Margin-constrained Random Projections And Very Sparse Random Projections

Ping Li PINGLI@STAT.STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Trevor J. Hastie HASTIE@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Kenneth W. Church CHURCH@MICROSOFT.COM
Microsoft Research
Microsoft Corporation
Redmond, WA 98052, USA

Editor: March 19, 2006

Abstract

We[1] propose methods for improving both the *accuracy* and *efficiency* of random projections, the popular dimension reduction technique in machine learning and data mining, particularly useful for estimating pairwise distances. Let $\mathbf{A} \in \mathbb{R}^{n \times D}$ be our n points in D dimensions. This method multiplies $\mathbf{A}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, reducing the D dimensions down to just k . $\mathbf{R}$ typically consists of i.i.d. entries in $N(0,1)$. The cost of the projection mapping is $O(nDk)$.

This study proposes an improved estimator of pairwise distances with provably smaller variances (errors) by taking advantage of the marginal information.

We also propose *very sparse random projections* by replacing the $N(0,1)$ entries in $\mathbf{R}$ with entries in $\{-1, 0, 1\}$ with probabilities $\{\frac{1}{2\sqrt{D}}, 1 - \frac{1}{\sqrt{D}}, \frac{1}{2\sqrt{D}}\}$, for achieving a significant $\sqrt{D}$-fold speedup, with little loss in accuracy. Previously, Achlioptas proposed *sparse random projections* by using entries in $\{-1, 0, 1\}$ with probabilities $\{\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\}$, achieving a threefold speedup.

Keywords: Random Projections, Sampling, Maximum Likelihood, Asymptotic Analysis

1. Introduction

There has been considerable interest in the method of random projections (Vempala, 2004), a popular technique in machine learning and data mining for dimension reduction. Random projections are particularly useful for estimating pairwise distances.

We define a data matrix $\mathbf{A}$ of size $n \times D$ to be a collection of n data points $\{u_i\}_{i=1}^n \in \mathbb{R}^D$. All pairwise distances, $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, can be computed at the cost of $O(n^2 D)$, which is often prohibitive in modern data mining and information retrieval applications. For example, $\mathbf{A}$ can be the *term-by-document* matrix with n as the total number of word types and D as the total number of documents. In modern search engines, $n \approx 10^6 \sim 10^7$ and $D \approx 10^{10} \sim 10^{11}$. Note that $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is called *Gram*

1. Part of the work will be presented in COLT, Pittsburgh, Pennsylvania, June 22-25, 2006.

matrix in machine learning (especially kernel-based methods). Several methods for approximating Gram matrix have been proposed, e.g., Achlioptas et al. (2001); Drineas and Mahoney (2005).

To speed up the computation and save the storage space, one can generate a random projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ and multiply it with the original matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ to get

$$\mathbf{B} = \frac{1}{\sqrt{k}} \mathbf{A}\mathbf{R} \in \mathbb{R}^{n \times k}, \quad k \ll \min(n, D). \tag{1}$$

The (much smaller) matrix $\mathbf{B}$ preserves all pairwise distances of $\mathbf{A}$ in the expectation, provided that $\mathbf{R}$ consists of i.i.d. entries with zero mean and constant variance. Thus, we can achieve a substantial cost reduction for computing $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, from $O(n^2 D)$ to $O(nDk + n^2 k)$, within which the cost of the projection mapping (also called the processing time) is $O(nDk)$.

In information retrieval, we often do not have to materialize $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. Instead, databases and search engines are interested in storing the projected data $\mathbf{B}$ in the main memory for efficiently responding to input queries. While the original data matrix $\mathbf{A}$ is often too large, the projected data matrix $\mathbf{B}$ can be small enough to reside in the main memory.

The entries of $\mathbf{R}$ (denoted by $\{r_{ji}\}_{j=1}^{D}{}_{i=1}^{k}$) should be i.i.d. with zero mean. In fact, this is the only necessary condition for preserving pairwise distances (Arriaga and Vempala, 1999). However, different choices of r_{ji} can change the variances (average errors) and error tail bounds. It is often convenient to let r_{ji} follow a symmetric distribution about zero with unit variance. A "simple" distribution is the standard normal[2], i.e., $r_{ji} \sim N(0, 1)$. It is "simple" in terms of theoretical analysis. We call this special case as the *normal random projections*.

When $\mathbf{R}$ consists of entries in a general projection distribution (assuming zero mean and unit variance), we name this case as the *general random projections*. In particular, when $\mathbf{R}$ is chosen to have i.i.d. entries in

$$r_{ji} = \sqrt{s} \times \begin{cases} 1 & \text{with prob. } \frac{1}{2s} \\ 0 & \text{with prob. } 1 - \frac{1}{s} \\ -1 & \text{with prob. } \frac{1}{2s} \end{cases}, \tag{2}$$

we call it the *sparse random projections* if $1 \leq s \leq 3$ (Achlioptas, 2003), or the *very sparse random projections* if s is some fraction of D. Typically we recommend $s = \sqrt{D}$.

1.1 Our Improvements

We improve both the *accuracy* and *efficiency* of random projections.

The *accuracy* can be improved by taking advantage of marginal norms, which are easy to compute. All marginal norms can be computed in time $O(nD)$ with no need of generating any random numbers. This cost is negligible compared with $O(nDk)$, the cost of the projection mapping.

The *efficiency* can be improved by using a random projection matrix in (2) with $s = \sqrt{D}$. Intuitively, (2) can be regarded as some kind of sampling procedure with a sampling rate of $\frac{1}{s}$. Statistical results tell us that when the data are normal-like, $\log D$ of the data probably suffice (i.e., $s = \frac{D}{\log D}$), because of the exponential tail error bounds, common in normal-like distributions, such as binomial, gamma, etc. For better robustness, we recommend choosing s less aggressively (e.g., $s = \sqrt{D}$). With $s = \sqrt{D}$, the cost of the projection mapping becomes $O(n\sqrt{D}k)$.

2. It has been suggested to use *2-stable* distributions for random projections. Normal distribution is the only known *2-stable* distribution that has a closed-form density (Indyk, 2000, 2001).

Achlioptas (2003) recommended $s = 3$, to get a threefold speedup, with rigorous error bounds. Since the multiplications with $\sqrt{s}$ can be delayed, no floating point arithmetic is needed and all computation amounts to highly optimized database aggregation operations. In addition, because the original data are usually stored on disks, multiplying $\mathbf{A}$ with $\mathbf{R}$ involve expensive disk IO's; it is always preferable if one can read as little data as possible.

First experimentally tested on image and text data by Bingham and Mannila (2001), this method of *sparse random projections* has gained its popularity, e.g., Fradkin and Madigan (2003); Lin and Gunopulos (2003); Tang et al. (2004); Sahin et al. (2005).

1.2 More Applications And Experiments

Random projections have been widely used in machine learning (Kaski, 1998; Arriaga and Vempala, 1999; Dasgupta, 1999; Arora and Kannan, 2001; Bingham and Mannila, 2001; Achlioptas et al., 2001; Fradkin and Madigan, 2003; Fern and Brodley, 2003; Balcan et al., 2004), VLSI layout (Vempala, 1998), analysis of Latent Semantic Indexing (LSI) (Papadimitriou et al., 1998), set intersections (Charikar, 2002; Ravichandran et al., 2005), finding motifs in bio-sequences (Buhler and Tompa, 2002; Leung et al., 2005), face recognition (Goel et al., 2005), privacy preserving distributed data mining (Liu et al., 2006), to name a few.

