]} = \Lambda(\boldsymbol{C}_r), \tag{153}$$

$$(\tilde{\lambda}_{r,i})_{i\in[r]} = \Lambda(\boldsymbol{G}_r). \tag{154}$$

The $(\lambda_{r,i})_{i\in[k]}$, $r \in [d]$, are called the outer eigensteps of $\boldsymbol{F}$, and we group them in the matrix

$$\Lambda^{out}(\boldsymbol{F}) = (\lambda_{r,i})_{i\in[k],r\in[d]} \in \mathbb{R}^{k\times d}.$$

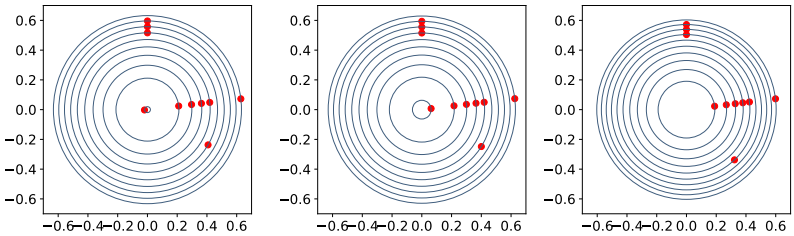
Similarly, the $(\tilde{\lambda}_{r,i})_{i\in[r]}$ are called inner eigensteps of $\boldsymbol{F}$.

Example 2 *For $k = 2$, $d = 4$, consider the full-rank matrix*

$$\boldsymbol{F} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \tag{155}$$

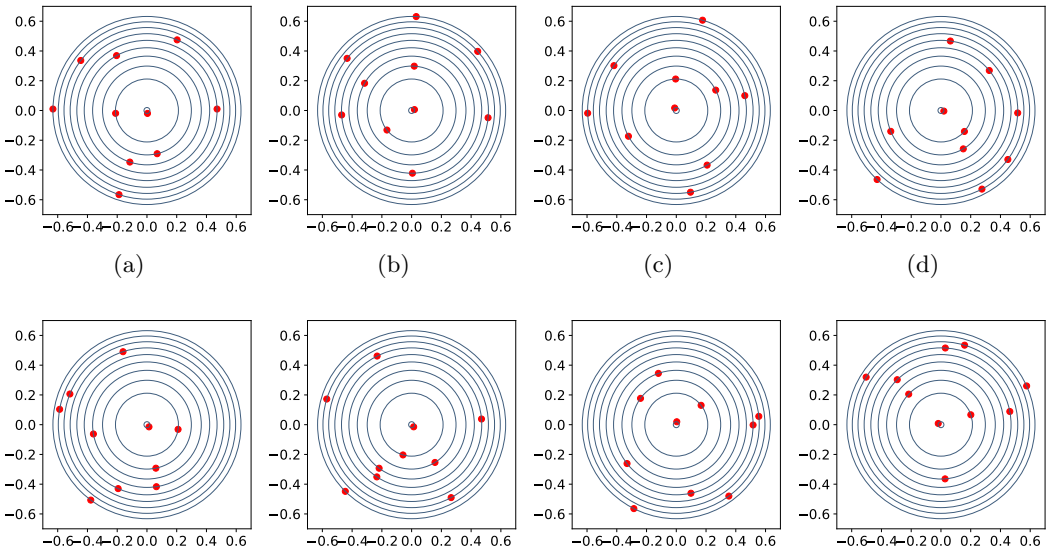
Then

$$\Lambda^{out}(\boldsymbol{F}) = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 2 \end{bmatrix}. \tag{156}$$



(a) (b) (c)

Figure 11: The output of the algorithm in Figure 10 for $k = 2$, $d = 10$, $\boldsymbol{\sigma} = (1, 1)$, and three different values of $\boldsymbol{\ell}$ that each add to k. Each red dot has coordinates a column of $\boldsymbol{F}$. The blue circles have for radii the prescribed $(\sqrt{\ell_i})$.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 12: The output of our algorithm for $k = 2$, $d = 10$, an input $\boldsymbol{\sigma} = (1, 1)$, and ℓ as in Figure 11a. Each red dot has coordinates a column of $\boldsymbol{F}$. The blue circles have for radii the prescribed $(\sqrt{\ell_i})$.

