Can matrix coherence be efficiently and accurately estimated?

Mehryar Mohri
Courant Institute and Google Research
New York, NY
mohri@cs.nyu.edu

Ameet Talwalkar
Computer Science Division
University of California, Berkeley
ameet@eecs.berkeley.edu

Abstract

Matrix coherence has recently been used to characterize the ability to extract global information from a subset of matrix entries in the context of low-rank approximations and other sampling-based algorithms. The significance of these results crucially hinges upon the possibility of efficiently and accurately testing this coherence assumption. This paper precisely addresses this issue. We introduce a novel sampling-based algorithm for estimating coherence, present associated estimation guarantees and report the results of extensive experiments for coherence estimation. The quality of the estimation guarantees we present depends on the coherence value to estimate itself, but this turns out to be an inherent property of sampling-based coherence estimation, as shown by our lower bound. In practice, however, we find that these theoretically unfavorable scenarios rarely appear, as our algorithm efficiently and accurately estimates coherence across a wide range of datasets, and these estimates are excellent predictors of the effectiveness of sampling-based matrix approximation on a case-by-case basis. These results are significant as they reveal the extent to which coherence assumptions made in a number of recent machine learning publications are testable.

1 Introduction

Very large-scale datasets are increasingly prevalent in a variety of areas, e.g., computer vision, natural language processing, computational biology. However, several standard methods in machine learning, such as spectral clustering, manifold learning techniques, kernel ridge regression or other kernel-based algorithms do not scale to such orders of magnitude. For large datasets, these algorithms would require storage and operation on matrices with thousands to millions of columns and rows, which is especially problematic since these matrices are often not sparse. An attractive solution to such problems involves efficiently generating low-rank approximations to the original matrix of interest. In particular, sampling-based techniques that operate on a subset of the columns of the matrix can be effective solutions to this problem, and have been widely studied within the machine learning and theoretical computer science communities (Drineas *et al.*, 2006; Frieze *et al.*, 1998; Kumar *et al.*, 2009b; Williams and Seeger, 2000). In the context of kernel matrices, the Nyström method (Williams and Seeger, 2000) has been shown to work particularly well in practice for various applications ranging from manifold learning to image segmentation (Fowlkes *et al.*, 2004; Talwalkar *et al.*, 2008).

A crucial assumption of these algorithms involves their sampling-based nature, namely that an accurate low-rank approximation of some matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ can be generated exclusively from information extracted from a small subset ($l \ll m$) of its columns. This assumption is not generally true for all matrices, and explains the negative results of Fergus *et al.* (2009). For instance, consider the extreme case:

$$\mathbf{X} = \begin{bmatrix} | & & | & | & & | \\ \mathbf{e}_1 & \dots & \mathbf{e}_r & \mathbf{0} & \dots & \mathbf{0} \\ | & & | & | & & | \end{bmatrix}, \qquad (1)$$

where $\mathbf{e}_i$ is the ith column of the n dimensional identity matrix and $\mathbf{0}$ is the n dimensional zero vector. Although this matrix has rank r, it cannot be well approximated by a random subset of l columns unless this subset includes $\mathbf{e}_1, \dots, \mathbf{e}_r$. In order to account for such pathological cases, previous theoretical bounds relied on sampling columns of $\mathbf{X}$ in an adaptive fash-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

ion (Bach and Jordan, 2005; Deshpande *et al.*, 2006; Kumar *et al.*, 2009b; Smola and Schölkopf, 2000) or from non-uniform distributions derived from properties of $\mathbf{X}$ (Drineas and Mahoney, 2005; Drineas *et al.*, 2006). Indeed, these bounds give better guarantees for pathological cases, but are often quite loose nonetheless, e.g., when dealing with kernel matrices using RBF kernels, and these sampling schemes are rarely utilized in practice.

More recently, Talwalkar and Rostamizadeh (2010) used the notion of *coherence* to characterize the ability to extract information from a small subset of columns, showing theoretical and empirical evidence that coherence is tied to the performance of the Nyström method. Coherence measures the extent to which the singular vectors of a matrix are correlated with the standard basis. Intuitively, if the dominant singular vectors of a matrix are incoherent, then the subspace spanned by these singular vectors is likely to be captured by a random subset of sampled columns of the matrix. In fact, coherence-based analysis of algorithms has been an active field of research, starting with pioneering work on compressed sensing (Candès *et al.*, 2006; Donoho, 2006), as well as related work on matrix completion (Candès and Recht, 2009; Keshavan *et al.*, 2009b) and robust principle component analysis (Candès *et al.*, 2009).

In Candès and Recht (2009), the use of coherence is motivated by results showing that several classes of randomly generated matrices have low coherence with high probability, one of which is the class of matrices generated from uniform random orthonormal singular vectors and arbitrary singular values. Unfortunately, these results do not help a practitioner compute coherence on a case-by-case basis to determine whether attractive theoretical bounds hold for the task at hand. Furthermore, the coherence of a matrix is by definition derived from its singular vectors and is thus expensive to compute: the prohibitive cost of calculating singular values and singular vectors is precisely the motivation behind sampling-based techniques. Hence, in spite of the numerous theoretical work based on related notions of coherence, the practical significance of these results largely hinges on the following open question: *Can we efficiently and accurately estimate the coherence of a matrix?* In this paper, we address this question by presenting a novel algorithm for estimating matrix coherence from a small number of columns.