1.3 Paper Organization

Some known results on *normal random projections* are reviewed in Section 2. Our contributions are summarized in Section 3. Section 4 describes our improved estimator and its statistical properties. Section 5 presents some new results on *general random projections*. Section 6 concerns *very sparse random projections*. All proofs are placed in the appendices after references.

2. Some Known Results On Normal Random Projections

The *normal random projections* multiply the data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ with a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ of i.i.d. $N(0,1)$ entries. Denote by $\{u_i\}_{i=1}^n \in \mathbb{R}^D$ the rows in $\mathbf{A}$ and by $\{v_i\}_{i=1}^n \in \mathbb{R}^k$ the rows of the projected data, i.e., $v_i = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}} u_i$. We focus on the leading two rows: u_1, u_2 and v_1, v_2. For convenience, we denote

$$m_1 = \|u_1\|^2 = \sum_{i=1}^D u_{1,i}^2, \qquad m_2 = \|u_2\|^2 = \sum_{i=1}^D u_{2,i}^2,$$

$$a = u_1^{\mathrm{T}} u_2 = \sum_{i=1}^D u_{1,i} u_{2,i}, \qquad d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a.$$

2.1 Moments

It is easy to show that (e.g., Lemma 1.3 of Vempala (2004))

$$\mathrm{E}\left(\|v_1\|^2\right) = \|u_1\|^2 = m_1, \qquad \mathrm{Var}\left(\|v_1\|^2\right) = \frac{2}{k} m_1^2, \tag{3}$$

$$\mathrm{E}\left(\|v_1 - v_2\|^2\right) = d, \qquad \mathrm{Var}\left(\|v_1 - v_2\|^2\right) = \frac{2}{k} d^2. \tag{4}$$

2.2 Distributions

It is easy to show that (e.g. Lemma 1.3 of Vempala (2004))

$$\frac{v_{1,j}}{\sqrt{m_1/k}} \sim N(0,1), \qquad \frac{\|v_1\|^2}{m_1/k} \sim \chi^2_k, \tag{5}$$

$$\frac{v_{1,j}-v_{2,j}}{\sqrt{d/k}} \sim N(0,1), \qquad \frac{\|v_1-v_2\|^2}{d/k} \sim \chi^2_k, \tag{6}$$

$$\left[\begin{array}{c} v_{1,j} \\ v_{2,j} \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \Sigma = \frac{1}{k}\left[\begin{array}{cc} m_1 & a \\ a & m_2 \end{array}\right]\right). \tag{7}$$

χ^2_k denotes a chi-squared random variable with k degrees of freedom. $v_{1,j}$ i.i.d. is any entry in $v_1 \in \mathbb{R}^k$.

2.3 The JL-embedding Bound

From the distributions of the projected data, various Johnson and Lindenstrauss (JL) embedding theorems[3] (Johnson and Lindenstrauss, 1984; Frankl and Maehara, 1987; Indyk and Motwani, 1998; Arriaga and Vempala, 1999; Dasgupta and Gupta, 2003; Achlioptas, 2003) have been proved for precisely determining k given some specified level of accuracy, for estimating the 2-norm distances.

The following Theorem is based on the best known result on JL-embedding.

Theorem 1 (Achlioptas (2003)) *Because* $\frac{\|v_1-v_2\|^2}{d/k} \sim \chi^2_k$*, the well-known chi-squared (Chernoff) tail bound gives (for any* $0 < \epsilon < 1$*)*

$$\mathbf{Pr}\left(\left|\|v_1-v_2\|^2 - d\right| \geq \epsilon d\right) \leq \exp\left(-\frac{k}{2}\left(\epsilon - \log(1+\epsilon)\right)\right) + \exp\left(-\frac{k}{2}\left(-\epsilon - \log(1-\epsilon)\right)\right) \tag{8}$$

$$\leq 2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right), \tag{9}$$

from which a JL-embedding bound follows immediately, using the union (i.e., Bonferroni) bound:

$$\frac{n^2}{2} 2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right) \leq n^{-\gamma} \Rightarrow k \geq k_0 = \frac{4+2\gamma}{\epsilon^2/2 - \epsilon^3/3}\log n, \tag{10}$$

i.e., if $k \geq k_0$*, then with probability at least* $1 - n^{-\gamma}$*, for any two rows* u_i*,* u_j *from the data matrix with* n *rows, we have* $(1-\epsilon)\|u_i - u_j\|^2 \leq \|v_i - v_j\|^2 \leq (1+\epsilon)\|u_i - u_j\|^2$*.*

These bounds also hold for sparse random projections in which r_{ji} *is sampled from (2) with* $s = 1$ *or* $s = 3$*.*

From a practical point of view, it appears a better idea to use the exact tail probabilities instead of the upper bounds. It is well-known that the Chernoff-type of tail bounds are in fact not tight when we compute the numerical numbers. For example, Figure 1(a) plots the ratio of the upper bound

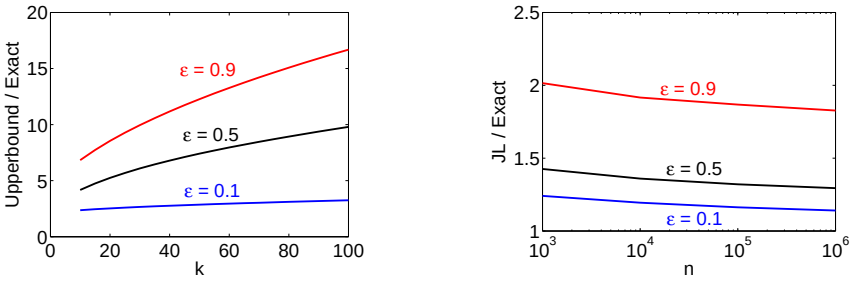
3. The JL-embedding bound (Johnson and Lindenstrauss, 1984) was defined much more generally than for estimating the 2-norm distances, the only case we consider.

(8) over the exact tail probability $\mathbf{Pr}\left(\left|\|v_1 - v_2\|^2 - d\right| \geq \epsilon d\right)$, indicating that the upper bound can easily magnify the exact tail probability by a factor of 5 or 10.

We suggest an alternative to the JL-embedding bound, by teratively solving an equation for k:

$$\frac{n^2}{2}\left(\mathbf{Pr}\left(\chi^2_k \geq (1+\epsilon)k\right) + \mathbf{Pr}\left(\chi^2_k \leq (1-\epsilon)k\right)\right) = \alpha \qquad \text{(e.g., } \alpha = 0.05\text{)}. \tag{11}$$

Figure 1(b) plots the ratio of the required sample size k given by the JL-embedding bound in (10) over the k computed from (11). The JL-embedding bound may increase the required sample size by (e.g.,) 50%, unncessarily.



(a) Tail prob. (b) Required k ($\alpha = 0.05$)

Figure 1: (a): The exact chi-squared tail probability can be seriously magnified by its Chernoff upper bound in (8). (b): The required sample size k given by the JL-embedding bound in Theorem 1 can be considerably larger (e.g., 50%) than the exact values computed from (11). When $\epsilon = 0.5$, the JL-embedding bound outputs $k = 404$, 514, 625 and 736, for $n = 10^3$, 10^4, 10^5 and 10^6, respectively.

2.4 Sign Random Projections

A variant of random projections is to store only the signs of the projected data, from which one can estimate the vector cosine angles, $\theta = \cos^{-1}\left(\frac{a}{\sqrt{m_1 m_2}}\right)$, by the following result (Goemans and Williamson, 1995; Charikar, 2002):

$$\mathbf{Pr}\left(\text{sign}(v_{1,j}) = \text{sign}(v_{2,j})\right) = 1 - \frac{\theta}{\pi}, \tag{12}$$

from which one can estimate $a = \cos(\theta)\sqrt{m_1 m_2}$, when m_1, m_2 are known, at the cost of some (small) bias.