Proposition 39 *The outer eigensteps satisfy the following constraints:*

$$\begin{cases} \forall i \in [k], \;\; \lambda_{0,i} = 0 \\ \forall i \in [k], \;\; \lambda_{d,i} = \sigma_i^2 \\ \forall r \in [d], \;\; (\lambda_{r,:}) \sqsubseteq (\lambda_{r+1,:}) \\ \forall r \in [d], \;\; \sum_{i \in [d]} \lambda_{r,i} = \sum_{i \in [r]} \ell_i \end{cases} . \tag{157}$$

In other words, the outer eigensteps are constrained to live in a polytope. We define the restricted Gelfand-Tsetlin polytope $\boldsymbol{GT}_{(k,d)}(\boldsymbol{\sigma}, \boldsymbol{\ell})$ to be the subset of $\mathbb{R}^{k \times d}$ defined by the equations (157). A more graphical summary of the interlacing and sum

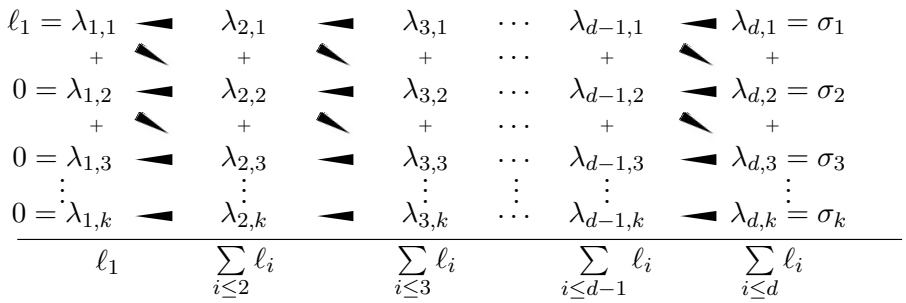


Figure 13: The interlacing relationships (157) satisfied by the outer eigensteps of a frame. Thick triangles are used in place of $\leq$ for improved readability.

constraints is given in Figure 13. The restricted GT polytope[6] allows a parametrization of $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$ by the following reconstruction result.

Theorem 40 (Theorem 3, Fickus et al., 2011a) *Every matrix $\boldsymbol{F} \in \mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$ can be constructed as follows:*

- *pick a valid sequence of outer eigensteps noted $\Lambda^{out} \in \boldsymbol{GT}_{(k,d)}(\boldsymbol{\sigma},\boldsymbol{\ell})$,*
- *pick $\boldsymbol{f}_1 \in \mathbb{R}^k$ such that*
$$\|\boldsymbol{f}_1\|^2 = \ell_1, \tag{158}$$
- *for $r \in [d]$, consider the polynomial $p_r(x) = \prod_{i\in[d]} (x - \lambda_{r,i})$, and for each $r \in [d-1]$, choose $\boldsymbol{f}_{r+1} \in \mathbb{R}^k$ such that*
$$\forall \lambda \in \{\lambda_{r,i}\}_{i\in[d]},\ \|\boldsymbol{P}_{r,\lambda}\boldsymbol{f}_{r+1}\|^2 = -\lim_{x\to\lambda}(x-\lambda)\frac{p_{r+1}(\lambda)}{p_r(\lambda)}, \tag{159}$$
where $\boldsymbol{P}_{r,\lambda}$ denotes the orthogonal projection onto the eigenspace $\mathrm{Ker}(\lambda\mathbb{I}_k - \boldsymbol{F}_r\boldsymbol{F}_r^T)$.

Conversely, any matrix $\boldsymbol{F}$ constructed by this process is in $\mathcal{M}_{(\boldsymbol{\ell},\boldsymbol{\sigma})}$.

Fickus et al. (2011a) propose an algorithm to construct a vector $\boldsymbol{f}_r$ satisfying Equation (159). Finally, an algorithm for the construction of a valid sequence of eigensteps $\Lambda^{\mathrm{out}} \in \boldsymbol{GT}_{(k,d)}(\boldsymbol{\sigma},\boldsymbol{\ell})$ was proposed in (Fickus et al., 2011b). This yields the following constructive result.

