SIAM J. MATH. DATA SCI.
Vol. 1, No. 1, pp. 144–160

Why Are Big Data Matrices Approximately Low Rank?*

Madeleine Udell† and Alex Townsend‡

Abstract. Matrices of (approximate) low rank are pervasive in data science, appearing in movie preferences, text documents, survey data, medical records, and genomics. While there is a vast literature on how to exploit low rank structure in these datasets, there is less attention paid to explaining why the low rank structure appears in the first place. Here, we explain the effectiveness of low rank models in data science by considering a simple generative model for these matrices: we suppose that each row or column is associated to a (possibly high dimensional) bounded latent variable, and entries of the matrix are generated by applying a piecewise analytic function to these latent variables. These matrices are in general full rank. However, we show that we can approximate every entry of an $m \times n$ matrix drawn from this model to within a fixed absolute error by a low rank matrix whose rank grows as $\mathcal{O}(\log(m+n))$. Hence any sufficiently large matrix from such a latent variable model can be approximated, up to a small entrywise error, by a low rank matrix.

Key words. big data, low rank matrices, Johnson–Lindenstrauss lemma

AMS subject classifications. 15A03, 68P99

DOI. 10.1137/18M1183480

1. Introduction.

Low rank matrices appear throughout the sciences. Numerous techniques have been developed to exploit low rank structure whenever it appears, whether in computational mathematics [7], computational fluid dynamics [40, 44, 11], genomics [10, 19, 27, 48], or, more surprisingly, movie preferences [18, 9], text documents [15, 16, 36], social networks [31, 34], medical records [39], and automated machine learning [50].

It is useful to know when a dataset can be approximated by a low rank matrix. A low rank approximation can be used to make filtering and statistics either computationally feasible or more efficient. In machine learning, low rank approximations to data tables are often employed to impute missing data, denoise noisy data, or perform feature extraction [45]. These techniques are also fundamental for many algorithms in recommender systems [28, 26] and can improve causal inference from survey data [25, 47, 5].

The broad applicability of low rank techniques is at first rather puzzling. Since the set of singular matrices is nowhere dense, random ("average") matrices are almost surely of full rank. In addition, the singular values of random Gaussian matrices are large with extraordinarily high probability [17]. We must conclude that matrices and datasets that appear in the real world must be far from average. We would like to understand the underlying phenomena that

*Received by the editors April 26, 2018; accepted for publication (in revised form) October 31, 2018; published electronically February 12, 2019.

http://www.siam.org/journals/simods/1-1/M118348.html

Funding: The work of the first author was supported by DARPA Award FA8750-17-2-0101. The work of the second author was supported by National Science Foundation grant 1818757.

†Department of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853 (udell@cornell.edu).

‡Department of Mathematics, Cornell University, Ithaca, NY 14853 (townsend@cornell.edu).

generate compressible datasets.

Let us begin with a statement about any matrix.

Theorem 1.0. *Let $X \in \mathbb{R}^{m\times n}$ with $m \geq n$ and $0 < \epsilon < 1$. Then, with $r = \lceil 72 \log(2n + 1)/\epsilon^2 \rceil$ we have*

$$\inf_{\text{rank}(Y)\leq r} \|X - Y\|_{\max} \leq \epsilon \|X\|_2, \tag{1.1}$$

where $\|\cdot\|_{\max}$ is the maximum absolute entry norm and $\|\cdot\|_2$ is the spectral matrix norm.

To avoid the theorem collapsing to a trivial statement, we need $r < n$, which only occurs when n is extremely large. Hence, roughly speaking, the theorem says that any sufficiently large matrix with a small spectral norm can be well approximated entrywise by a low rank matrix.

It is important to appreciate that Theorem 1.0 above holds for *any* matrix, including the identity matrix, and that the result is trivially false if $\|\cdot\|_{\max}$ is replaced by $\|\cdot\|_2$. Spectral norm approximations are generally preferred in linear algebra; however, for data science applications, an entrywise approximation is generally much more important. In a data science setting, one often wants to compress a dataset while perturbing each entry as little as possible—this is exactly what the maximum absolute entry norm captures.

Theorem 1.0 is simple, but the upper bound in (1.1) depends on $\|X\|_2$, which typically grows rapidly with n. However, a simple model for many matrices generated in data science explains why we should expect some matrices that appear in data science to have a small spectral norm. We suppose that X is generated by sampling columns and rows from a so-called nice latent variable model (intuitively, smooth; see Definition 4.1 for a formal definition) or a piecewise nice model. A nice latent variable model has a simple parametrization, but not a *linear* parametrization. One might worry that a low rank approximation could overlook this structure, and a more complicated approximation scheme would be required to compress such datasets. However, our main theorem suggests that low rank approximation is a remarkably powerful technique for approximating datasets from nice latent variable models. Hence this paper provides one explanation for the prevalence of low rank matrices in data science.

Our main result can be informally summarized as follows:

Nice latent variables models are of log-rank.

After formally defining "nice" latent variable models and log-rank in section 2, we state a precise version of this informal statement (see Theorem 4.2). Theorem 4.5 extends this result to piecewise nice latent variable models, while Theorem 4.7 considers symmetric latent variable models, i.e., graphons.

Our main tool is the Johnson–Lindenstrauss lemma (see Lemma 2.3), which says that given any point cloud in a high dimensional Euclidean space there exists an embedding onto a low dimensional Euclidean space that approximately preserves pairwise distances between points.

This result has ramifications for how to interpret an underlying low rank structure in datasets. In particular, we have good news for those designing algorithms: sufficiently large datasets tend to have low rank structure, which can be used to design faster algorithms.

Conversely, we have bad news for those who attempt to find meaning in low rank structure. Researchers often give post hoc explanations for why a particular dataset is approximately of low rank. For example, typical arguments are as follows: customers' movie preferences are low rank because movies are well parametrized by a few meaningful *genres*, or word document matrices are low rank because they are well parametrized by a handful of meaningful *topics*. Our main theorem shows that low rank structure can persist even without an underlying physical reason. In particular, a dataset from a nice latent variable model has an ϵ-rank (see Definition 2.1) that grows slowly with its dimensions, *no matter how many* genres or topics generate the data.

Throughout, we use $\|v\|^2 = \sum_{i=1}^N v_i^2$ to denote the Euclidean length of a vector $v \in \mathbb{R}^N$ and $\|f\| = \sup_{x\in\Omega} |f(x)|$ to denote the supremum norm of $f : \Omega \to \mathbb{R}$ over its domain Ω.

2. Background material.

We review some necessary background material and introduce the concept of ϵ-rank in the maximum absolute entry norm.

2.1. Rank.

A nonzero matrix $X \in \mathbb{R}^{m\times n}$ is said to be of rank 1 if X can be written as an outer-product of two column vectors, i.e., $X = uv^T$ for $u \in \mathbb{R}^{m\times 1}$ and $v \in \mathbb{R}^{n\times 1}$. Moreover, a matrix X is of rank k if k is the smallest integer so that X can be written as a sum of k rank 1 matrices. That is,

$$X = u_1 v_1^T + \cdots + u_k v_k^T, \qquad u_1, \ldots, u_k \in \mathbb{R}^{m\times 1}, \quad v_1, \ldots, v_k \in \mathbb{R}^{n\times 1}.$$

Generically, a matrix is of full rank; however, we find in data science that a full rank matrix can often be well approximated by a low rank matrix in the sense that $X \approx u_1 v_1^T + \cdots + u_k v_k^T$, where $k \ll \min(m, n)$. If one finds that a matrix X can be well approximated by a rank k matrix, X_k, then one can perform diagnostics directly on X_k instead of X.