The remainder of this paper is organized as follows. Section 2.1 introduces basic definitions, and provides a brief background on low-rank matrix approximation and matrix coherence. In Section 3 we introduce our sampling-based algorithm to estimate matrix coherence. We then formally analyze its behavior in Section 4 presenting both upper and lower bounds on performance. We also use this analysis to derive a novel coherence-based bound for matrix projection reconstruction via Column-sampling (defined in Section 2.2). Finally, in Section 5 we present extensive experimental results on synthetic and real datasets. In contrast to our worst-case theoretical analysis in the previous section, these results provide strong support for the use of our proposed algorithm whenever sampling-based matrix approximation is being considered. Empirically, our algorithm effectively estimates matrix coherence across a wide range of datasets, and these coherence estimates are excellent predictors of the effectiveness of sampling-based matrix approximation on a case-by-case basis.

2 Background

2.1 Notation

Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ be an arbitrary matrix. We define $\mathbf{X}^{(j)}$, $j = 1 \dots m$, as the jth column vector of $\mathbf{X}$, $\mathbf{X}_{(i)}$, $i = 1 \dots n$, as the ith row vector of $\mathbf{X}$ and $\mathbf{X}_{ij}$ as the ijth entry of $\mathbf{X}$. Furthermore, $\mathbf{X}^{(i:j)}$ refers to the ith through jth columns of $\mathbf{X}$ and $\mathbf{X}_{(i:j)}$ refers to the ith through jth rows of $\mathbf{X}$. We denote by $\|\mathbf{X}\|_F$ the Frobenius norm of $\mathbf{X}$ and by $\|\mathbf{v}\|$ the l_2 norm of the vector $\mathbf{v}$. If $\text{rank}(\mathbf{X}) = r$, we can write the thin Singular Value Decomposition (SVD) as $\mathbf{X} = \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top$. $\mathbf{\Sigma}_X$ is diagonal and contains the singular values of $\mathbf{X}$ sorted in decreasing order, i.e., $\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \dots \geq \sigma_r(\mathbf{X})$. $\mathbf{U}_X \in \mathbb{R}^{n\times r}$ and $\mathbf{V}_X \in \mathbb{R}^{m\times r}$ have orthogonal columns that contain the left and right singular vectors of $\mathbf{X}$ corresponding to its singular values. We define $\mathbf{P}_X = \mathbf{U}_X\mathbf{U}_X^\top$ as the orthogonal projection matrix onto the column space of $\mathbf{X}$, and denote the projection onto its orthogonal complement as $\mathbf{P}_{X,\perp} = \mathbf{I} - \mathbf{P}_X$. We further define $\mathbf{X}^+ \in \mathbb{R}^{m\times n}$ as the Moore-Penrose pseudoinverse of $\mathbf{X}$, with $\mathbf{X}^+ = \mathbf{V}_X\mathbf{\Sigma}_X^+\mathbf{U}_X^\top$. Finally, we define $\mathbf{K} \in \mathbb{R}^{n\times n}$ as a symmetric positive semidefinite (SPSD) matrix with $\text{rank}(\mathbf{K}) = r \leq n$, i.e. a symmetric matrix with non-negative eigenvalues.

2.2 Low-rank matrix approximation

Starting with an $n \times m$ matrix $\mathbf{X}$, we are interested in algorithms that generate a low-rank approximation, $\widetilde{\mathbf{X}}$, from a sample of $l \ll n$ of its columns. The accuracy of this approximation is often measured using the Frobenius $\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F$ or the Spectral distance $\|\mathbf{X} - \widetilde{\mathbf{X}}\|_2$. We next briefly describe two of the most common algorithms of this form, the Column-sampling and the Nyström methods.

The Column-sampling method generates approxima-

tions to arbitrary rectangular matrices. We first sample l columns of $\mathbf{X}$ such that $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$, where $\mathbf{X}_1$ has l columns, and then use the SVD of $\mathbf{X}_1$, $\mathbf{X}_1 = \mathbf{U}_{X_1}\mathbf{\Sigma}_{X_1}\mathbf{V}_{X_1}^\top$, to approximate the SVD of $\mathbf{X}$ (Frieze *et al.*, 1998). This method is most commonly used to generate a 'matrix projection' approximation (Kumar *et al.*, 2009b) of $\mathbf{X}$ as follows:

$$\widetilde{\mathbf{X}}^{col} = \mathbf{U}_{X_1}\mathbf{U}_{X_1}^\top\mathbf{X}. \tag{2}$$

The runtime of the Column-sampling method is dominated by the SVD of $\mathbf{X}_1$ which takes $\mathrm{O}(nl^2)$ time to perform and is feasible for small l.

In contrast to the Column-sampling method, the Nyström method deals only with SPSD matrices. We start with an $n \times n$ SPSD matrix, sampling l columns such that $\mathbf{K} = \begin{bmatrix}\mathbf{K}_1 & \mathbf{K}_2\end{bmatrix}$, where $\mathbf{K}_1$ has l columns, and define $\mathbf{W}$ as the $l \times l$ matrix consisting of the intersection of these l columns with the corresponding l rows of $\mathbf{K}$. Since $\mathbf{K}$ is SPSD, $\mathbf{W}$ is also SPSD. Without loss of generality, we can rearrange the columns and rows of $\mathbf{K}$ based on this sampling such that:

$$\mathbf{K} = \begin{bmatrix}\mathbf{W} & \widehat{\mathbf{K}}_1^\top \\ \widehat{\mathbf{K}}_1 & \widehat{\mathbf{K}}_2\end{bmatrix} \tag{3}$$

where

$$\mathbf{K}_1 = \begin{bmatrix}\mathbf{W} \\ \widehat{\mathbf{K}}_1\end{bmatrix} \quad \text{and} \quad \mathbf{K}_2 = \begin{bmatrix}\widehat{\mathbf{K}}_1^\top \\ \widehat{\mathbf{K}}_2\end{bmatrix}. \tag{4}$$

The Nyström method uses $\mathbf{W}$ and $\mathbf{K}_1$ from (3) to generate a 'spectral reconstruction' approximation of $\mathbf{K}$ as $\widetilde{\mathbf{K}}^{nys} = \mathbf{K}_1\mathbf{W}^+\mathbf{K}_1^\top$. Since the running time complexity of SVD on $\mathbf{W}$ is in $O(l^3)$ and matrix multiplication with $\mathbf{K}_1$ takes $O(nl^2)$, the total complexity of the Nyström approximation computation is also in $O(nl^2)$.