The advantage of sign random projections is the saving in storing the projected data because only one bit per sign is needed. With sign random projections, we can compare vectors using hamming distances for which efficient algorithms are available (Indyk and Motwani, 1998; Charikar, 2002; Ravichandran et al., 2005).

3. Summary of Our Contributions

We derive some new theoretical results on random projections and propose a couple of improvements. The *accuracy* can be improved using the marginal norms; and the *efficiency* can be significantly improved by our *very sparse random projections*.

3.1 Normal Random Projections

In this case, $r_{ji} \sim N(0,1)$. We derive the variance formula and moment generating function for the estimator of inner products, not seen in prior literature. We derive a maximum likelihood estimator (MLE) of the inner products, taking advantage of the marginal norms. Extensive analysis on this estimator (e.g., asymptotic variance, bias, and third moment) is conducted.

In a practical sense, this new estimator can improve the JL-embedding bound in Theorem 1.

3.2 General Random Projections

In this case, r_{ji} follows some rather general distribution. We develop exact variance formulas for estimating both 2-norm distances and inner products. We show that if r_{ji} follows a *subgaussian* distribution, the same JL-embedding bound in Theorem 1 still holds.

3.3 Very Sparse Random Projections

In this case, r_{ji} follows the projection distribution defined in (2) with s being a fraction of D, the original data dimension. Typically, we recommend $s = \sqrt{D}$ to achieve a significant $\sqrt{D}$-fold speedup. Assuming bounded third or fourth moments on the original data, we show that when $s = \sqrt{D}$, the distributions and variances of the projected data converge to those of *normal random projections* at the rate of $O\left(\frac{1}{D^{1/4}}\right)$. Because D has to be large (otherwise there would be no need of seeking approximate answers), using *very sparse random projections* incurs little loss in accuracy.

We will explain how *very sparse random projections* can still be useful in heavy-tailed data.

4. Normal Random Projections Using Marginal Norms

This section is devoted to deriving an improved estimator of distances using margins.

Recall $u_i \in \mathbb{R}^D$ denotes data vectors in the original space and $v_i = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}} u_i \in \mathbb{R}^k$ denotes vectors in the projection space, $\mathbf{R} \in \mathbb{R}^{n \times D}$ consists of i.i.d $N(0,1)$ entries. We assume that the marginal norms, $\|u_i\|^2$ are known. As $\|u_1 - u_2\|^2 = \|u_1\|^2 + \|u_2\|^2 - 2u_1^{\mathrm{T}}u_2$, we only need to estimate the inner product $u_1^{\mathrm{T}}u_2$. Also recall that, for convenience, we denote $a = u_1^{\mathrm{T}}u_2$, $m_1 = \|u_1\|^2$, $m_2 = \|u_2\|^2$, and $d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a$.

4.1 A Basic Estimator Of Inner Products

The following lemma presents some basic results on the inner product $v_1^{\mathrm{T}}v_2$, proved in Appendix A.

Lemma 2 *Given $u_1, u_2 \in \mathbb{R}^D$, and a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ consisting of i.i.d. $N(0,1)$ entries, if we let $v_1 = \frac{1}{\sqrt{k}}\mathbf{R}^T u_1$, and $v_2 = \frac{1}{\sqrt{k}}\mathbf{R}^T u_2$, then*[4]

$$E\left(v_1^T v_2\right) = a, \quad Var\left(v_1^T v_2\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right), \quad E\left(v_1^T v_2 - a\right)^3 = \frac{2a}{k^2}\left(3m_1 m_2 + a^2\right), \quad (13)$$

with the moment generating function

$$E\left(\exp(v_1^T v_2 t)\right) = \left(1 - \frac{2}{k}at - \frac{1}{k^2}\left(m_1 m_2 - a^2\right)t^2\right)^{-\frac{k}{2}}, \quad (14)$$

4. A recent proof by (Liu et al., 2006, Lemma 5.4) showed that Var $\left(v_1^{\mathrm{T}}v_2\right) \leq \frac{2}{k}\left(\|u_1\|^2\|u_2\|^2\right)$.

where $\frac{-k}{\sqrt{m_1 m_2}-a} \leq t \leq \frac{k}{\sqrt{m_1 m_2}+a}$.

The moment generating function is convenient for deriving tail bounds using Chernoff inequality. However, it is more difficult to derive practically useful tail bounds for $v_1^{\mathrm{T}} v_2$ than for $\|v_1 - v_2\|^2$. One intuitive way to see this is via the coefficients of variations:

$$\frac{\sqrt{\mathrm{Var}\left(\|v_1 - v_2\|^2\right)}}{\|u_1 - u_2\|^2} = \sqrt{\frac{2}{k}} \quad \text{(constant)}, \qquad \frac{\sqrt{\mathrm{Var}\left(v_1^{\mathrm{T}} v_2\right)}}{u_1^{\mathrm{T}} u_2} \geq \sqrt{\frac{2}{k}} \quad \text{(unbounded)}.$$

A straightforward unbiased estimator of the inner product $a = u_1^{\mathrm{T}} u_2$ would be

$$\hat{a}_{MF} = v_1^{\mathrm{T}} v_2, \qquad \mathrm{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right), \tag{15}$$

where the subscript "MF" stands for "margin-free."

It is expected that if the marginal norms, $m_1 = \|u_1\|^2$ and $m_2 = \|u_2\|^2$, are given, one can do better. For example,

$$\hat{a}_{SM} = \frac{1}{2}\left(m_1 + m_2 - \|v_1 - v_2\|^2\right), \qquad \mathrm{Var}\left(\hat{a}_{SM}\right) = \frac{1}{2k}\left(m_1 + m_2 - 2a\right)^2, \tag{16}$$

where the subscript "SM" stands for "simple margin (method)." Note that

$$\mathrm{Var}\left(\hat{a}_{SM}\right) \geq \mathrm{Var}\left(\hat{a}_{MF}\right) \qquad \text{if } a \leq (m_1 + m_2) - \sqrt{\frac{1}{2}(m_1^2 + m_2^2) + 2m_1 m_2}. \tag{17}$$

4.2 A Maximum Likelihood Estimator (MLE) Of Inner Products

Assuming that marginal norms, m_1 and m_2 are known, we propose an estimator based on maximum likelihood (ML) in the following lemma, proved in Appendix B.