2.2. The ϵ-rank of a matrix.

A matrix X can be approximated by a rank k matrix, up to an absolute accuracy of $\epsilon > 0$, if the ϵ-rank of X is less than or equal to k.

Definition 2.1 (ϵ-rank). *Let $X \in \mathbb{R}^{m\times n}$ be a matrix and $\epsilon > 0$ a tolerance. The (absolute) ϵ-rank of X is given by*

$$\operatorname{rank}_\epsilon(X) = \min\left\{\operatorname{rank}(A) : A \in \mathbb{R}^{m\times n}, \ \|X - A\|_{\max} \le \epsilon\right\},$$

where $\|\cdot\|_{\max}$ is the absolute maximum matrix entry. That is, $k = \operatorname{rank}_\epsilon(X)$ is the smallest integer for which X can be approximated by a rank k matrix, up to an accuracy of ϵ.

There are several alternative definitions of ϵ-rank in the literature [8].

2.3. A log-rank family of matrices.

We are interested in families of matrices $\mathcal{X} = \{X^{(m\times n)}\}_{m,n\ge 1}$, where the ϵ-rank of $X^{(m\times n)} \in \mathbb{R}^{m\times n}$ grows more slowly than a polylogarithm in m and n. We use the notation $X^{(m\times n)} \in \mathcal{X}$ to emphasize that $X^{(m\times n)}$ is a matrix of size $m \times n$.

Definition 2.2. *An infinite family of matrices $\mathcal{X} = \{X^{(m\times n)}\}_{m,n\ge 1}$ is of log-rank if there is a polynomial p such that for any fixed $\epsilon > 0$ we have*

$$\operatorname{rank}_\epsilon(X) = \mathcal{O}(p(\log(m + n))),$$

where X is any $m \times n$ matrix in the family $\mathcal{X}$.

In many settings (including the results in this paper), the polynomial p is simply $p(x) = x$.

In machine learning, $\mathcal{X}$ might represent a family of datasets. One can generate datasets of varying dimensions by sampling more examples (rows m) or features (columns n) from a data distribution: say, by collecting the required number of text documents, patient records, customer preferences, or movie reviews. A log-rank family of matrices contains datasets for which the ϵ-rank grows only slowly as we collect more examples and features. This structure could potentially be exploited for data science applications.

2.4. Latent variable models. Latent variable models (LVMs) are a particularly interesting class of families of matrices. An LVM is parametrized by a continuous function f and two sets $\mathcal{A}$ and $\mathcal{B}$. A family of matrices $\mathcal{X}_{f,\mathcal{A},\mathcal{B}} = \{X^{(m\times n)}\}_{m,n\geq 1}$ is an LVM (depending on f, $\mathcal{A}$, and $\mathcal{B}$) if for every $X^{(m\times n)} \in \mathcal{X}$,

$$(X^{(m\times n)})_{ij} = f(\alpha_i, \beta_j), \qquad 1 \leq i \leq m, \quad 1 \leq j \leq n,$$

where $\alpha_i \in \mathcal{A}$ and $\beta_j \in \mathcal{B}$.

LVMs have a natural relationship to low rank matrices. Let us consider two particular well-studied LVMs to understand how these models lead to low rank matrices.

- *Inner products.* Suppose $\mathcal{A} = \mathcal{B} = \mathbb{R}^r$ and $f(\alpha, \beta) = \alpha^T\beta$ is an inner product. Then, the rank of any matrix in the family $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ is at most r. Note that this bound is independent of the dimension of the matrix.
- *Smooth scalar functions.* Suppose $\mathcal{A} = \mathcal{B} = [-1, 1]$, and $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is an analytic function with bounded derivatives in α: there is a constant M such that $|f^{(k)}(\alpha_0, 0)| \leq M$ for every k and every $\alpha_0 \in [-1, 1]$. Then the ϵ-rank of any matrix in the family $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ is at most $\log(\frac{2M}{\epsilon})$. To see this, expand $f(\alpha, \beta)$ around $\beta = 0$ as

$$f(\alpha, \beta) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(\alpha, 0)\beta^k.$$

  For any matrix X in the family $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$, we can truncate this expansion after the Kth term to obtain a rank K approximation to X. To understand the quality of this approximation, consider the tail sum

$$\sum_{k=K}^{\infty} \frac{1}{k!} f^{(k)}(\alpha, 0)\beta^k \leq \sum_{k=K}^{\infty} \frac{M}{k!} \leq \frac{2M}{K!}.$$

  Using Stirling's approximation $K! \geq \sqrt{2\pi K}(\frac{K}{e})^K$ [38], we find

$$\sum_{k=K}^{\infty} \frac{1}{k!} f^{(k)}(\alpha, 0)\beta^k \leq \frac{2M}{K!} \leq 2M \left(\frac{e}{K}\right)^K \leq 2M \left(\frac{1}{2}\right)^K \leq \epsilon$$

  when $K \geq \max\{2e, \log(\frac{2M}{\epsilon})\}$. Hence we see that the ϵ-rank of any matrix in the family $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ is at most $\log(\frac{2M}{\epsilon})$. Note that this bound is again independent of the dimension of the matrix.
- *Smooth vector functions.* The previous argument used a Taylor expansion of the function in the parameter β. If $\mathcal{A}$ and $\mathcal{B}$ are both bounded sets in $\mathbb{R}^N$, and $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is

an analytic function with bounded derivatives, we can use the same argument to expand f in the vector β to again obtain a bound on the ϵ-rank independent of the dimension of the matrix. A similar idea is used in the literature on random feature maps to construct a low dimensional mapping for the Gaussian kernel using its Fourier expansion [37]. The bound again depends logarithmically on $\frac{1}{\epsilon}$; however, the bound grows exponentially in the dimension N of the latent variables. See Lemma 4.3 for the formal argument. Our main result, Theorem 4.2 eliminates the dependence on the dimension N of the latent variables by introducing a dependence on the dimension of the matrix.

LVMs can also be used to model more complex families of matrices. For example, f might be a kernel function, and $\mathcal{A}$ and $\mathcal{B}$ might be sets of extremely high dimensional vectors.

2.5. The Johnson–Lindenstrauss lemma. A key tool in theoretical computer science is the Johnson–Lindenstrauss lemma [24]. Roughly, it says that a high dimensional point cloud can be projected onto a low dimensional space while approximately preserving all pairwise distances between the points. There are several alternative forms and proofs [33].

Lemma 2.3 (the Johnson–Lindenstrauss lemma). *Let $0 < \epsilon_{JL} < 1$, $x_1, \ldots, x_n$ be n points in $\mathbb{R}^N$, and let $r = \lceil 8(\log n)/\epsilon_{JL}^2 \rceil$. Then, there is a linear map $Q : \mathbb{R}^N \to \mathbb{R}^r$ such that*

$$(1 - \epsilon_{JL})\|x_i - x_j\|^2 \le \|Q(x_i - x_j)\|^2 \le (1 + \epsilon_{JL})\|x_i - x_j\|^2, \qquad 1 \le i, j \le n.$$

Here, $\lceil a \rceil$ is the smallest integer larger than a.

Proof. See [33, Thm. 1.1]. Also, see [24]. ∎

A slight reformulation of the Johnson–Lindenstrauss lemma is useful for us; it roughly says that a high dimensional point cloud can be projected onto a low dimensional space while approximately preserving inner-products between vectors.