2.3 Matrix Coherence

Matrix coherence measures the extent to which the singular vectors of a matrix are correlated with the standard basis. As previously mentioned, coherence has been used to analyze techniques such as compressed sensing, matrix completion, robust PCA, and the Nyström method. These analyses have used a variety of related notions of coherence. If we let $\mathbf{e}_i$ be the ith column of the standard basis, we can define three basic notions of coherence as follows:

Definition 1 (μ-Coherence)**.** *Let $\mathbf{U} \in \mathbb{R}^{n\times r}$ contain orthonormal columns with $r < n$. Then the μ-coherence of $\mathbf{U}$ is:*

$$\mu(\mathbf{U}) = \sqrt{n}\max_{i,j}\left|\mathbf{U}_{ij}\right|. \tag{5}$$

Definition 2 (μ_0-Coherence)**.** *Let $\mathbf{U} \in \mathbb{R}^{n\times r}$ contain orthonormal columns with $r < n$ and define $\mathbf{P}_U = \mathbf{U}\mathbf{U}^\top$ as its associated orthogonal projection matrix. Then the μ_0-coherence of $\mathbf{U}$ is:*

$$\mu_0(\mathbf{U}) = \frac{n}{r}\max_{1\le i\le n}\|\mathbf{P}_U\mathbf{e}_i\|^2 = \frac{n}{r}\max_{1\le i\le n}\|\mathbf{U}_{(i)}\|^2. \tag{6}$$

Definition 3 (μ_1-Coherence)**.** *Given the matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ with rank r, left and right singular vectors, $\mathbf{U}_X$ and $\mathbf{V}_X$, and define $\mathbf{T} = \sum_{1\le k\le r}\mathbf{U}_X^{(k)}{\mathbf{V}_X^{(k)}}^\top$. Then, the μ_1-coherence of $\mathbf{X}$ is:*

$$\mu_1(\mathbf{X}) = \sqrt{\frac{nm}{r}}\max_{ij}\left|\mathbf{T}_{ij}\right|. \tag{7}$$

In Talwalkar and Rostamizadeh (2010), $\mu(\mathbf{U})$ is used to provide coherence-based bounds for the Nyström method, where $\mathbf{U}$ corresponds to the singular vectors of a low-rank SPSD kernel matrix. Low-rank matrices are also the focus of work on matrix completion by Candès and Recht (2009) and Keshavan *et al.* (2009b), though they deal with more general rectangular matrices with SVD $\mathbf{X} = \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top$, and they use $\mu_0(\mathbf{U}_X)$, $\mu_0(\mathbf{V}_X)$ and $\mu_1(\mathbf{X})$ to bound the performance of two different matrix completion algorithms. Note that a stronger, more complex notion of coherence is used in Candès and Tao (2009) to provide tighter bounds for the matrix completion algorithm presented in Candès and Recht (2009) (definition omitted here). Moreover, coherence has also been used to analyze algorithms dealing with low-rank matrices in the presence of noise, e.g., Candès and Plan (2009); Keshavan *et al.* (2009a) for noisy matrix completion and Candès *et al.* (2009) for robust PCA. In these analyses, the coherence of the underlying low-rank matrix once again appears in the form of $\mu_0(\cdot)$ and $\mu_1(\cdot)$.

In this work, we choose to focus on μ_0. In comparison to μ, μ_0 is a more robust measure of coherence, as it deals with row norms of $\mathbf{U}$, rather than the maximum entry of $\mathbf{U}$, and the two notions are related by a simple pair of inequalities: $\mu^2/r \le \mu_0 \le \mu^2$. Furthermore, since we focus on coherence in the context of algorithms that sample columns of the original matrix, μ_0 is a more natural choice than μ_1, since existing coherence-based bounds for these algorithms (both in Talwalkar and Rostamizadeh (2010) and in Section 4 of this work) only depend on the left singular vectors of the matrix.

3 Estimate-Coherence Algorithm

As discussed in the previous section, matrix coherence has been used to analyze a variety of algorithms, under the assumption that the input matrix is either exactly

Input: $n \times l$ matrix ($\mathbf{X}_1$) storing l columns of arbitrary $n \times m$ matrix $\mathbf{X}$, low-rank parameter (r)
Output: An estimate of the coherence of $\mathbf{X}$

ESTIMATE-COHERENCE($\mathbf{X}_1, r$)
1 $\mathbf{U}_{X_1} \leftarrow \text{SVD}(\mathbf{X}_1)$ ▷ keep left singular vectors
2 $q \leftarrow \min\big(\text{rank}(\mathbf{X}_1), r\big)$
3 $\widetilde{\mathbf{U}} \leftarrow$ TRUNCATE($\mathbf{U}_{X_1}, q$) ▷ keep top q singular vectors of $\mathbf{X}_1$
4 $\gamma(\mathbf{X}_1) \leftarrow$ CALCULATE-GAMMA($\widetilde{\mathbf{U}}$) ▷ see equation (8)
5 **return** $\gamma(\mathbf{X}_1)$