Lemma 3 *Suppose the margins, $m_1 = \|u_1\|^2$ and $m_2 = \|u_2\|^2$, are known; a maximum likelihood estimator (MLE), denoted by $\hat{a}_{MLE}$, is the solution to a cubic equation:*

$$a^3 - a^2\left(v_1^T v_2\right) + a\left(-m_1 m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1 m_2 v_1^T v_2 = 0, \tag{18}$$

which admits more than one real root with a (small) positive probability expressed as

$$\mathbf{Pr}\left(\textit{multiple real roots}\right) = \mathbf{Pr}\left(P^2(11 - Q^2/4 - 4Q + P^2) + (Q-1)^3 \leq 0\right), \tag{19}$$

where $P = \frac{v_1^T v_2}{\sqrt{m_1 m_2}}$, $Q = \frac{\|v_1\|^2}{m_1} + \frac{\|v_2\|^2}{m_2}$. This probability can be (crudely) bounded by

$$\mathbf{Pr}\left(\textit{multiple real roots}\right) \leq e^{-0.0085k} + e^{-0.0966k}. \tag{20}$$

When $a = m_1 = m_2$, this probability can be (sharply) bounded by

$$\mathbf{Pr}\left(\textit{multiple real roots} \mid a = m_1 = m_2\right) \leq e^{-1.5328k} + e^{-0.4672k}. \tag{21}$$

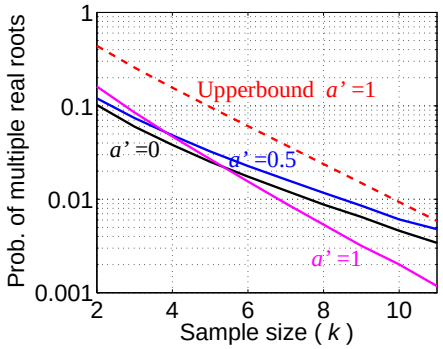


Figure 2: Simulations show that $\mathbf{Pr}$ (multiple real roots) decreases exponentially fast with respect to increasing sample size k. After $k \geq 8$, the probability of multiple roots becomes so small ($\leq 1\%$) that it can be safely ignored in practice. Here $a' = \frac{u_1^T u_2}{\sqrt{\|u_1\|^2 \|u_2\|^2}} = \frac{a}{\sqrt{m_1 m_2}}$.

Remark The cubic MLE equation (18) admits more than one real root with exponentially diminishing probability. Although the bound (20) is very crude, the probability (19) can be easily simulated. Figure 2 shows that the probability of having multiple real roots drops quickly to $< 1\%$ when $k \geq 8$.

To the best of our knowledge, there is no consensus on how to deal with multiple real roots (Barnett, 1966; Small et al., 2000); for example, a theoretically consistent solution is not always real; and a global maximum of the likelihood is not always consistent, e.g., (Kraft and LeCam, 1956).[5] For the large-scale applications we are interested in, the sample size k should be $\gg 10$; and the probability of multiple real roots will be negligible. Therefore, we suggest not to worry about multiple roots.

We expect that $\hat{a}_{MLE}$ will outperform both $\hat{a}_{MF}$ and $\hat{a}_{SM}$. We have the following lemma addressing the asymptotic behavior of MLE, proved in Appendix C.

Lemma 4 *$\hat{a}_{MLE}$ is asymptotically unbiased and converges to a normal random variable with mean a and variance (up to $O(k^{-2})$ terms)*

$$Var\left(\hat{a}_{MLE}\right) = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} \leq \min\left(Var\left(\hat{a}_{MF}\right), Var\left(\hat{a}_{SM}\right)\right). \tag{22}$$

The (asymptotic) bias $E\left(\hat{a}_{MLE} - a\right) = O(k^{-2})$. The $O(k^{-2})$ variance correction would be

$$Var\left(\hat{a}_{MLE}\right)_2^c = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} + \frac{1}{k^2}\frac{4(m_1 m_2 - a^2)^4}{(m_1 m_2 + a^2)^4} m_1 m_2 + O(k^{-3}). \tag{23}$$

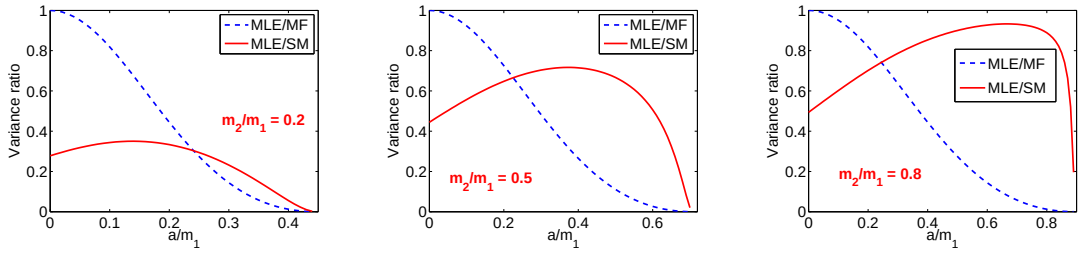
The asymptotic third moment is given by

$$E\left(\left(\hat{a}_{MLE} - a\right)^3\right) = \frac{-2a(3m_1 m_2 + a^2)(m_1 m_2 - a^2)^3}{k^2(m_1 m_2 + a^2)^3} + O(k^{-3}). \tag{24}$$

5. The regularity conditions in (Wald, 1949) to ensure that the global maximum likelihood estimate is consistent are diffi cult to check for models involving multiple roots.(Small et al., 2000)

Remark Maximum likelihood estimators can be seriously biased (hence not useful) in some cases, but usually the bias of an MLE is on the order of $O(k^{-1})$, which may be corrected by "Bartlett correction." (Bartlett, 1953) We are able to show that our MLE only has $O(k^{-2})$ bias hence no need for bias corrections. However, due to the existence of multiple real roots at very small k ($k < 8$), some small bias will be observable, as well as some small discrepancies between the observed variance (and third moment) and the theoretical asymptotic variance (and third moment).

Figure 3 verifies the inequality in (22) by plotting $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})}$ and $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{SM})}$. The improvement can be quite substantial. For example, $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})} = 0.2$ implies that in order to achieve the same mean square accuracy, the proposed MLE estimator needs only 20% of the samples required by the current margin-free (MF) estimator.



(a) (b) (c)

Figure 3: The ratios, $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})}$ and $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{SM})}$ verify that our proposed MLE has smaller variances than both the margin-free (MF) estimator and the simple margin (SM) method.

The ratio $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})} = \frac{\left(m_1 m_2 - a^2\right)^2}{\left(m_1 m_2 + a^2\right)^2} = \frac{(1-cos^2(\theta))^2}{(1+cos^2(\theta))^2}$ ranges from 0 to 1. When $\cos(\theta) \approx 1$ (i.e., $a^2 \approx m_1 m_2$), the improvement will be substantial. When $\cos(\theta) \approx 0$ (i.e., $a \approx 0$), we do not benefit from $\hat{a}_{MLE}$. Note that many studies (e.g., duplicate detection) are mostly interested in data points that are quite similar (e.g., $\cos(\theta) > 0.5$).

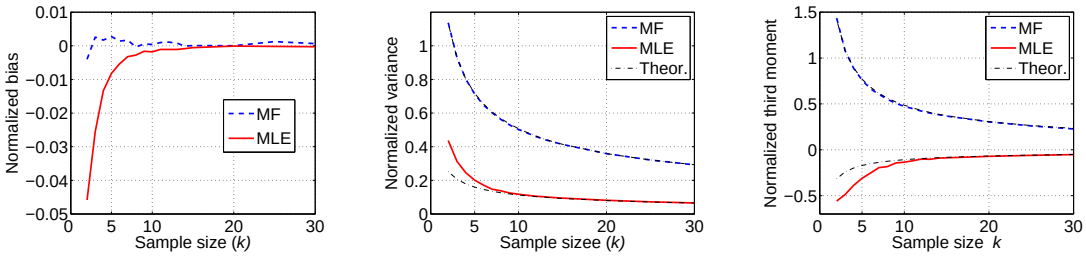
4.3 A Numerical Example

Figure 4 presents some numerical results, using two words "THIS" and "HAVE," from a chunk of MSN Web crawl data ($D = 2^{16}$). That is, $u_{1,j}$ ($u_{2,j}$) is the number of occurrences of word "THIS" (word "HAVE") in the jth document (Web page), $j = 1$ to D. The numerical results are consistent with our theoretical analysis.