Lemma 2.4 (variant of the Johnson–Lindenstrauss lemma). *Let $0 < \epsilon_{JL} < 1$, $x_1, \ldots, x_n$ be n points in $\mathbb{R}^N$, and let $r = \lceil 8 \log(n+1)/\epsilon_{JL}^2 \rceil$. Then, there is a linear map $Q : \mathbb{R}^N \to \mathbb{R}^r$ such that*

$$\left| x_i^T x_j - x_i^T Q^T Q x_j \right| \le \epsilon_{JL} \left(\|x_i\|^2 + \|x_j\|^2 - x_i^T x_j \right), \qquad 1 \le i, j \le n.$$

Proof. Consider the point set $\{x_1, \ldots, x_n, 0\} \subset \mathbb{R}^N$. Since $r = \lceil 8(\log(n+1))/\epsilon_{\mathrm{JL}}^2 \rceil$, the Johnson–Lindenstrauss lemma says that there exists a linear map $Q : \mathbb{R}^N \to \mathbb{R}^r$ such that

$$\begin{aligned} (1 - \epsilon_{\mathrm{JL}})\|x_i\|^2 \le &\|Qx_i\|^2 \le (1 + \epsilon_{\mathrm{JL}})\|x_i\|^2, && 1 \le i \le n, \\ (1 - \epsilon_{\mathrm{JL}})\|x_i - x_j\|^2 \le &\|Q(x_i - x_j)\|^2 \le (1 + \epsilon_{\mathrm{JL}})\|x_i - x_j\|^2, && 1 \le i, j \le n. \end{aligned}$$

Therefore, from the identity $2a^T b = \|a\|^2 + \|b\|^2 - \|b - a\|^2$ we find that

$$\begin{aligned} (1 - \epsilon_{\mathrm{JL}})(\|x_i\|^2 + \|x_j\|^2) - (1 + \epsilon_{\mathrm{JL}})\|x_i - x_j\|^2 \le 2x_i^T Q^T Q x_j \\ \le (1 + \epsilon_{\mathrm{JL}})(\|x_i\|^2 + \|x_j\|^2) - (1 - \epsilon_{\mathrm{JL}})\|x_i - x_j\|^2. \end{aligned}$$

Using the identity $2a^T b = \|a\|^2 + \|b\|^2 - \|b - a\|^2$ again, we obtain

$$-\epsilon_{\mathrm{JL}}(\|x_i\|^2 + \|x_j\|^2 - x_i^T x_j) \le x_i^T x_j - x_i^T Q^T Q x_j \le \epsilon_{\mathrm{JL}}(\|x_i\|^2 + \|x_j\|^2 - x_i^T x_j),$$

as required. ∎

2.6. Extremely large matrices are low rank in the max norm. The variant of the Johnson–Lindenstrauss lemma in Lemma 2.4 allows us to prove Theorem 1.0.

Proof of Theorem 1.0. The singular value decomposition of X is $X = U\Sigma V^T$. We can write $X = \tilde{U}\tilde{V}^T$, where $\tilde{U} = U\sqrt{\Sigma}$ and $\tilde{V} = V\sqrt{\Sigma}$. Applying Lemma 2.4 with $\epsilon_{\text{JL}} = \epsilon/3$ to the set $\{\tilde{u}_1, \ldots, \tilde{u}_n, \tilde{v}_1, \ldots \tilde{v}_n, 0\}$ with $\tilde{u}_j$ and $\tilde{v}_j$ being the jth column of $\tilde{U}$ and $\tilde{V}$, respectively, we find that for $r = \lceil 72 \log(2n+1)/\epsilon^2 \rceil$ there exists a $Q \in \mathbb{R}^{n \times r}$ such that

$$\left|\tilde{u}_i^T \tilde{v}_j - \tilde{u}_i^T Q^T Q \tilde{v}_j\right| \leq \epsilon_{\text{JL}} \left(\|\tilde{u}_i\|^2 + \|\tilde{v}_j\|^2 - \tilde{u}_i^T \tilde{v}_j\right).$$

Since $X_{ij} = \tilde{u}_i^T \tilde{v}_j$, $\|\tilde{u}_i\|^2 = \sigma_i(X) \leq \|X\|_2$, and $\|\tilde{v}_i\|^2 = \sigma_i(X) \leq \|X\|_2$, we find that

$$\begin{aligned}\left|X_{jk} - \tilde{u}_i^T Q^T Q \tilde{v}_j\right| &\leq \epsilon_{\text{JL}} \left(2\|X\|_2 + \|X\|_{\max}\right) \\ &\leq 3\epsilon_{\text{JL}} \|X\|_2,\end{aligned}$$

where the last inequality uses the fact that $\|X\|_{\max} \leq \|X\|_2$. The result follows by setting $Y_{ij} = \tilde{u}_i^T Q^T Q \tilde{v}_j$ and noting that $\epsilon = 3\epsilon_{\text{JL}}$. ∎

3. Related work. The majority of the literature focuses on either how to find low rank matrices or how to exploit low rank structure after it has been found. This trend is set to continue with the emerging field of multilinear algebra and the increasing use of tensor factorizations in machine learning and data analysis [21, 22, 35, 42]. This keen practical interest in low rank structure lends urgency to the quest to understand why and when low rank techniques work well on real datasets.

3.1. Bounds on ϵ-rank. The work of Alon and his coauthors is closest in spirit to our paper [3, 4]. These papers use the Johnson–Lindenstrauss lemma to show that the identity matrix, and any positive semidefinite matrix, has an ϵ-rank that grows logarithmically with the number of columns and rows. An earlier result in the same spirit by Barvinok shows that for any positive semidefinite matrix A, one can find a low rank matrix B that matches the inner-product between A and each of k other positive semidefinite matrices to within ϵ [6]. The rank of B is at most $\lceil 8/\epsilon^2 \log(4k) \rceil$.

One can obtain a result similar to our Theorem 1.0 using a technique similar to the proof of the approximate Carathéodory theorem in [46, Theorem 0.1.2]. The analogous result for matrices is as follows: for any matrix $X \in \mathbb{R}^{m \times n}$ there is a matrix Y of rank $\leq \lceil 1/\epsilon^2 \rceil$ satisfying

$$\|X - Y\|_F \leq \epsilon \|X\|_*,$$

where $\|X\|_*$ is the sum of the singular values of X. Since $\|X\|_* \leq n\|X\|$ and $\|X - Y\|_{\max} \leq \|X - Y\|_F$, we can use this to show

$$\|X - Y\|_{\max} \leq n\epsilon \|X\|.$$

Chatterjee shows that any matrix with bounded entries can be well approximated by thresholding all singular values lower than a given value to 0 [13]. His main theorem implies that the ϵ-rank of a matrix of size $n \times n$ grows like $\mathcal{O}(\sqrt{n})$. Our theorem improves this result to $\mathcal{O}(\log n)$ when the matrix comes from a nice LVM.

In [8, 43], bounds were derived on a slightly different ϵ-rank of certain matrices $X \in \mathbb{R}^{m\times n}$ with displacement structure, i.e., a matrix that satisfies $AX - XB = F$. For example, [8, Thm. 3.1] showed that all $n \times n$ positive-definite Hankel matrices, $(H_n)_{ij} = h_{i+j}$, have an ϵ-rank that grows logarithmically in n. These results from linear algebra are considering matrices that have more rapidly decaying singular values than the LVMs we study in this paper.

3.2. Exchangeable families of matrices. LVMs are related to so-called *exchangeable* families of matrices. We say that an infinite matrix $\mathcal{X}$ is exchangeable if for any permutations σ and π on $\mathbb{N}$ and any integers m and n, we have

$$\mathcal{X}_{i,j} \sim \mathcal{X}_{\sigma(i),\pi(j)}, \qquad 1 \le i \le m, \quad 1 \le j \le n,$$

where $\sim$ denotes equality in distribution. A celebrated result by Aldous [2] states that if $\mathcal{X}$ is exchangeable, then

$$\mathcal{X}_{ij} \sim f(\omega, \alpha_i, \beta_j, \eta_{ij}),$$

where f is a measurable function; ω, α_i, β_j, η_{ij} are scalar-valued; and the ω, α_i's, β_j's, and η_{ij}'s are mutually independent and uniformly distributed random variables on $[0, 1]$. One can generate a family of matrices from $\mathcal{X}$ by taking the leading $m \times n$ principal submatrices.