Figure 1: The proposed sampling-based algorithm to estimate matrix coherence. Note that r is only required when $\mathbf{X}$ is perturbed by noise.

low-rank or is low-rank with the presence of noise. In this section, we present a novel algorithm to estimate the coherence of matrices following the same assumption. Starting with an arbitrary $n \times m$ matrix, $\mathbf{X}$, we are ultimately interested in an estimate of $\mu_0(\mathbf{U}_X)$, which contains the scaling factor n/r as shown in Definition 2. However, our estimate will also involve singular vectors in dimension n, and as we mentioned above, r is assumed to be small. Hence, neither of these scaling terms has a significant impact on our estimation. As such, our algorithm focuses on the closely related expression:

$$\gamma(\mathbf{U}) = \max_{1 \le i \le n} \|\mathbf{P}_U \mathbf{e}_i\|^2 = \frac{r}{n}\mu_0 \,. \tag{8}$$

Our proposed algorithm is quite similar in flavor to the Column-sampling algorithm discussed in Section 2.2. It estimates coherence by first sampling l columns of the matrix and subsequently using the left singular vectors of this submatrix to obtain an estimate. Note that our algorithm applies both to exact low-rank matrices as well as low-rank matrices perturbed by noise. In the latter case, the algorithm requires a user-defined low-rank parameter r. The runtime of this algorithm is dominated by the singular value decomposition of the $n \times l$ submatrix, and hence is in $\text{O}(nl^2)$. The details of the ESTIMATE-COHERENCE algorithm are presented in Figure 1.

4 Theoretical Analysis

In this section we analyze the performance of ESTIMATE-COHERENCE when used with low-rank matrices. In Section 4.1, we present an upper bound on the convergence of our algorithm and we detail the proof of this bound in Section 4.3. In Section 4.2 we present a lower bound using an adversarially constructed class of matrices.

4.1 Upper Bound

The upper bound presented in Theorem 1 shows that ESTIMATE-COHERENCE produces a monotonically increasing estimate of $\gamma(\cdot)$, and the convergence rate of the estimate is a function of coherence.

Theorem 1 (Upper Bound)**.** *Define $\mathbf{X} \in \mathbb{R}^{n \times m}$ with $\text{rank}(\mathbf{X}) = r \ll n$, and denote by $\mathbf{U}_X$ the r left singular vectors of $\mathbf{X}$ corresponding to its non-zero singular values. Let $\mathbf{X}_1$ be a set of l columns of $\mathbf{X}$ sampled uniformly at random, let the orthogonal projection onto span($\mathbf{X}_1$) be denoted by $\mathbf{P}_{X_1} = \mathbf{U}_{X_1}\mathbf{U}_{X_1}^\top$ and define the projection onto its orthogonal complement as $\mathbf{P}_{X_1,\perp}$. Let $\mathbf{x}$ be a column of $\mathbf{X}$ that is not in $\mathbf{X}_1$ that is sampled uniformly at random. Then the following statements can be made about $\gamma(\mathbf{X}_1)$, which is the output of* ESTIMATE-COHERENCE*($\mathbf{X}_1$):*

1. *$\gamma(\mathbf{X}_1)$ is a monotonically increasing estimate of $\gamma(\mathbf{X})$. Furthermore, if $\mathbf{X}_1' = \begin{bmatrix}\mathbf{X}_1 & \mathbf{x}\end{bmatrix}$ with $\mathbf{x}_\perp = \mathbf{P}_{X_1,\perp}\mathbf{x}$, then $0 \le \gamma(\mathbf{X}_1') - \gamma(\mathbf{X}_1) \le \gamma(\mathbf{z})$, where $\mathbf{z} = \mathbf{x}_\perp / \|\mathbf{x}_\perp\|$.*
2. *$\gamma(\mathbf{X}_1) = \gamma(\mathbf{X})$ when $\text{rank}(\mathbf{X}_1) = \text{rank}(\mathbf{X})$. For any $\delta > 0$, this equality holds with probability $1-\delta$ for $l \ge r^2\mu_0(\mathbf{U}_X)\max\big(C_1\log(r), C_2\log(3/\delta)\big)$ for positive constants C_1 and C_2.*

The second statement in Theorem 1 leads to Corollary 1, which relates matrix coherence to the performance of the Column-sampling algorithm when used for matrix projection on a low-rank matrix.

Corollary 1. *Assume the same notation as defined in Theorem 1, and let $\widetilde{\mathbf{X}}^{col}$ be the matrix projection approximation generated by the Column-sampling method using $\mathbf{X}_1$, as described in (2). Then, for any $\delta > 0$, $\widetilde{\mathbf{X}}^{col} = \mathbf{X}$ with probability $1-\delta$, for $l \ge r^2\mu_0(\mathbf{U}_X)\max\big(C_1\log(r), C_2\log(3/\delta)\big)$ for positive constants C_1 and C_2.*

Proof. When $\text{rank}(\mathbf{X}_1) = \text{rank}(\mathbf{X})$, the columns of $\mathbf{X}_1$

span the columns of $\mathbf{X}$. Hence, when this event occurs, projecting $\mathbf{X}$ onto the span of the columns of $\mathbf{X}_1$ leaves $\mathbf{X}$ unchanged. The second statement in Theorem 1 bounds the probability of this event. □

4.2 Lower Bound

Theorem 1 suggests that the ability to estimate matrix coherence is dependent on the coherence of the matrix itself. The following result proves that this is in fact the case: it shows for any large γ_0 the existence of matrices $\mathbf{X}$ with $\gamma(\mathbf{X}) = \gamma_0$, for which an estimate $\gamma(\mathbf{X}_1)$ based on a random sample $\mathbf{X}_1$ is almost always significantly different from $\gamma(\mathbf{X})$.