4.4 Some Approximate Tail Bounds

We propose some approximate tail bounds, e.g., $\mathbf{Pr}\left(|\hat{a}_{MLE} - a|^2 \leq \epsilon a\right)$. Although it is in general not possible to describe the exact tail behavior for $\hat{a}_{MLE}$, it is also well-known that for "small deviations," (e.g., small ϵ) we can well approximate $\hat{a}_{MLE}$ by the asymptotic normality.[6]

6. More accurate approximations are also possible through *saddlepoint approximations* (Daniels, 1954; Goutis and Casella, 1999).



(a) Bias (b) Variance (c) Third moment

Figure 4: Estimations of the inner products between two vectors "THIS" and "HAVE." (a): $\frac{\text{bias}}{a}$, (b): $\frac{\sqrt{\text{Var}(\hat{a})}}{a}$, (c): $\frac{\sqrt[3]{\text{E}(\hat{a}-a)^3}}{a}$. This experiment verifies that (A): Marginal information can improve the estimations considerably. (B): As soon as $k > 8$, $\hat{a}_{MLE}$ is essentially unbiased and the asymptotic variance and third moment match simulations remarkably well. (C): The margin-free estimator ($\hat{a}_{MF}$) is unbiased and the theoretical variance and third moment are indistinguishable from the empirical values even for $k = 2$.

We often care about the "small deviation" behavior because we would like the estimated value to be close to the truth. However, when we estimate all pairwise distances simultaneously (as is considered in deriving the JL-embedding bounds), the common Bonferroni union bound[7] may push the tail to the "large deviation" range hence assuming asymptotic normality could be a concern. On the other hand, using the Bonferroni bound leads to larger k values; and larger k improves the accuracy of the asymptotic normality.

The next two subsections are devoted to normal approximations and generalized gamma approximations, respectively.

4.4.1 NORMAL APPROXIMATIONS

We propose approximate tail bounds for estimating both inner products and 2-norm distances.

The usual (margin-free) estimator for $d = \|u_1 - u_2\|^2$ is

$$\hat{d}_{MF} = \|v_1 - v_2\|^2 = d, \qquad \text{Var}\left(\hat{d}_{MF}\right) = \frac{2}{k}\|v_1 - v_2\|^4 = \frac{2d^2}{k}. \tag{25}$$

We can also estimate d from $\hat{a}_{MLE}$ as

$$\hat{d}_{MLE} = m_1 + m_2 - 2\hat{a}_{MLE}, \qquad \text{Var}\left(\hat{d}_{MLE}\right) = \frac{4}{k}\frac{(m_1 m_2 - a^2)^2}{m_1 m_2 + a^2} + O(k^{-2}). \tag{26}$$

Assuming normality, $\hat{a}_{MLE} \sim N\left(a, \text{Var}\left(\hat{a}_{MLE}\right)\right)$, the well-known normal bound yields

$$\mathbf{Pr}\left(|\hat{a}_{MLE} - a| \geq \epsilon a\right) \leq 2\exp\left(-\frac{k\epsilon^2}{2}\frac{a^2(m_1 m_2 + a^2)}{(m_1 m_2 - a^2)^2}\right). \tag{27}$$

7. The Bonferroni bound is well-known for being way too conservative. The method of false discovery rate (FDR) (Benjamini and Hochberg, 1995) has become a popular alternative.

Similarly, assuming $\hat{d}_{MLE} \sim N\left(d, \text{Var}\left(\hat{d}_{MLE}\right)\right)$ yields

$$\mathbf{Pr}\left(\left|\hat{d}_{MLE} - d\right| \geq \epsilon d\right) \leq 2\exp\left(-\frac{k}{4}\epsilon^2 \frac{d^2}{2}\frac{m_1 m_2 + a^2}{(m_1 m_2 - a^2)^2}\right). \tag{28}$$

Note that $\frac{d^2}{2}\frac{m_1 m_2 + a^2}{(m_1 m_2 - a^2)^2} \geq 1$ (unbounded), with equality holds when $m_1 = m_2 = a$. In practice, we have to choose some reasonable values for m_1, m_2 and a based on prior knowledge of the data, or for the regions we are most interested in. For example, if the reasonable value of $\frac{d^2}{2}\frac{m_1 m_2 + a^2}{(m_1 m_2 - a^2)^2}$ is around 3, the required sample size can be reduced by a factor of 3.

It would be interesting to see how normal approximation on $\hat{d}_{MF}$ affects its tail bound. Assuming normality, i.e., $\hat{d}_{MF} \sim N\left(d, \frac{2d^2}{k}\right)$, we can get

$$\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| \geq \epsilon d\right) \leq 2\exp\left(-\frac{k}{4}\epsilon^2\right), \tag{29}$$

which agrees with the exact bound in (9) on the dominating ϵ^2 term.

When applying normal approximations, it is important to watch out for the third central moments, which, to an extent, affect the rate of convergence:

$$\text{E}\left(\hat{d}_{MF} - d\right)^3 = \frac{8d^3}{k^2}, \quad \text{E}\left(\hat{d}_{MLE} - d\right)^3 = 8\frac{-2a(3m_1 m_2 + a^2)(m_1 m_2 - a^2)^3}{k^2(m_1 m_2 + a^2)^3}.$$

Some algebra can verify that

$$\left|\frac{\text{E}\left(\hat{d}_{MLE} - d\right)^3}{\text{E}\left(\hat{d}_{MF} - d\right)^3}\right| \leq \left(\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})}\right)^{\frac{3}{2}} \leq 1, \tag{30}$$

i.e., the third moment of $\hat{d}_{MLE}$ is well-behaved.

4.4.2 GENERALIZED GAMMA APPROXIMATION

The normal approximation matches the first two (asymptotic) moments. The accuracy can be further improved by matching the third moment. For example, (Li et al., 2006b) used a generalized gamma distribution to accurately approximate the finite-dimensional behavior of random matrix eigenvalues in some wireless communication channels.

For convenience, we assume $a \geq 0$, true in most applications. Assuming $-\hat{a}_{MLE} \sim G(\alpha, \beta, \xi)$, a generalized gamma random variable (Hougaard, 1986) with three parameters (α, β, ξ), we have

$$\text{E}\left(-\hat{a}_{MLE}\right) = \alpha\beta, \quad \text{Var}\left(-\hat{a}_{MLE}\right) = \alpha\beta^2, \quad \text{E}\left(-\hat{a}_{MLE} + a\right)^3 = (\xi + 1)\alpha\beta^3, \tag{31}$$

from which we can compute (α, β, ξ):

$$\begin{aligned}
&\alpha = \frac{ka^2(m_1 m_2 + a^2)}{(m_1 m_2 - a^2)^2} = k\alpha', \quad \beta = \frac{-(m_1 m_2 - a^2)^2}{k(m_1 m_2 + a^2)a} = \frac{-1}{k}\beta', \\
&\xi = \frac{2a^2(3m_1 m_2 + a^2)}{(m_1 m_2 + a^2)(m_1 m_2 - a^2)} - 1.
\end{aligned} \tag{32}$$

The generalized gamma distribution does not have a closed-form density, but it does have closed-form moment generating functions (Li et al., 2006b, (69)(70)):

$$\mathrm{E}\left(\exp\left(-\hat{a}_{MLE}t\right)\right) = \begin{cases} \exp\left(\frac{\alpha}{\xi-1}\left(1-\left(1-\beta\xi t\right)^{\frac{\xi-1}{\xi}}\right)\right) & \text{when } \xi > 1 \\ \exp\left(\frac{\alpha}{1-\xi}\left(\left(\frac{1}{1-\beta\xi t}\right)^{\frac{1-\xi}{\xi}}-1\right)\right) & \text{when } \xi < 1 \\ (1-\beta t)^{-\alpha} & \text{when } \xi = 1 \end{cases}$$