There are two significant differences here from the LVM: (1) There is an intrinsic noise term η_{ij}, and (2) the latent variables ω, α_i, and β_j are scalar-valued and uniform random variables on $[0, 1]$. Our result on LVMs can be extended to exchangeable families of matrices under additional smoothness assumptions on f.

The symmetric analogue of an exchangeable array is a graphon. Graphons can be seen as the continuous limit of a sequence of (dense) graphs [32]. Many authors have proposed methods for graphon estimation from samples of the entries [14, 1, 49, 12]. For example, Airoldi, Costa, and Chan required that the graphon be piecewise Lipshitz and provided an approximate graphon that gives a complexity that grows linearly in the number of pieces [1]. Our theory shows that this procedure overestimates the complexity required to model a graphon when the graphon is nice. Indeed, Theorem 4.5 shows that the ϵ-rank of a nice graphon grows with the maximum complexity of each piece (see section 4.2). For reasonable models, the maximum complexity grows sublinearly in the number of pieces. Choi, Wolfe, and Airoldi showed that it is possible to find a consistent estimator for the graphon when the number of classes in a stochastic block model grows at most like the square root of the dimension [14]. Our theory shows that a low rank model for the graphon (which generalizes a stochastic block model) only requires a rank that grows like the logarithm of the dimension. Whether it is possible to find statistically consistent estimators that obtain this threshold is an important question for future research.

The theory of exchangeable matrices has been used to motivate the use of LVMs for collaborative filtering and other applications in machine learning. For example, many authors have used the assumption that the LVM is Lipschitz to design efficient estimators [41, 30, 29]. We show a connection between this approach and the standard low rank model.

4. Any nice latent variable model is log-rank.

Our result applies to any *nice* LVM, which we now define.

Definition 4.1. *A latent variable model* $\mathcal{X} = \mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ *is called* nice *with parameters* (N, R, C, M) *if the following conditions hold:*

- *The associated sets* $\mathcal{A}$ *and* $\mathcal{B}$ *are supported on a closed ball* $B_R \subset \mathbb{R}^N$ *for some* $N \geq 1$ *of radius* $R > 0$*, i.e.,* $B_R = \{x \in \mathbb{R}^N : \|x\| \leq R\}$*. Here,* N *is allowed to be extremely large.*
- *The associated function* $f : B_R \times B_R \to \mathbb{R}$ *is bounded and sufficiently smooth in the sense that* $f(\alpha, \cdot)$ *is uniformly analytic in* B_R *for every* $\alpha \in B_R$ *and for all* $\mu \in \mathbb{N}^N$ *we have*

$$\|D^\mu f(\alpha, \beta)\| \leq C M^{|\mu|} \|f\|.$$

Here, $\mu = (\mu_1, \ldots, \mu_N)$ *is a multi-index,* $|\mu| = \sum_{i=1}^N \mu_i$*,* $D^\mu f = \frac{\partial^{|\mu|}}{\partial^{\mu_1}\beta_1 \cdots \partial^{\mu_N}\beta_N}$*, and* $C \geq 0$ *and* $M \geq 0$ *are nonnegative constants.*

This parametrization is not unique as one can always rescale the latent parameters so that they are drawn from a smaller ball by making f steeper. Except for possibly constants, our bounds only depend on the product MR, which is preserved by any such reparametrization.

Nice LVMs are common in machine learning and data analysis. Functions that give rise to nice LVMs include the following:

- *Linear functions.* If $f(\alpha, \beta) = \alpha^T \beta$ and the sets $\mathcal{A} = \mathcal{B} = B_R \subset \mathbb{R}^N$ for $R \geq 1$, then $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ is a nice LVM with $M = C = 1$. In this case, $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ has a rank bounded by N. Theorem 4.2 shows that when N is sufficiently large the ϵ-rank is actually smaller than N for $\epsilon > 0$.
- *Polynomials.* If f is a polynomial in $2N$-variables, then there is a constant M that depends on N, R, and the degree of the polynomial so that $\|D^\mu f(\alpha, \beta)\| \leq C M^{|\mu|} \|f\|$. For simplicity, consider $N = 1$ and $f(\alpha, \beta) = \beta^d$. Then, for $k < d$ we have

$$\|D^k f(\alpha, \beta)\| = d(d-1)\cdots(d-k+1) \sup_{|\beta| \leq R} |\beta|^{d-k} \leq d^k R^{-k} \|f\|.$$

  So, $M = d/R$ and $C = 1$ suffices.
- *Kernels.* If $f(\alpha, \beta) = e^{p(\alpha,\beta)}$ for a $2N$-variable polynomial p, then $\|D^\mu f(\alpha, \beta)\| \leq C M^{|\mu|} \|f\|$ for some constants C and M that depend on p. (This follows from a simple inductive argument using the chain rule.) This functional form includes most kernels typically used in machine learning. For example, consider the radial basis function kernel $f(\alpha, \beta) = \exp(-\|\alpha - \beta\|^2)$ with $R > 1/2$. Then, $\|D^\mu f(\alpha, \beta)\| \leq N(4R)^{N+|\mu|} \|f\|$.

We see that the bound on the derivatives of f allows for many relevant examples.

Many probabilistic models give rise to piecewise nice LVMs. For example, suppose that the sets α and β are generated by taking m and n draws, respectively, from a sub-Gaussian distribution on $\mathbb{R}^N$. For each α and β, $\Pr[\|\alpha\|^2 > 5N] \leq \exp(-N)$ [23]. So we see that with probability $\geq 1/2$, α and β are both supported on a ball of radius $5N$ so long as $m + n < 1/2 \exp(N)$.

Our framework can also handle the case of piecewise nice LVMs, which we treat below in Theorem 4.5.

We are now ready to formally state our main result. An alternative theorem with the analytic assumptions of f on the first variable is also possible with an analogous proof.

Theorem 4.2. *Let* $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ *be a nice LVM and* $0 < \epsilon < 1$*. Then, for each* $X^{(m \times n)} \in \mathcal{X}_{f,\mathcal{A},\mathcal{B}}$*,*

the $\epsilon\|f\|$-rank of $X^{(m\times n)}$ is no more than

$$r = \left\lceil 8\log(m+n+1)\left(1+\frac{2(C_u+C_v+1)}{\epsilon}\right)^2\right\rceil,$$

where C_u and C_v are constants defined below that depend on the LVM $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$.

We state Theorem 4.2 in terms of the $\epsilon\|f\|$-rank to show that we achieve a natural sort of relative-error guarantee. Consider the LVM $\mathcal{X}'_{f',\mathcal{A},\mathcal{B}}$, where $f' = cf$ for some positive constant c. The entries of a matrix drawn from $\mathcal{X}'_{f',\mathcal{A},\mathcal{B}}$ are about a factor of c larger in expectation than the entries of a matrix drawn from $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$. It is natural to compare the c-rank of a matrix from $\mathcal{X}'_{f',\mathcal{A},\mathcal{B}}$ with the 1-rank of a matrix from $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$. Theorem 4.2 shows both satisfy the same bound, since $\|f'\| = c\|f\|$.