Theorem 2 (Lower Bound)**.** *Fix positive integers n, m and r, with $r \ll \min(n, m)$ and let $\frac{C\bar{r}}{n} \ll \gamma_0 \leq 1$, where $\bar{r} = \max(r, \log n)$ and C is a constant. Then, there exists a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ with $\operatorname{rank}(\mathbf{X}) = r$ and $\gamma(\mathbf{X}) = \gamma_0$ such that the following holds for any set of l columns, $\mathbf{X}_1$, sampled from $\mathbf{X}$:*

$$\begin{cases} \gamma(\mathbf{X}_1) \leq C\frac{\bar{r}}{n} & \text{if } \mathbf{X}_1 \text{ does not include } \mathbf{X}^{(1)}, \\ \gamma(\mathbf{X}_1) = \gamma_0 & \text{otherwise.} \end{cases} \tag{9}$$

Proof. Let $\mathbf{X}_0$ be a matrix formed by r orthonormal $n-1$ dimensional vectors such that $\gamma(\mathbf{X}_0) \leq C\bar{r}/n$. Such a matrix exists. In fact, by Lemma 2.2 of Candès and Recht (2009) and the so-called 'random orthogonal model', sampling uniformly at random from the set of all possible r orthonormal vectors leads to a matrix $\mathbf{X}_0$ with $\gamma(\mathbf{X}_0) \leq C\bar{r}/n$, with high probability.

Next, we rescale the first column of $\mathbf{X}_0$ such that $\|\mathbf{X}_0^{(1)}\|^2 = 1 - \gamma_0$ and let $\mathbf{v}$ be an r dimensional vector with $\mathbf{v}_1 = \sqrt{\gamma_0}$ and $\mathbf{v}_i = 0$ for $i > 1$. To construct $\mathbf{X}$ with properties described in the statement of the theorem, we first let $\mathbf{X}^{(r+1:m)}$ be all zeros. We then set the first row of $\mathbf{X}^{(1:r)}$ equal to $\mathbf{v}^\top$, and set the remaining $(n-1) \times r$ submatrix equal to $\mathbf{X}_0$. Overall, the construction is:

$$\mathbf{X} = \begin{bmatrix} \sqrt{\gamma_0} & \mathbf{0} & | & & | \\ & & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{X}_0^{(1)} & \mathbf{X}_0^{(2:r)} & | & & | \end{bmatrix}. \tag{10}$$

Observe that the first r columns of $\mathbf{X}$ are its top left singular vectors. Now, for a sample $\mathbf{X}_1$ extracted from $\mathbf{X}$, $\gamma(\mathbf{X}_1)$ has precisely the properties indicated in the statement of the theorem. □

Theorem 2 implies that in the worst case, all columns of the original matrix could be required when sampling randomly, and this lower bound on the number of samples holds for all column-sampling based methods that rely on the coherence of the sample to generate an estimate.

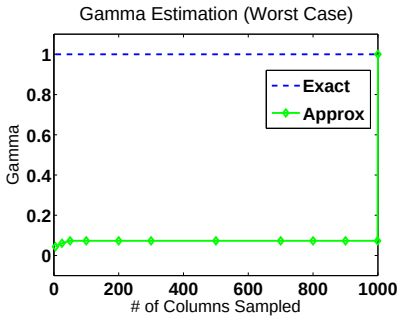

Figure 2: Synthetic dataset illustrating worst-case performance of Estimate-Coherence.

A simple and extreme unfavorable case is illustrated by Figure 2 based on the following construction: generate a synthetic matrix with $n = 1000$ and $k = 50$ using the Rand function in Matlab, and then replace its first diagonal with an arbitrarily large value, leading to a very high coherence matrix. Then, estimating coherence using Estimate-Coherence with a sample that does not include the first column of the matrix cannot be successful, as illustrated in Figure 2.

4.3 Proof of Theorem 1

We first present Lemmas 1 and 2, and then complete the proof of Theorem 1 using these lemmas.

Lemma 1. *Assume the same notation as defined in Theorem 1. Further, let $\mathbf{P}_{X_1'}$ be the orthogonal projection onto span($\mathbf{X}_1'$) and define $s = \|\mathbf{x}_\perp\|$. Then, for any $l \in [1, n-1]$, the following equalities relate the projection matrix $\mathbf{P}_{X_1'}$ to $\mathbf{P}_{X_1}$:*

$$\mathbf{P}_{X_1'} = \begin{cases} \mathbf{P}_{X_1} + \mathbf{z}\mathbf{z}^\top & \text{if } s > 0; \\ \mathbf{P}_{X_1} & \text{if } s = 0. \end{cases} \tag{11}$$

Proof. First assume that $s = 0$, which implies that $\mathbf{x}$ is in the span of the columns of $\mathbf{X}_1$. Since orthogonal projections are unique, then clearly $\mathbf{P}_{X_1'} = \mathbf{P}_{X_1}$ in this case. Next, assume that $s > 0$, in which case the span of the columns of $\mathbf{X}_1'$ can be viewed as the subspace spanned by the columns of $\mathbf{X}_1$ along with the subspace spanned by the residual of $\mathbf{x}$, i.e., $\mathbf{x}_\perp$. Observe that $\mathbf{z}\mathbf{z}^\top$ is the orthogonal projection onto span($\mathbf{x}_\perp$). Since these two subspaces are orthogonal and since orthogonal projection matrices are unique, we can write $\mathbf{P}_{X_1'}$ as the sum of orthogonal projections onto these subspaces, which matches the statement of the lemma for $s > 0$. □

Lemma 2. *Assume the same notation as defined in Theorem 1. Then, if $l \geq$*

$r^2\mu_0(\mathbf{U}_X)\max\big(C_1\log(r), C_2\log(3/\delta)\big)$, *where* C_1 *and* C_2 *are positive constants, then for any* $\delta > 0$, *with probability at least* $1-\delta$, $\text{rank}(\mathbf{X}_1) = r$.