$\xi > 1$ happens when $\frac{a^2}{m_1 m_2} > \frac{\sqrt{17}-3}{4} = 0.2808$. Using the Chernoff inequality and assuming $\xi > 1$ (other cases are similar) , we obtain

$$\mathbf{Pr}\left(\hat{d}_{MLE} \geq (1+\epsilon)d\right) \leq \exp\left(-k\left(\left(\frac{2a}{2a-\epsilon d}\right)^{\xi-1}\left(\frac{\alpha'}{\xi-1}-\frac{a}{\beta'\xi}\right)-\frac{\alpha'}{\xi-1}+\frac{2a-\epsilon d}{2\beta'\xi}\right)\right),$$

$$\mathbf{Pr}\left(\hat{d}_{MLE} \leq (1-\epsilon)d\right) \leq \exp\left(-k\left(\left(\frac{2a}{2a+\epsilon d}\right)^{\xi-1}\left(\frac{\alpha'}{\xi-1}-\frac{a}{\beta'\xi}\right)-\frac{\alpha'}{\xi-1}+\frac{2a+\epsilon d}{2\beta'\xi}\right)\right),$$

which appear quite complicated, but still computable.

5. Some New Results On General Random Projections

In this section, we consider r_{ji} follows some distribution symmetric about zero with unit variance.

It is practically meaningful to study other projection distributions besides normals. For example, it is often more convenient and less expensive to sample form continuous or discrete uniform distributions. Using the sparse projection distribution defined in (2) can speedup the mapping significantly. Also we will soon show that using a subgaussian projection distribution can actually lead to some (slight) improvement. Previously, Achlioptas (2003) showed this improvement for a special case, i.e., the sparse projection distribution in (2) with $s = 1$ or $s = 3$.

Recall $v_i = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}}u_i$, $m_1 = \|u_1\|^2$, $m_2 = \|u_2\|^2$, $d = \|u_1 - u_2\|^2$, and $a = u_1^{\mathrm{T}}u_2$. We first derive the general variance formulas for $\|v_1\|^2$, $\|v_1 - v_2\|^2$, and $v_1^{\mathrm{T}}v_2$, in the following lemma, proved in Appendix D.

Lemma 5 *Suppose the projection matrix $\mathbf{R} \in \mathbb{R}^{D\times k}$ consists of i.i.d entries r_{ji} following any distribution symmetric about zero with unit variance, then*

$$\mathit{Var}\left(\|v_1\|^2\right) = \frac{1}{k}\left(2m_1^2 + (E(r_{ji}^4)-3)\sum_{j=1}^{D}u_{1,j}^4\right), \tag{33}$$

$$\mathit{Var}\left(\|v_1 - v_2\|^2\right) = \frac{1}{k}\left(2d^2 + (E(r_{ji}^4)-3)\sum_{j=1}^{D}\left(u_{1,j}-u_{2,j}\right)^4\right), \tag{34}$$

$$\mathit{Var}\left(v_1^T v_2\right) = \frac{1}{k}\left(m_1 m_2 + a^2 + (E(r_{ji}^4)-3)\sum_{j=1}^{D}u_{1,j}^2 u_{2,j}^2\right). \tag{35}$$

Compared with the corresponding variances in *normal random projections* (i.e., $r_{ji} \sim N(0,1)$), the general variances all have extra terms involving $(\mathrm{E}(r_{ji}^4) - 3)$, which is the "kurtosis" of r_{ji}.[8]

The kurtosis for $N(0,1)$ is zero. If we would like to achieve strictly smaller variances than *normal random projections*, we can choose r_{ji} with negative kurtosis. A couple of examples are:

- A continuous uniform distribution. It's kurtosis = $-\frac{6}{5}$.
- A discrete uniform distribution with T points. Its kurtosis = $-\frac{6}{5}\frac{T^2+1}{T^2-1}$, ranging between -2 (when $T = 2$) and $-\frac{6}{5}$ (when $T \to \infty$). The case with $T = 2$ is the same as (2) with $s = 1$.
- The sparse projection distribution defined in (2) with $1 \leq s < 3$.
- Discrete and continuous U-shaped distributions.

Besides variances, we are also interested in the higher order properties, such as the moment generating function (MGF). While it is difficult to derive the MGFs exactly, we can analyze the upper bounds when r_{ji} follows a subgaussian distribution.

We call a random variable r_{ji} follows a subgaussian distribution if

$$\text{for all } t > 0, \qquad \mathrm{E}\left(\exp(r_{ji}t)\right) \leq \exp\left(ct^2\right) \text{ for some positive constant } c. \tag{36}$$

In particular, we let $c = \frac{1}{2}$ (i.e., standard normal), and $\mathrm{E}\left(r_{ji}^p\right) \leq \mathrm{E}\left(Z^p\right)$, for any integer p, where $Z \sim N(0,1)$. This condition is satisfied by the distributions we are interested in, e.g., continuous and discrete uniform, the sparse projection distribution defined in (2) with $1 \leq s \leq 3$.

Appendix E proves the following lemma concerning the MGF of the projected data.

Lemma 6 *When r_{ji} follows a subgaussian satisfying the above conditions, then the MGFs of $\|v_1\|^2$ and $\|v_1 - v_2\|^2$ are bounded above by the MGF of χ_k^2 when $t > 0$, i.e.,*

$$E\left(\exp\left(\frac{\|v_1\|^2}{m_1/k}t\right)\right) \leq (1-2t)^{-\frac{k}{2}}, \tag{37}$$

$$E\left(\exp\left(\frac{\|v_1 - v_2\|^2}{d/k}t\right)\right) \leq (1-2t)^{-\frac{k}{2}}. \tag{38}$$

Consequently, the JL-embedding bound in Theorem 1 still holds.

6. Very Sparse Random Projections

This section is devoted to *very sparse random projections*, i.e., using i.i.d. $r_{ji} \in \sqrt{s} \times \{-1, 0, 1\}$ with probabilities $\{\frac{1}{2s}, 1 - \frac{1}{s}, \frac{1}{2s}\}$, as defined in (2). With this r_{ji}, we only need to sample $\frac{1}{s}$ of the data. Achlioptas (2003) proposed using $s = 3$ to get a threefold speedup. We suggest $s = \sqrt{D}$ to achieve a significant $\sqrt{D}$-fold speedup.

We show that under the assumption of bounded third or fourth moments on the original data, the variances and distributions of the projected data converge to those when $r_{ji} \sim N(0,1)$, at the rate of $O\left(\sqrt{s/D}\right)$, which is $O\left(\frac{1}{D^{1/4}}\right)$ when $s = \sqrt{D}$. As D is very large, this rate of convergence is so fast that we can basically treat *vary sparse random projections* the same as *normal random projections*, with little loss in accuracy.

8. The kurtosis $\gamma_2(r_{ji}) = \frac{\mathrm{E}((r_{ji} - \mathrm{E}(r_{ji}))^4)}{\mathrm{E}^2((r_{ji} - \mathrm{E}(r_{ji}))^2)} = \mathrm{E}(r_{ji}^4) - 3$, since we restrict r_{ji} to have zero mean and unit variance. Note that the kurtosis can not be smaller than -2 because of the Cauchy-Schwarz inequality: $\mathrm{E}^2(r_{ji}^2) \leq \mathrm{E}(r_{ji}^4)$.