Theorem 41 (Theorem 4.1, Fickus et al., 2011b) *Every matrix $\boldsymbol{F} \in \mathcal{M}(\boldsymbol{\sigma},\boldsymbol{\ell})$ can be constructed as follows:*

- *Set $\forall i \in [k]$, $\tilde{\lambda}_{d,i} = \sigma_i^2$,*
- *For $r \in \{d-1,\dots,1\}$, construct $\{\tilde{\lambda}_{r,:}\}$ as follows. For each $i \in \{k,\dots,1\}$, pick*
$$\tilde{\lambda}_{r-1,i} \in [B_{i,r}(\boldsymbol{\ell},\boldsymbol{\sigma}), A_{i,r}(\boldsymbol{\ell},\boldsymbol{\sigma})],$$

[6] Note the difference with the Gelfand-Tsetlin polytope in the random matrix literature (Baryshnikov, 2001), where only the spectrum is constrained, not the diagonal.

RandomEigensteps$(\boldsymbol{\ell}, \boldsymbol{\sigma})$

$\Lambda^{\text{out}} \longleftarrow \mathbb{0} \in \mathbb{R}^{k \times d}$

$\forall i \in [k],\ \tilde{\lambda}_{d,i} \longleftarrow \sigma_i$

for $r \in \{d-1, \dots, 1\}$

  for $i \in \{k, \dots, 1\}$

    Pick $\tilde{\lambda}_{r-1,i} \sim \mathcal{U}([B_{i,r}(\boldsymbol{\ell}, \boldsymbol{\sigma}), A_{i,r}(\boldsymbol{\ell}, \boldsymbol{\sigma})])$

return Λ^{out}

Figure 14: The pseudocode of the generator of random valid eigensteps taking as input $(\boldsymbol{\ell}, \boldsymbol{\sigma})$.

where

$$\begin{aligned} A_{i,r}(\boldsymbol{\ell}, \boldsymbol{\sigma}) &= \max\left\{\tilde{\lambda}_{r+1,i+1}, \sum_{t=i}^{k} \tilde{\lambda}_{r+1,t} - \sum_{t=i+1}^{k} \tilde{\lambda}_{r,t} - \ell_{r+1}\right\} \\ B_{i,r}(\boldsymbol{\ell}, \boldsymbol{\sigma}) &= \min\left\{\tilde{\lambda}_{r+1,i}, \min_{z=1,\dots,i}\left\{\sum_{t=z}^{r} \ell_t - \sum_{t=z+1}^{i} \tilde{\lambda}_{r+1,t} - \sum_{t=i+1}^{k} \tilde{\lambda}_{r,t}\right\}\right\}. \end{aligned} \tag{160}$$

Furthermore, any sequence constructed by this algorithm is a valid sequence of inner eigensteps.

Based on these results we propose an algorithm for the generation of orthogonal random matrices with a given profile of leverage scores.

E.4 Our algorithm

We consider a randomization of the algorithm given in Theorem 41. First, we generate a random sequence of valid inner eigensteps Λ^{in} using Algorithm 14. Then we proceed to the reconstruction a frame that admits Λ^{in} as a sequence of eigensteps using the Algorithm proposed in (Fickus et al., 2011a).

Note that Equations (158) and (159) admit several solutions. For example, for $r \in [d]$, and if $\boldsymbol{f}_{r+1}$ satisfies (159), $-\boldsymbol{f}_{r+1}$ satisfies this equation too. Fickus et al. (2011a) actually prove that the set of solutions of these equations is invariant under a specific action of the orthogonal group $\mathbb{O}(\rho(r,k))$ where $\rho(r,k) \in \mathbb{N}$ nontrivially depends on the eigensteps. In the reconstruction step of our algorithm, we apply a random Haar-distributed orthogonal matrix as soon as such an invariance is provable. Namely, we a random orthogonal matrix sampled from the Haar measure on $\mathbb{O}(d)$ to the vector $\boldsymbol{f}_1$ and, then, we apply an independent random orthogonal matrix sampled from the Haar measure on $\mathbb{O}(\rho(r,k))$ to each reconstructed vector $\boldsymbol{f}_{r+1}$.

Figure 12 displays a few samples from our algorithm, which display diversity and no apparent structure, as required for a generator of toy datasets. The question of fully characterizing the distribution of the output of our algorithm is an open question.