Proof. Assuming uniform sampling at random, Talwalkar and Rostamizadeh (2010) shows that $\Pr[\text{rank}(\mathbf{X}_1) = r] \geq \Pr\left(\|c\mathbf{V}_{X,l}^\top \mathbf{V}_{X,l} - \mathbf{I}\|_2 < 1\right)$ for any $c \geq 0$, where $\mathbf{V}_{X,l} \in \mathbb{R}^{l\times r}$ corresponds to the first l components of the r right singular vectors of $\mathbf{X}$. Applying Theorem 1.2 in Candès and Romberg (2007) and using the identity $r\mu_0 \geq \mu^2$ yields the statement of the lemma. □

Now, to prove Theorem 1 we analyze the difference:

$$\begin{aligned}\Delta_l &= \left|\gamma(\mathbf{X}_1') - \gamma(\mathbf{X}_1)\right| \\ &= \left|\max_j \mathbf{e}_j^\top \mathbf{P}_{X_1'}\mathbf{e}_j - \max_i \mathbf{e}_i^\top \mathbf{P}_{X_1}\mathbf{e}_i\right|. \qquad (12)\end{aligned}$$

If $s = \|\mathbf{x}_\perp\| = 0$, then by Lemma 1, $\Delta_l = 0$. If $s > 0$, then using Lemma 1 and (12) yields:

$$\begin{aligned}\Delta_l &= \max_j \mathbf{e}_j^\top\left(\mathbf{P}_{X_1} + \mathbf{z}\mathbf{z}^\top\right)\mathbf{e}_j - \max_i \mathbf{e}_i^\top \mathbf{P}_{X_1}\mathbf{e}_i \qquad (13)\\ &\leq \max_j \mathbf{e}_j^\top \mathbf{z}\mathbf{z}^\top \mathbf{e}_j = \gamma(\mathbf{z}). \qquad (14)\end{aligned}$$

In (13), we use the fact that orthogonal projections are always SPSD, which means that $\mathbf{e}_j^\top \mathbf{z}\mathbf{z}^\top \mathbf{e}_j \geq 0$ for all j and ensures that $\Delta_l \geq 0$. In (14) we decouple the $\max(\cdot)$ over $\mathbf{P}_{X_1}$ and $\mathbf{z}\mathbf{z}^\top$ to obtain the inequality and then apply the definition of $\gamma(\cdot)$, which yields the first statement of Theorem 1. Finally, the second statement of Theorem 1 follows directly from Lemma 1 when $s = 0$ along with Lemma 2, as the former shows that $\Delta_l = 0$ if $\text{rank}(\mathbf{X}_1) = \text{rank}(\mathbf{X})$ and the latter gives a coherence-based finite-sample bound on the probability of this event occurring.

5 Experiments

In contrast to the lower bound presented in Section 4.2, our extensive empirical studies show that Estimate-Coherence performs quite well in practice on a variety of synthetic and real datasets with varying coherence, suggesting that the adversarial matrices used in the lower bounds are rarely encountered in practice. We present these empirical results in this section.

5.1 Experiments with synthetic data

We first generated low-rank synthetic matrices with varying coherence and singular value spectra, with $n = m = 1000$, and $r = 50$. To control the low-rank structure of the matrix, we generated datasets with exponentially decaying eigenvalues with differing decay rates, i.e., for $i \in \{1, \ldots, r\}$ we defined the ith singular value as $\sigma_i = \exp(-i\eta)$, where η controls the rate of decay and $\eta_{slow} = .01$, $\eta_{medium} = .1$, $\eta_{fast} = .5$. To control coherence, we independently generated left and right singular vectors with varying coherences by manually defining one singular vector and then using QR to generate $r-1$ additional orthogonal vectors. We associated this coherence-inducing singular vector with the $r/2$ largest singular value. We defined our 'low' coherence model by forcing the coherence-inducing singular vector to have minimal coherence, i.e., setting each component equal to $1/\sqrt{n}$. Using this as a baseline, we used 3 and 8 times this baseline to generate 'mid' and 'high' coherences (see Figure 3(a)). We then used Estimate-Coherence with varying numbers of sampled columns to estimate matrix coherence. Results reported in Figure 3(b-d) are means and standard deviations of 10 trials for each value of l. Although the coherence estimate converges faster for the low coherence matrices, the results show that even in the high coherence matrices, Estimate-Coherence recovers the true coherence after sampling only r columns. Further, we note that the singular value spectrum influences the quality of the estimate. This observation is due to the fact that the faster the singular values decay, the greater the impact of the $r/2$ largest singular value, which is associated with the coherence-inducing singular vector, and hence the more likely it will be captured by sampled columns.

Next, we examined the scenario of low-rank matrices with noise, working with the 'MEDIUM' decaying matrices used in the low-rank experiments. To create a noisy matrix from each original low-rank matrix, we first used the QR algorithm to find a full orthogonal basis containing the r left singular vectors of the original matrix, and used it as our new left singular vectors (we repeated this procedure to obtain right singular vectors). We then defined each of the remaining $n-r$ singular values of our noisy matrix to equal some fraction of the rth singular value of the original matrix (0.1 for 'SMALL' noise and 0.9 'LARGE' noise). The performance of Estimate-Coherence on these noisy matrices is presented in Figure 3(e-f), where results are means and standard deviations of 10 trials for each value of l. The presence of noise clearly has a negative affect on performance, yet the estimates are quite accurate for $l = 2r$ in the 'LOW' noise scenario, and even for the high coherence matrices with 'LARGE' noise, the estimate is fairly accurate when $l \geq 4r$.