6.1 Asymptotic Variances

We have already derived the variances for $\|v_1\|^2$, $\|v_1 - v_2\|^2$, and $v_1^{\mathrm{T}} v_2$ in Lemma 5. For example (note that in this case $\mathrm{E}(r_{ji}^4) = s$)

$$\mathrm{Var}\left(\|v_1\|^2\right) = \frac{1}{k}\left(2m_1^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^4\right) = \frac{2m_1}{k}\left(1 + (s-3)\frac{\sum_{j=1}^{D} u_{1,j}^4}{(\sum_{j=1}^{D} u_{1,j}^2)^2}\right).$$

Assuming that $u_{1,j}$ has a bounded fourth moment, by the strong law of large numbers (Durrett, 1995, (1.7.1)), we know

$$\frac{\sum_{j=1}^{D} u_{1,j}^4}{D} \to \mathrm{E}\left(u_{1,j}^4\right) \quad a.s., \qquad \frac{\sum_{j=1}^{D} u_{1,j}^2}{D} \to \mathrm{E}\left(u_{1,j}^2\right) \quad a.s., \tag{39}$$

$$(s-3)\frac{\sum_{j=1}^{D} u_{1,j}^4}{(\sum_{j=1}^{D} u_{1,j}^2)^2} = \frac{(s-3)}{D}\frac{\sum_{j=1}^{D} u_{1,j}^4/D}{(\sum_{j=1}^{D} u_{1,j}^2/D)^2} \to \frac{(s-3)}{D}\frac{\mathrm{E}\left(u_{1,j}^4\right)}{\left(\mathrm{E}\left(u_{1,j}^2\right)\right)^2} \to 0, \tag{40}$$

as long as we let $s = o(D)$ (e.g., $s = \frac{D}{\log D}$ or $s = \sqrt{D}$). When $s = \sqrt{D}$, the rate of convergence (for the variance) is $O\left(\frac{1}{\sqrt{D}}\right)$. In terms of the standard error (i.e., square root of variance), the rate of convergence would be $O\left(\frac{1}{D^{1/4}}\right)$. Therefore, asymptotically

$$\mathrm{Var}\left(\|v_1\|^2\right) \doteq \frac{2m_1}{k}, \qquad \mathrm{Var}\left(\|v_1 - v_2\|^2\right) \doteq \frac{2d}{k}, \qquad \mathrm{Var}\left(v_1^{\mathrm{T}} v_2\right) \doteq \frac{1}{k}\left(m_1 m_2 + a^2\right). \tag{41}$$

Here $\doteq$ denotes asymptotic equivalence as $D \to \infty$.

6.2 Asymptotic Distributions

The distributions of the projected data converge to the same distributions as when $r_{ji} \sim N(0,1)$. We only need to assume bounded $2+\delta$ (for any $\delta > 0$) moments to ensure the convergence. For convenience, especially for analyzing the rate of convergence, we assume bounded third moments.

Lemma 7 and Lemma 8 present the asymptotic distributions of v_1 and $v_1 - v_2$, respectively. Lemma 8 is strictly analogous to Lemma 7. We present them both because it is more straightforward to analyze v_1 and u_1, but we actually use the results for $v_1 - v_2$.

Lemma 7 *As $D \to \infty$,*

$$\frac{v_{1,j}}{\sqrt{m_1/k}} \overset{\mathcal{L}}{\Longrightarrow} N(0,1), \qquad \frac{\|v_1\|^2}{m_1/k} \overset{\mathcal{L}}{\Longrightarrow} \chi_k^2, \tag{42}$$

with the rate of convergence

$$|F_{v_{1,j}}(y) - \Phi(y)| \le 0.8\sqrt{s}\frac{\sum_{i=1}^{D}|u_{1,i}|^3}{m_1^{3/2}} \to 0.8\sqrt{\frac{s}{D}}\frac{E|u_{1,i}|^3}{\left(E\left(u_{1,i}^2\right)\right)^{3/2}} \to 0,$$

where $\overset{\mathcal{L}}{\Longrightarrow}$ denotes "convergence in distribution;" $F_{v_{1,j}}(y)$ is the empirical cumulative density function (CDF) of $v_{1,j}$ and $\Phi(y)$ is the standard normal $N(0,1)$ CDF.

Lemma 8 *As $D \to \infty$,*

$$\frac{v_{1,j} - v_{2,j}}{\sqrt{d/k}} \xrightarrow{\mathcal{L}} N(0,1), \qquad \frac{\|v_1 - v_2\|^2}{d/k} \xrightarrow{\mathcal{L}} \chi^2_k, \tag{43}$$

with the rate of convergence

$$|F_{v_{1,j}-v_{2,j}}(y) - \Phi(y)| \leq 0.8\sqrt{s}\frac{\sum_{i=1}^{D}|u_{1,i} - u_{2,i}|^3}{d^{3/2}} \to 0. \tag{44}$$

The above two lemmas show that both $v_{1,j}$ and $v_{1,j} - v_{2,j}$ are approximately normal, with the rate of convergence determined by $\sqrt{s/D}$, which is $O\left(\frac{1}{D^{1/4}}\right)$ when $s = \sqrt{D}$.

The next lemma concerns the joint distribution of $(v_{1,j}, v_{2,j})$.

Lemma 9 *As $D \to \infty$,*

$$\Sigma^{-\frac{1}{2}} \begin{bmatrix} v_{1,j} \\ v_{2,j} \end{bmatrix} \xrightarrow{\mathcal{L}} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right), \tag{45}$$

and

$$\mathbf{Pr}\left(\textit{sign}(v_{1,j}) = \textit{sign}(v_{2,j})\right) \to 1 - \frac{\theta}{\pi}. \tag{46}$$

where[9]

$$\Sigma = \frac{1}{k}\begin{bmatrix} m_1 & a \\ a & m_2 \end{bmatrix}, \qquad \theta = \cos^{-1}\left(\frac{a}{\sqrt{m_1 m_2}}\right).$$

The asymptotic normality allows us to approximately apply the JL-embedding and sign random projections. In particular, we suggest using the maximum likelihood estimator (MLE) for the inner products as derived in Lemma 3, when the margins (m_1 and m_2) are known.[10]

6.3 Heavy-tail and Term Weighting

The *very sparse random projections* are still useful in heavy-tailed data, mainly due to term weighting. The assumption of bounded third or fourth moments may not hold in heavy-tailed data. In fact, even the second moment may not exist if the data are modeled by a Pareto distribution.[11]

Heavy-tailed data are ubiquitous in large-scale applications (Leland et al., 1994; Faloutsos et al., 1999; Newman, 2005). The pairwise distances computed from heavy-tailed data are usually dominated by "outliers," i.e., exceptionally large entries. However, vector distances are meaningful only

9. Strictly speaking, we should write $\theta = \cos^{-1}\left(\frac{\mathrm{E}(u_{1,j}u_{2,j})}{\sqrt{\mathrm{E}(u_{1,j}^2)\mathrm{E}(u_{2,j}^2)}}\right)$.

10. Computing all margins cost $O(nD)$ while *very sparse random projections* cost $O(n\sqrt{D}k)$ (when $s = \sqrt{D}$), which is probably less than $O(nD)$. We consider that during various processing stages (e.g., data-collection, normalization, term-weighting), all entries of the data matrix $\mathbf{A}$ will be touched at least once. Also, as an important summary statistic, the marginal norms are likely already computed and stored in databases and search engines.

11. It is also common to model the data by other distributions such as lognormal, which has finite moments. In practice, we always deal with data of finite sizes, i.e., $D < \infty$, hence the empirical moments are always finite.

when all dimensions are more or less equally important. Therefore, in practice, various term weighting schemes have been proposed, e.g., (Manning and Schutze, 1999, Chapter 15.2)(Yu et al., 1982; Salton and Buckley, 1988; Dumais, 1991; Liu et al., 2001; Greiff, 2003).