The proof proceeds in two main steps. The first is to find an explicit (possibly high) rank factorization of some approximation $\hat{X}$ to a matrix $X^{(m\times n)} \in \mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ drawn from the LVM. We use a Taylor expansion of the function $f(\alpha,\cdot)$ about 0 to show that $f(\alpha_i,\beta_j) \approx u_i^T v_j$. That is, f can be well approximated as the inner product between two (high-dimensional) vectors, u_i and v_j, with bounded Euclidean norms. The second step is to use the Johnson–Lindenstrauss lemma to reduce the dimensionality of the set of vectors $\{0, u_1, \ldots, u_m, v_1, \ldots, v_n\}$ while approximately preserving the inner products $u_i^T v_j$.

We present the first step as a lemma.

Lemma 4.3 (bounded rank approximation). *Let $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ be a nice LVM with parameters (N,R,C,M), and let $0<\epsilon<1$. Then, for each $X^{(m\times n)} \in \mathcal{X}_{f,\mathcal{A},\mathcal{B}}$, there is a matrix $\hat{X}\in\mathbb{R}^{m\times n}$ such that $\|X-\hat{X}\|_{max} \le \epsilon\|f\|$ and*

$$rank(\hat{X}) \le (K+1)N^K, \qquad K \le \max(2eNRM, \log_2(C/\epsilon)) + 1.$$

Furthermore, $\hat{X}$ admits a rank $\tilde{N} \le (K+1)N^K$ factorization as

$$\hat{X}_{ij} = u_i^T v_j \quad 1\le i\le m, \quad 1\le j\le n,$$

where each $u_i \in \mathbb{R}^{\tilde{N}}$ and $v_j \in \mathbb{R}^{\tilde{N}}$ obey

$$\|u_i\| \le C_u\|f\|, \qquad \|v_j\| \le C_v\|f\|.$$

Here, C_u and C_v are constants depending on the LVM $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ but not on the dimensions m or n.

Notice that the vectors u_i and v_j may have an extremely large number of entries when the dimension N of the LVM is large: this bound on the rank of $\hat{X}$ grows as N^N.

Proof of Lemma 4.3. We begin by showing that $f(\alpha_i,\beta_j) \approx u_i^T v_j$. By Taylor expanding $f(\alpha_i,\beta_j)$ in the second variable about 0 with K terms, we find that

$$\left|X_{ij}-\hat{X}_{ij}\right| \le \frac{N^{K+1}R^{K+1}}{(K+1)!}\max_{|\tau|=K+1}\sup_{z\in B_R}\left|D^\tau f(\alpha_i,z)\right|, \quad \hat{X}_{ij} = \sum_{|\mu|\le K}\frac{D^\mu f(\alpha_i,0)}{\mu!}\beta_j^\mu,$$

where $D^\mu f = \frac{\partial^{|\mu|}}{\partial^{\mu_1}\beta_1 \cdots \partial^{\mu_N}\beta_N}$, $\mu! = \mu_1! \cdots \mu_N!$, and $\beta_j^\mu = (\beta_j)_1^{\mu_1} \cdots (\beta_j)_N^{\mu_N}$. Here, the N^{K+1} term in the Taylor error comes from the fact that there are fewer than N^{K+1} μ's with $|\mu| = K+1$: to get a term with $|\mu| = K+1$, we must choose $K+1$ elements from the N coordinates (with replacement).

From the formula for $\hat{X}_{ij}$, there are vectors u_i and v_j with $\tilde{N} := \sum_{|\mu| \le K} 1$ entries, such that $\hat{X}_{ij} = u_i^T v_j$. From the simple counting argument above, we can find that

$$\tilde{N} = \sum_{|\mu| \le K} 1 = \sum_{k=0}^{K} \sum_{|\mu|=K} 1 \le \sum_{k=0}^{K} N^k \le (K+1)N^K.$$

The vectors u_i and v_j are indexed by $|\mu| \le K$ and can be taken to be

$$(u_i)_\mu = \frac{1}{\sqrt{\mu!}\sqrt{\|f\|}} D^\mu f(\alpha_i, 0), \qquad (v_j)_\mu = \frac{1}{\sqrt{\mu!}} \sqrt{\|f\|} \beta_j^\mu.$$

Hence, we write

$$\hat{X} = UV, \qquad U = \left[u_1 | \cdots | u_m\right]^T, \quad V = \left[v_1 | \cdots | v_n\right].$$

This result immediately gives a bound on the rank of $\hat{X}$. For example, if $N = 1$, we have $\mathrm{Rank}(\hat{X}) \le \tilde{N} \le (K+1)N^K = K+1$.

Now, we select K to be sufficiently large so that

$$\begin{aligned} \left|X_{ij} - \hat{X}_{ij}\right| &\le \frac{N^{K+1}R^{K+1}}{(K+1)!} \max_{|\tau|=K+1} \sup_{z \in B_R} |D^\tau f(\alpha_i, z)| \\ &\le C \frac{N^{K+1}R^{K+1}M^{K+1}}{(K+1)!} \|f\| \\ &\le \epsilon \|f\|. \end{aligned}$$

Since the denominator grows superexponentially in K, there is always a sufficiently large K for the bound above for any $0 < \epsilon < 1$.

To find a concrete bound on K, we use Stirling's formula: $K! \ge \sqrt{2\pi K}(\frac{K}{e})^K$ [38]. Pick $K \ge 2eNRM$, so $\frac{eNRM}{K+1} \le \frac{1}{2}$. Substituting Stirling's formula into the previous display, we find that

$$\begin{aligned} \left|X_{ij} - \hat{X}_{ij}\right| &\le \frac{1}{\sqrt{2\pi(K+1)}} \left(\frac{NRMe}{K+1}\right)^{K+1} C\|f\| \\ &\le \left(\frac{NRMe}{K+1}\right)^{K+1} C\|f\| \\ &\le \left(\frac{1}{2}\right)^{K+1} C\|f\| \\ &\le \epsilon \|f\| \end{aligned}$$

when $K \geq \log_2(C/\epsilon)$. Hence $K \geq \max(2eNRM, \log_2(C/\epsilon))$ suffices to achieve an $\epsilon\|f\|$-approximation to X.

Therefore, we have the approximation

$$|X_{ij} - \hat{X}_{ij}| \leq \epsilon\|f\|, \quad \hat{X}_{ij} = u_i^T v_j, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n,$$

where $u_i \in \mathbb{R}^{\tilde{N}}$ and $v_j \in \mathbb{R}^{\tilde{N}}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Let us remark on the norms of u_i and of v_j. We suppress the indices i and j in this discussion.

Let $u^{(\infty)} = (u_\mu)_{|\mu|\geq 0}$ and $v^{(\infty)} = (v_\mu)_{|\mu|\geq 0}$ be infinite dimensional vectors. Then,

$$\|u\|^2 \leq \|u^{(\infty)}\|^2 = C_u\|f\| < \infty, \quad \|v\|^2 \leq \|v^{(\infty)}\|^2 = C_v\|f\| < \infty,$$

where C_u and C_v are constants that depend only on the properties of the nice LVM.

For C_v we have

$$\|v^{(\infty)}\|^2 \leq \sum_\mu \frac{1}{\mu!}\left|\beta^{2\mu}\right|\|f\| \leq \sum_{s=0}^{\infty} \frac{1}{s!}(N+s)^N R^{2s}\|f\| \leq C_v\|f\|,$$

showing that C_v is finite.

The constant C_u depends on how quickly the derivatives of f grow; it is bounded as long as they grow no faster than exponentially. Since $\mu! \geq (\lfloor|\mu|/N\rfloor)!$, we have

$$|u_\mu|^2 = \frac{1}{\mu!\|f\|}|D^\mu f(\alpha, 0)|^2 \leq C^2 M^{2|\mu|}\|f\|\frac{1}{(\lfloor|\mu|/N\rfloor)!}.$$

Hence, we see that

$$\|u^{(\infty)}\|^2 \leq \sum_{s=0}^{\infty}(N+s)^N \frac{C^2M^{2s}}{(\lfloor s/N\rfloor)!}\|f\| \leq C_u\|f\|,$$

showing that C_u is finite. ∎

We are now ready to prove our main theorem.