5.2 Experiments with real data

We next performed experiments using the datasets listed in Table 1. In these experiments, we implicitly assume that we are interested in the coherence of an underlying low-rank matrix that is perturbed by noise.

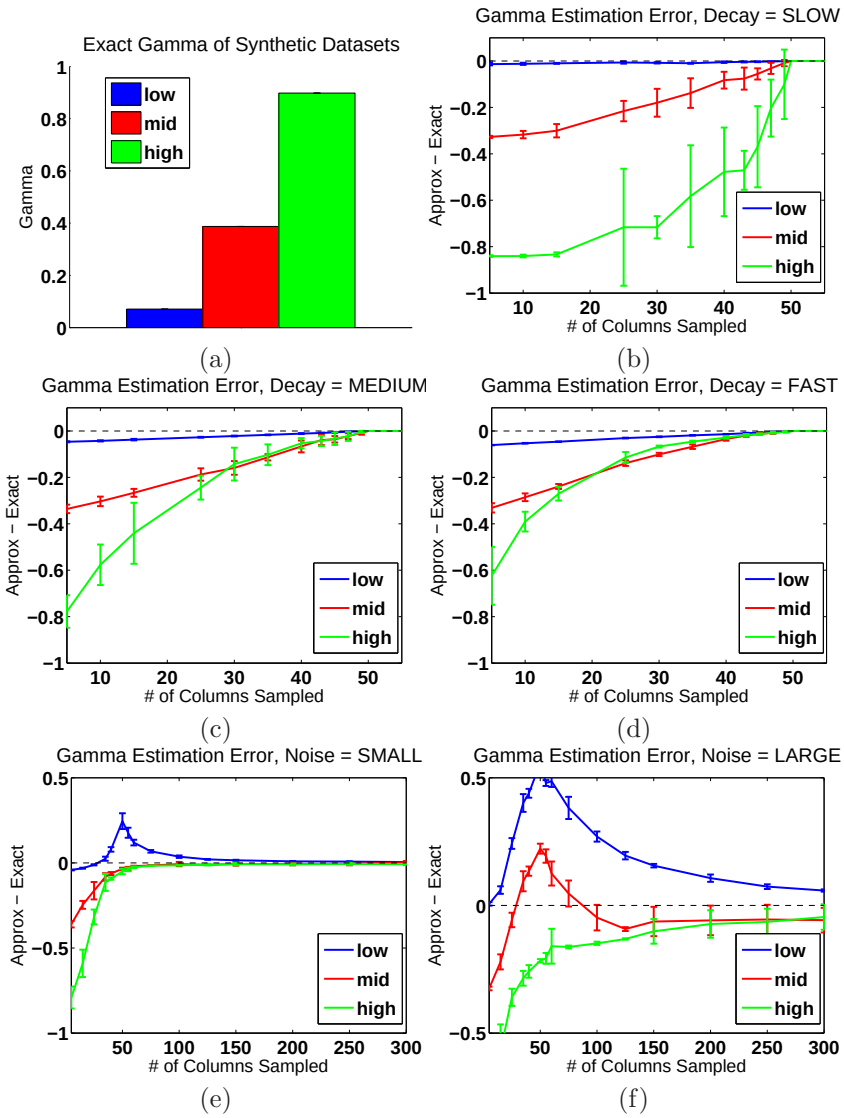

Figure 3: Experiments with synthetic matrices. (a) True coherence associated with 'low', 'mid' and 'high' coherences. (b-d) Exact low-rank experiments measuring difference between the exact coherence and the estimate by ESTIMATE-COHERENCE. (e-f) Experiments with low-rank matrices in the presence of noise, comparing exact and estimated coherence with two different levels of noise.

Dataset	Type of data	# Points (n)	# Features (d)	Kernel
NIPS	bag of words	1500	12419	linear
PIE	face images	2731	2304	linear
MNIS	digit images	4000	784	linear
Essential	proteins	4728	16	RBF
Abalone	abalones	4177	8	RBF
Dexter	bag of words	2000	20000	linear
KIN-8nm	kinematics of robot arm	2000	8	polynomial

Table 1: Description of real datasets used in our coherence experiments, including the type of data, the number of points (n), the number of features (d) and the choice of kernel (Asuncion and Newman, 2007; Gustafson *et al.*, 2006; LeCun and Cortes, 1998; Sim *et al.*, 2002).