Proof of Theorem 4.2. Suppose $X \in \mathcal{X}_{f,\mathcal{A},\mathcal{B}} \cap \mathbb{R}^{m\times n}$ has entries $X_{ij} = f(\alpha_i, \beta_j)$ for each $1 \leq i \leq m$ and $1 \leq j \leq n$.

The proof proceeds in two steps. First, we use Lemma 4.3 to show that for each $1 \leq i \leq m$ and $1 \leq j \leq n$, $|f(\alpha_i, \beta_j) - u_i^T v_j| \leq \epsilon/2$ for two (extremely high dimensional) vectors, $u_i \in \mathbb{R}^{\tilde{N}}$ and $v_j \in \mathbb{R}^{\tilde{N}}$, with Euclidean norms bounded by $C_u\|f\|$ and $C_v\|f\|$, respectively. Second, we use the Johnson–Lindenstrauss lemma to show that $u_i^T v_j \approx (Qu_i)^T Qv_j$ for $Q \in \mathbb{R}^{r\times\tilde{N}}$.

Let $r = \lceil 8(\log(m+n+1)/\epsilon_{\text{JL}}^2\rceil$. Then, by Lemma 2.4 we know that there exists a linear map $Q \in \mathbb{R}^{r\times\tilde{N}}$ such that

$$|u_i^T v_j - u_i^T Q^T Q v_j| \leq \epsilon_{\text{JL}}(\|u_i\|^2 + \|v_j\|^2 - u_i^T v_j), \qquad 1 \leq i \leq m, \quad 1 \leq j \leq n.$$

Now, using our bound on $\|u_i\|^2$ and $\|v_j\|^2$ from above, we obtain the following inequalities for every $u \in \{u_1, \ldots, u_m\}$ and $v \in \{v_1, \ldots, v_n\}$:

$$\begin{aligned} |u^T v - u^T Q^T Q v| &\le \epsilon_{\mathrm{JL}} \left(\|u^{(\infty)}\|^2 + \|v^{(\infty)}\|^2 + |f(\alpha,\beta)| + \tfrac{\epsilon}{2}\|f\| \right) \\ &\le \epsilon_{\mathrm{JL}}((C_u + C_v)\|f\| + (1 + \tfrac{\epsilon}{2})\|f\|), \end{aligned}$$

where we have used the fact that $|f(\alpha,\beta)| \le \|f\|$ and $|u^T v - f(\alpha,\beta)| \le \epsilon/2\|f\|$.

The total error in each entry of our approximation is thus

$$\begin{aligned} |f(\alpha_i,\beta_j) - x_i^T y_j| &\le |f(\alpha_i,\beta_j) - u_i^T v_j| + |u_i^T v_j - x_i^T y_j| \\ &\le \epsilon/2\|f\| + \epsilon_{\mathrm{JL}}(C_u + C_v + 1 + \epsilon/2)\|f\|. \end{aligned}$$

Thus, if we select ϵ_{JL} to be

$$\epsilon_{\mathrm{JL}} = \frac{\epsilon/2}{C_u + C_v + 1 + \epsilon/2},$$

then we have $|f(\alpha_i,\beta_j) - x_i^T y_j| \le \epsilon\|f\|$, as desired.

Therefore, the $\epsilon\|f\|$-rank of X is at most the rank of the matrix $\tilde{X}_{ij} = x_i^T y_j$, which is of rank at most r. Here, r is the integer given by

$$r = \left\lceil 8\log(m+n+1)\left(1 + \frac{2(C_u + C_v + 1)}{\epsilon}\right)^2 \right\rceil.$$

Theorem 4.2 provides a nontrivial bound when

$$\min(m,n) > \left\lceil 8\log(m+n+1)\left(1 + \frac{2(C_u + C_v + 1)}{\epsilon}\right)^2 \right\rceil,$$

since the rank of a matrix is always bounded by its smallest dimension. Hence, we see Theorem 4.2 is only interesting for sufficiently large matrices.

4.1. Piecewise nice latent variable models. The requirement that the function f associated to the LVM be analytic can be relaxed to piecewise analytic. We call such models piecewise nice LVMs.

Definition 4.4. *The family of matrices $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ is call a piecewise nice LVM if there exists a finite partition of the sets*

$$\mathcal{A} \times \mathcal{B} = \cup_{\ell=1}^{P} (\mathcal{A}_\ell \times \mathcal{B}_\ell), \quad (\mathcal{A}_\ell \times \mathcal{B}_\ell) \cap (\mathcal{A}_{\ell'} \times \mathcal{B}_{\ell'}) = \emptyset, \quad \ell \ne \ell',$$

so that

$$f(\alpha,\beta) = f_\ell(\alpha,\beta), \qquad (\alpha,\beta) \in \mathcal{A}_\ell \times \mathcal{B}_\ell,$$

with $\mathcal{X}_{f_\ell,\mathcal{A}_\ell,\mathcal{B}_\ell}$ being nice LVMs for $1 \le \ell \le P$.

We find that any piecewise nice LVM is also of log-rank.

Theorem 4.5. *Let $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$ be a piecewise nice LVM. Then, for each $0 < \epsilon < 1$ and for any $X^{(m\times n)} \in \mathcal{X}_{f,\mathcal{A},\mathcal{B}}$, the $\epsilon\|f\|$-rank of $X^{(m\times n)}$ is no more than*

$$r = \left\lceil 8\log(m+n+1)\left(1+\frac{2(C_u+C_v+1)}{\epsilon}\right)^2\right\rceil,$$

where C_u and C_v are constants that depend on properties of the LVM $\mathcal{X}_{f,\mathcal{A},\mathcal{B}}$.

The proof of this theorem is an easy modification of the proof of Theorem 4.2 because the dimension of the projected vectors in the Johnson–Lindenstrauss lemma is independent of the dimension of the original vectors. For example, we can take

$$u_i = (0,\dots,0,\overbrace{u_i^{(\ell)}}^{\alpha_i\in\mathcal{A}_\ell},0,\dots,0), \qquad v_j = (0,\dots,0,\overbrace{v_j^{(\ell)}}^{\beta_j\in\mathcal{B}_\ell},0,\dots,0),$$

where $1 \le i \le m$ and $1 \le j \le n$. Note that it is possible that $\alpha_i \in \mathcal{A}_\ell$ (resp., $\alpha_i \in \mathcal{B}_\ell$) for multiple ℓs, so u_i (resp., v_i) may have more than one nonzero block. We can also take

$$\hat{X}_{ij} = u_i^T v_j = \sum_{l:(\alpha_i,\beta_j)\in\mathcal{A}_\ell\times\mathcal{B}_\ell} \left(u_i^{(\ell)}\right)^T v_j^{(\ell)} = \left(u_i^{(\ell_{ij})}\right)^T v_j^{(\ell_{ij})},$$

where ℓ_{ij} is the unique ℓ so that $(\alpha_i,\beta_j) \in \mathcal{A}_\ell \times \mathcal{B}_\ell$. (It is unique because $\{\mathcal{A}_\ell \times \mathcal{B}_\ell\}_{\ell=1}^P$ partitions $\mathcal{A}\times\mathcal{B}$.) Finally, the norms of u_i and v_j are just the sum of the norms of $u_i^{(\ell)}$ and $v_j^{(\ell)}$ so the constants C_u and C_v in the proof are replaced by $\max_\alpha \sum_{\ell:\ \alpha\in\mathcal{A}_\ell} C_u^{(\ell)}$ and $\max_\beta \sum_{\ell:\ \beta\in\mathcal{B}_\ell} C_v^{(\ell)}$.