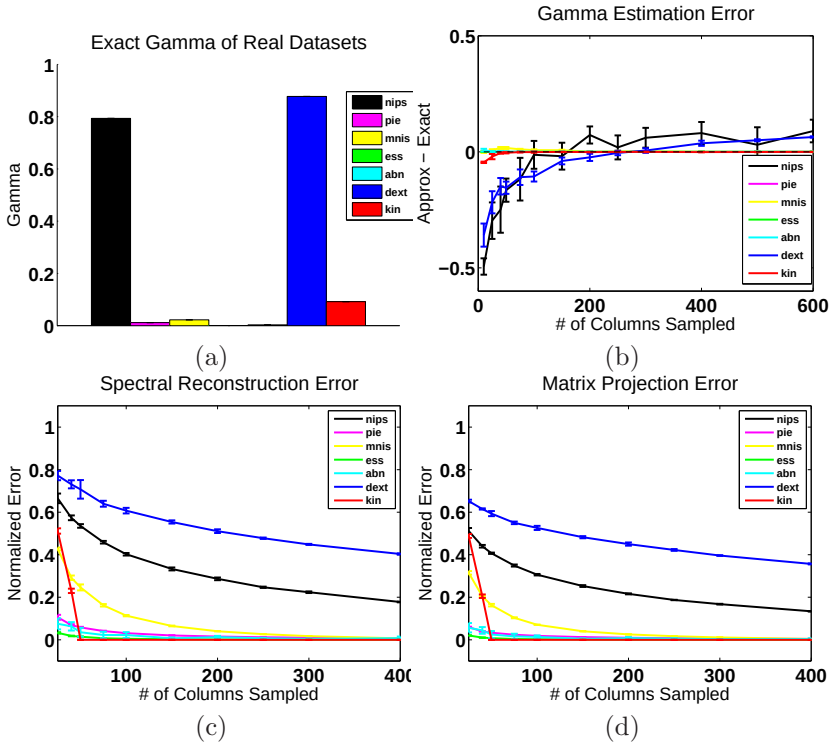

Figure 4: Experiments with real data. (a) True coherence of each kernel matrix $\mathbf{K}$. (b) Difference between the true coherence and the estimated coherence. (c-d) Quality of two types of low-rank matrix approximations ($\widetilde{\mathbf{K}}$), where 'Normalized Error' equals $\|\mathbf{K} - \widetilde{\mathbf{K}}\|_F / \|\mathbf{K}\|_F$.

We used a variety of kernel functions to generate SPSD kernel matrices from these datasets, with the resulting kernel matrices being quite varied in coherence (see Figure 4(a)). We used Estimate-Coherence with r set to equal the number of singular values needed to capture 99% of the spectral energy of each kernel matrix. Note that in practice, when we do not know the exact spectrum of the matrix, r can be estimated based on the spectrum of the sampled matrix.[1]

Figure 4(b) shows the estimation error over 10 trials. Although the coherence is well estimated across datasets when $l \geq 100$, the estimates for the two high coherence datasets (*nips* and *dext*) converge most slowly and exhibit the most variance across trials. Next, we performed spectral reconstruction using the Nyström method and matrix projection reconstruction using the Column-sampling method, and report results over 10 trials in Figure 4(c-d). The results clearly illustrate the connection between matrix coherence and the quality of these low-rank approximation techniques, as the two high coherence datasets exhibit significantly worse performance than the remaining datasets.

[1]The choice of r does indeed affect results, as can be seen by comparing the experimental results in this paper with those of Talwalkar and Rostamizadeh (2010) in which r is set to a fixed constant across all datasets, independent of the spectra of the various matrices.

6 Conclusion

We proposed a novel algorithm to estimate matrix coherence. Our theoretical analysis shows that Estimate-Coherence provides good estimates for relatively low-coherence matrices, and more generally, its effectiveness is tied to coherence itself. We corroborate this finding by presenting a lower bound derived from an adversarially constructed class of matrices. Empirically, however, our algorithm efficiently and accurately estimates coherence across a wide range of datasets, and these estimates are excellent predictors of the effectiveness of sampling-based matrix approximation. These results are quite significant as they reveal the extent to which coherence assumptions made in a number of recent machine learning publications are testable. We believe that our algorithm should be used whenever low-rank matrix approximation is being considered to determine its applicability on a case-by-case basis. Moreover, the variance of coherence estimates across multiple samples may provide further information, and the use of multiple samples fits nicely in the framework of ensemble methods for low-rank approximation, e.g., Kumar *et al.* (2009a).

References

A. Asuncion and D.J. Newman. UCI machine learning repository. `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 2007.

Francis R. Bach and Michael I. Jordan. Predictive low-rank decomposition for kernel methods. In *International Conference on Machine Learning*, 2005.

Emmanuel J. Candès and Yaniv Plan. Matrix completion with noise. `arXiv:0903.3131v1`, 2009.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–986, 2007.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. `arXiv:0903.1476v1[cs.IT]`, 2009.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? `arXiv:0912.3599v1[cs.IT]`, 2009.

Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Symposium on Discrete Algorithms*, 2006.

David L. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal of Computing*, 36(1), 2006.

Rob Fergus, Yair Weiss, and Antonio Torralba. Semi-supervised learning in gigantic image collections. In *Neural Information Processing Systems*, 2009.

Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the Nyström method. *Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. In *Foundation of Computer Science*, 1998.

A. Gustafson, E. Snitkin, S. Parker, C. DeLisi, and S. Kasif. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC:Genomics*, 7:265, 2006.

Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. In *Neural Information Processing Systems*, 2009.

Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion with a few entries. `arXiv:0901.3150v4[cs.LG]`, 2009.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nyström method. In *Neural Information Processing Systems*, 2009.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. On sampling-based approximate spectral decomposition. In *International Conference on Machine Learning*, 2009.

Yann LeCun and Corinna Cortes. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998.

Terence Sim, Simon Baker, and Maan Bsat. The CMU pose, illumination, and expression database. In *Conference on Automatic Face and Gesture Recognition*, 2002.

Alex J. Smola and Bernhard Schölkopf. Sparse Greedy Matrix Approximation for machine learning. In *International Conference on Machine Learning*, 2000.

Ameet Talwalkar and Afshin Rostamizadeh. Matrix coherence and the Nyström method. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

Ameet Talwalkar, Sanjiv Kumar, and Henry Rowley. Large-scale manifold learning. In *Conference on Vision and Pattern Recognition*, 2008.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Neural Information Processing Systems*, 2000.