4.2. Symmetric latent variable models. Above, we noticed a connection between LVMs and exchangeable families of matrices. To understand the rank of *symmetric* exchangeable families of matrices (e.g., graphons), and the rank of *symmetric* matrices, we define a symmetric notion of LVMs.

Definition 4.6. *A family of matrices $\mathcal{X}_{f,\mathcal{A}}$ is a symmetric latent variable model (depending on f and $\mathcal{A}$) if for every $X^{(n\times n)} \in \mathcal{X}_{f,\mathcal{A}}$,*

$$(X^{(n\times n)})_{ij} = f(\alpha_i,\alpha_j), \qquad 1 \le i,j \le n.$$

If $\mathcal{A}$ is compact and $|D^\mu f(\alpha,\alpha')| \le CM^{|\mu|}\|f\|$, we say the symmetric LVM is nice. If $\mathcal{A} = [0,1]$ and $f : [0,1]\times[0,1] \to [0,1]$, then $\mathcal{X}_{f,\mathcal{A}}$ is a graphon [32]. Graphons are often used to model processes that generate random graphs, by interpreting the entries of $X^{(n\times n)} \in \mathcal{X}_{f,\mathcal{A}}$ as the probability that a graph on n nodes has an edge between node i and node j.

We show any symmetric LVM is of log-rank.

Theorem 4.7. *Let $\mathcal{X}_{f,\mathcal{A}}$ be a nice symmetric LVM, and let $0 < \epsilon < 1$. Then, for $X^{(n\times n)} \in \mathcal{X}_{f,\mathcal{A}}$, the $\epsilon\|f\|$-rank of $X^{(n\times n)}$ is no more than*

$$r = \left\lceil 8\log(2n+1)\left(1+\frac{2(C_u+C_v+1)}{\epsilon}\right)^2\right\rceil,$$

where C_u and C_v are constants which depend on the LVM $\mathcal{X}_{f,\mathcal{A}}$.

The proof of this theorem is nearly identical to the proof of Theorem 4.2, since we never use independence of α_i and β_j.

5. Numerical experiments. Our theory shows that a matrix generated from a nice LVM is often well approximated by a matrix of low rank, even if the true latent structure is high dimensional or nonlinear. However, computing the ϵ-rank for $0 < \epsilon < 1$ is provably NP-hard [20], where

$$\text{rank}_\epsilon(X) = \min \left\{ \text{rank}(A) : A \in \mathbb{R}^{m \times n}, \ \ \|X - A\|_{\max} \leq \epsilon \right\}.$$

This makes numerical experiments difficult as our theory is only meaningful for large matrices.

A simple approach to crudely computing $\text{rank}_\epsilon(X)$ is to approximate X by its truncated singular value decomposition, using whatever truncation is necessary so that $\|X - A\|_{\max} \leq \epsilon$. More formally, define $[X]_r = \text{argmin}_{\text{rank}(Y) \leq r} \|X - Y\|_2$, and define $\mu_r(X)$ as

$$\mu_r(X) = \|X - [X]_r\|_{\max}.$$

An upper bound on $\text{rank}_\epsilon(X)$ can be found by selecting the small integer r so that $\mu_r(X) \leq \epsilon$.

This paper provides three different bounds on the ϵ-rank for a matrix $X \in \mathbb{R}^{n \times n}$ drawn from a nice LVM with latent factors of dimension N. Lemma 4.3 shows that $\text{rank}_\epsilon(X) = \mathcal{O}(N^N \log(1/\epsilon))$. Our main result, Theorem 4.2, shows that $\text{rank}_\epsilon(X) = \mathcal{O}(\log n/\epsilon^2)$. And, of course, we have the trivial bound of $\text{rank}_\epsilon(X) \leq n$. Based on these bounds, we should expect that when N is large, then for sufficiently large n, $\text{rank}_\epsilon(X)$ grows like $\log n$. On the other hand, for small n or ϵ, we can have $\log n/\epsilon^2 \gtrsim n$, and hence we may see that $\text{rank}_\epsilon(X)$ grows linearly with n.

Figure 5.1 shows both of these behaviors. We realize a matrix by drawing from a nice LVM with $N = 1000$: each latent variable is generated as a random point on the N-dimensional unit sphere, and we use the function $f(\alpha, \beta) = \exp(-\|\alpha - \beta\|^2)$ to generate matrix entries. We plot our crude upper bound on $\text{rank}_\epsilon(X)$ using the values of $\mu_r(X)$ by generating matrices for a range of tolerances ϵ and dimensions n. For each value of ϵ and n, we randomly draw five matrices and plot the maximum obtained upper bound. We can see that for small n or ϵ, our upper bound on $\text{rank}_\epsilon(X)$ grows linearly in the dimension n. On the other hand, we can see that for large n and ϵ, the growth of $\text{rank}_\epsilon(X)$ is approximately logarithmic in n.

Conclusion. This paper seeks to partially answer the question "Why are low rank techniques so effective for solving problems in data analysis and machine learning?" Theorem 4.2 provides a partial explanation for its effectiveness: when rows and columns of the data are drawn in a consistent way from a nice set, the rank of the resulting matrix cannot increase very quickly. Formally, we have shown that nice latent variable models give rise to matrices that have an ϵ-rank that grows only logarithmically with the matrix dimensions, with respect to the maximum absolute entry norm. This suggests that low rank structure in large datasets is a universal feature and provides a broad motivation for low rank techniques in data science and machine learning.

Acknowledgments. We would like to thank Siddhartha Banerjee, Lijun Ding, Gilbert Strang, and Joel Tropp for useful discussions.

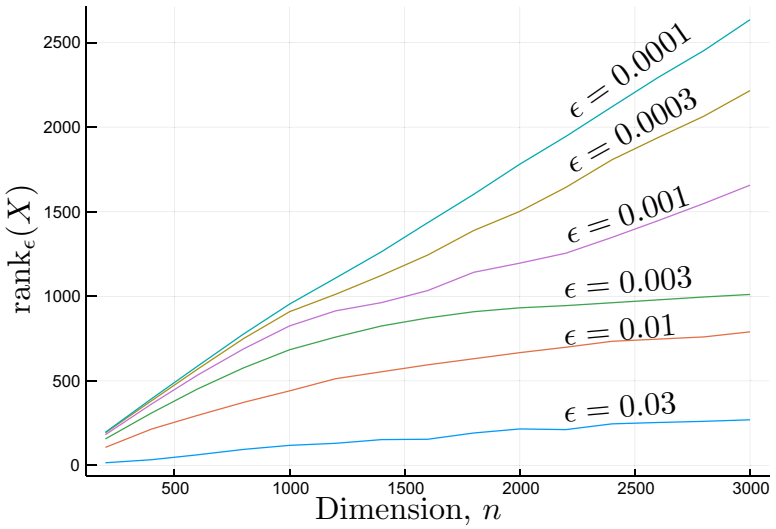

Figure 5.1. *An upper bound on* $\text{rank}_\epsilon(X)$ *for* $0.0001 \le \epsilon \le 0.03$ *and* $300 \le n \le 3000$.

REFERENCES

[1] E. M. Airoldi, T. B. Costa, and S. H. Chan, *Stochastic blockmodel approximation of a graphon: Theory and consistent estimation*, in Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2013, pp. 692–700.

[2] D. J. Aldous, *Representations for partially exchangeable arrays of random variables*, J. Multivariate Anal., 11 (1981), pp. 581–598.

[3] N. Alon, *Perturbed identity matrices have high rank: Proof and applications*, Combin. Probab. Comput., 18 (2009), pp. 3–15.

[4] N. Alon, T. Lee, A. Shraibman, and S. Vempala, *The approximate rank of a matrix and its algorithmic applications*, in Proceedings of the 45th Annual ACM Symposium on Theory of Computing, ACM, New York, 2013, pp. 675–684.

[5] S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi, *Matrix Completion Methods for Causal Panel Data Models*, Tech. rep., National Bureau of Economic Research, Cambridge, MA, 2018.

[6] A. Barvinok, *A Course in Convexity*, Grad. Stud. Math. 54, AMS, Providence, RI, 2002.

[7] M. Bebendorf, *Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems*, Lect. Notes Comput. Sci. Eng. 63, Springer, Berlin, 2008.

[8] B. Beckermann and A. Townsend, *On the singular values of matrices with displacement structure*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1227–1248, https://doi.org/10.1137/16M1096426.

[9] R. M. Bell and Y. Koren, *Lessons from the Netflix prize challenge*, ACM SIGKDD Explorations Newsletter, 9 (2007), pp. 75–79.

[10] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, *Metagenes and molecular pattern discovery using matrix factorization*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 4164–4169.

[11] J. Calhoun, F. Cappello, L. N. Olson, M. Snir, and W. D. Gropp, *Exploring the feasibility of lossy compression for PDE simulations*, Internat. J. High Performance Comput. Appl., to appear, https://doi.org/10.1177/1094342018762036.

[12] S. Chan and E. Airoldi, *A consistent histogram estimator for exchangeable graph models*, in Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014, pp. 208–216.

[13] S. Chatterjee, *Matrix estimation by universal singular value thresholding*, Ann. Statist., 43 (2015), pp. 177–214.

[14] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, *Stochastic blockmodels with a growing number of classes*, Biometrika, 99 (2012), pp. 273–284.
[15] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, *Indexing by latent semantic analysis*, J. Amer. Soc. Inform. Sci., 41 (1990), pp. 391–407.
[16] I. S. Dhillon, *Co-clustering documents and words using bipartite spectral graph partitioning*, in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2001, pp. 269–274.
[17] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 543–560, https://doi.org/10.1137/0609045.
[18] S. Funk, *Netflix Update: Try This At Home*, 2006, https://sifter.org/simon/journal/20061211.html.
[19] Y. Gao and G. Church, *Improving molecular cancer class discovery through sparse non-negative matrix factorization*, Bioinformatics, 21 (2005), pp. 3970–3975.
[20] N. Gillis and Y. Shitov, *Low-Rank Matrix Approximation in the Infinity Norm*, preprint, https://arxiv.org/abs/1706.00078, 2017.
[21] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, *Limestone: High-throughput candidate phenotype generation via tensor factorization*, J. Biomed. Informatics, 52 (2014), pp. 199–211.
[22] J. C. Ho, J. Ghosh, and J. Sun, *Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization*, in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2014, pp. 115–124.
[23] D. Hsu, S. Kakade, and T. Zhang, *A tail inequality for quadratic forms of subgaussian random vectors*, Electron. Commun. Probab., 17 (2012), 52.
[24] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, in Conference in Modern Analysis and Probability (New Haven, CT, 1982), Contemp. Math. 26, AMS, Providence, RI, 1984, pp. 189–206.
[25] N. Kallus, X. Mao, and M. Udell, *Causal inference with noisy and missing covariates via matrix factorization*, in Advances in Neural Information Processing Systems, Montreal, Canada, 2018.
[26] N. Kallus and M. Udell, *Revealed preference at scale: Learning personalized preferences from assortment choices*, in Proceedings of the 2016 ACM Conference on Economics and Computation, ACM, New York, 2016, pp. 821–837.
[27] H. Kim and H. Park, *Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis*, Bioinformatics, 23 (2007), pp. 1495–1502.
[28] Y. Koren, R. Bell, and C. Volinsky, *Matrix factorization techniques for recommender systems*, Computer, 42 (2009), pp. 30–37.
[29] C. Lee, Y. Li, D. Shah, and D. Song, *Blind Regression via Nearest Neighbors under Latent Variable Models*, preprint, https://arxiv.org/abs/1705.04867, 2017.
[30] C. Lee and D. Shah, *Reducing Crowdsourcing to Graphon Estimation, Statistically*, preprint, https://arxiv.org/abs/1703.08085, 2017.
[31] D. Liben-Nowell and J. Kleinberg, *The link-prediction problem for social networks*, J. Assoc. Inform. Sci. Tech., 58 (2007), pp. 1019–1031.
[32] L. Lovász and B. Szegedy, *Limits of dense graph sequences*, J. Combin. Theory Ser. B, 96 (2006), pp. 933–957.
[33] J. Matoušek, *On variants of the Johnson–Lindenstrauss lemma*, Random Structures Algorithms, 33 (2008), pp. 142–156.
[34] A. Menon and C. Elkan, *Link prediction via matrix factorization*, in Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 2011, pp. 437–452.
[35] L. Omberg, G. H. Golub, and O. Alter, *A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 18371–18376.
[36] J. Pennington, R. Socher, and C. Manning, *Glove: Global vectors for word representation*, in Proceedings of the Empiricial Methods in Natural Language Processing Conference (EMNLP 2014), Doha, Qatar, 2014, 12.

[37] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in Advances in Neural Information Processing Systems, Whistler, BC, Canada, 2008, pp. 1177–1184.
[38] H. Robbins, *A remark on Stirling's formula*, Amer. Math. Monthly, 62 (1955), pp. 26–29.
[39] A. Schuler, V. Liu, J. Wan, A. Callahan, M. Udell, D. E. Stark, and N. H. Shah, *Discovering patient phenotypes using generalized low rank models*, in Pacific Symposium on Biocomputing, Pacific Symposium on Biocomputing, vol. 21, NIH Public Access, 2016, p. 144.
[40] S. W. Son, Z. Chen, W. Hendrix, A. Agrawal, W.-k. Liao, and A. Choudhary, *Data compression for the exascale computing era-survey*, Supercomput. Frontiers Innovations, 1 (2014), pp. 76–88.
[41] D. Song, C. E. Lee, Y. Li, and D. Shah, *Blind regression: Nonparametric regression for latent variable models via collaborative filtering*, in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2016, pp. 2155–2163.
[42] Z. Song, D. P. Woodruff, and P. Zhong, *Relative Error Tensor Low Rank Approximation*, preprint, https://arxiv.org/abs/1704.08246, 2017.
[43] A. Townsend and H. Wilber, *On the singular values of matrices with high displacement rank*, Linear Algebra Appl., 548 (2018), pp. 19–41.
[44] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, *More Practical Sketching Algorithms for Low-Rank Matrix Approximation*, Tech. rep. 2018-01, California Institute of Technology, Pasadena, CA, 2018.
[45] M. Udell, C. Horn, R. Zadeh, and S. Boyd, *Generalized low rank models*, Found. Trends Mach. Learn., 9 (2016), pp. 1–118.
[46] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Ser. Statist. Probab. Math. 47, Cambridge University Press, Cambridge, UK, 2018.
[47] Y. Wang and D. M. Blei, *The Blessings of Multiple Causes*, preprint, https://arxiv.org/abs/1805.06826, 2018.
[48] D. M. Witten, R. Tibshirani, and T. Hastie, *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis*, Biostatistics, 10 (2009), pp. 515–534.
[49] P. J. Wolfe and S. C. Olhede, *Nonparametric Graphon Estimation*, preprint, https://arxiv.org/abs/1309.5936, 2013.
[50] C. Yang, Y. Akimoto, D. W. Kim, and M. Udell, *OBOE: Collaborative filtering for automl initialization*, in NIPS Workshop on Automated Machine Learning, Montreal, Canada, 2018.