It is well-known in practice that term weighting is vital. For example, as shown in (Leopold and Kindermann, 2002; Lan et al., 2005), in text categorization using support vector machine (SVM), choosing an appropriate term weighting scheme is far more important than tuning kernel functions of SVM. See similar comments in (Rennie et al., 2003) for the work on Naive Bayes text classifier.

We only list two simplest weighting methods. One variant of the *logarithmic weighting* keeps zero entries and replaces any non-zero count with $1 + \log(\text{original count})$. Another scheme is the *square root weighting*. In the same spirit of the Box-Cox transformation (Venables and Ripley, 2002, Chapter 6.8), these weighting schemes significantly reduce the tail thickness and make the data resemble normal.

Therefore, it is fair to say that assuming finite moments (third or fourth) is reasonable whenever the computed distances are meaningful.

However, there are also applications in which "distances" do not have to bear any clear meaning. For example, using random projections to estimate the joint sizes (set intersections). If it is expected that the original data are severely heavy-tailed (with no second moment or even first moment) and for some reason one prefers not to apply any term weighting, we recommend using $s \leq 3$, or any other subgaussian projection distributions previously discussed.

Finally, we shall point out that *very sparse random projections* can be fairly robust against (mildly) heavy-tailed data, especially when the second moments exist.

For example, instead of assuming finite fourth moments, as long as $D\frac{\sum_{i=1}^D u_{1,i}^4}{\left(\sum_{i=1}^D u_{1,i}^2\right)^2}$ grows slower than $O(\sqrt{D})$, the variances still convergence if $s = \sqrt{D}$. Similarly, analyzing the rate of converge to normality only requires that $\sqrt{D}\frac{\sum_{i=1}^D |u_{1,i}|^3}{\left(\sum_{i=1}^D u_{1,i}^2\right)^{3/2}}$ grows slower than $O(D^{1/4})$. We provide some additional theoretical analysis on heavy-tailed data in Appendix G.

6.4 Experimental Results

Some experimental results are presented for a sanity check, using one pair of words, "THIS" and "HAVE," provided by MSN. $D = 2^{16}$. Some summary statistics are listed in Table 1.

The data are certainly heavy-tailed as the empirical kurtoses for $u_{1,j}$ and $u_{2,j}$ are 178 and 190, respectively, far above zero. Therefore we do not expect that *very sparse random projections* with $s = \frac{D}{\log D} \approx 6000$ work well, though the results are actually not disastrous as shown in Figure 5(d).

We first test random projections on the original data, for $s = 1, 3, 256 = \sqrt{D}$ and $6000 \approx \frac{D}{\log D}$, presented in Figure 5. We then apply square root weighting and logarithmic weighting before random projections, with results presented in Figure 6 only for $s = 6000$.

These results are consistent with what we would expect:

- When s is small, i.e., $O(1)$, sparse random projections perform very similarly to normal random projections as shown in panels (a) and (b) of Figure 5 .
- With increasing s, the variances of very sparse random projections increase. With $s = \frac{D}{\log D}$, the errors are large (but not disastrous), because the data are heavy-tailed. With $s = \sqrt{D}$, very sparse random projections are robust.

Table 1: Some summary statistics of the word pair, "THIS" (u_1) and "HAVE" (u_2). γ_2 denotes the kurtosis. These expectations are computed empirically from the data. Two popular term weighting schemes are applied. The "square root weighting" replaces $u_{1,j}$ with $\sqrt{u_{1,j}}$ and the "logarithmic weighting" replaces any non-zero $u_{1,j}$ with $1 + \log u_{1,j}$.

	Unweighted	Square root	Logarithmic
$\gamma_2(u_{1,j})$	178.1	8.79	-0.159
$\gamma_2(u_{2,j})$	190.0	11.41	1.533
$\frac{\mathrm{E}(u_{1,j}^4)}{\mathrm{E}^2(u_{1,j}^2)}$	157.0	9.46	3.728
$\frac{\mathrm{E}(u_{2,j}^4)}{\mathrm{E}^2(u_{2,j}^2)}$	176.6	13.47	5.800
$\cos(\theta(u_1, u_2))$	0.797	0.761	0.722

- Since $\cos(\theta(u_1, u_2)) \approx 0.7 \sim 0.8$ in this case, marginal information can improve the estimation accuracy quite substantially. The asymptotic variances of $\hat{a}_{MLE}$ match the empirical variances of the asymptotic MLE estimator quite well, even for $s = \sqrt{D}$.

- After applying term weighting on the original data, very sparse random projections are almost as accurate as normal random projections, even for $s \approx \frac{D}{\log D}$, as shown in Figure 6.

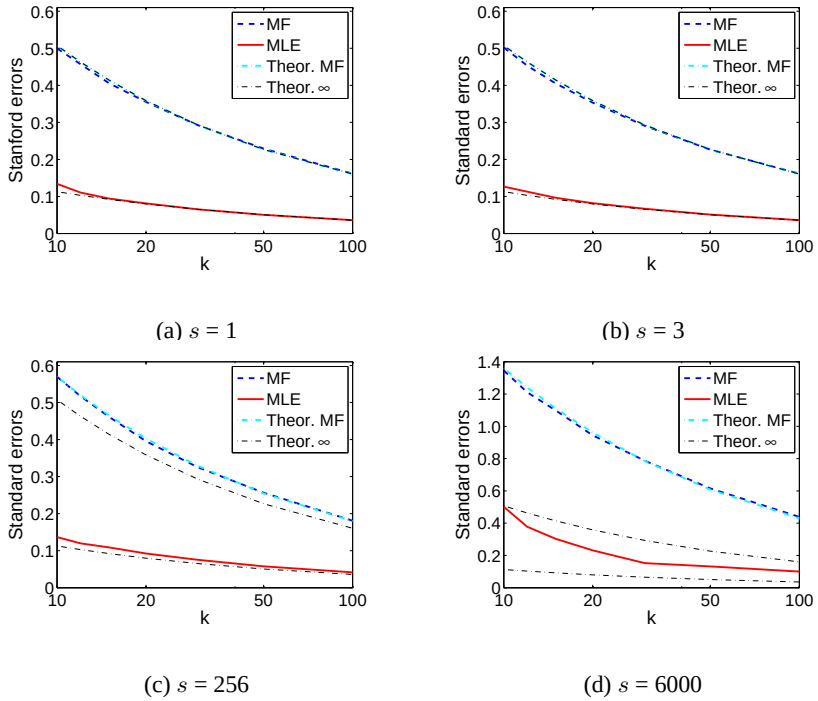
7. Conclusion

We provide some new theoretical results on random projections, a popular randomized approximate algorithm widely used in machine learning and data mining. The *accuracy* of random projections can be considerably improved by taking advantage of the marginal information. The *efficiency* can be significantly improved using a sparse random projection scheme. Our theoretical analysis suggests that we can achieve a significant $\sqrt{D}$-fold speedup of the projection mapping with little loss in accuracy, where D is the original data dimension. When the data are free of "outliers" (e.g., after careful term weighting), a cost of reduction by a factor of $\frac{D}{\log D}$ is also possible.

The limitation of random projections, shared by many other sketching algorithms, is that they are designed for specific summary statistics such as pairwise 2-norm distances. The authors' concurrent work (Li and Church, 2005; Li et al., 2006a) has proposed a new sketching-based sampling algorithm designed for sparse data, which is capable of estimating pairwise and multi-way distances (in either 2-norm or 1-norm). Theoretical comparisons indicate that this algorithm outperforms random projections in boolean data.

Acknowledgment

We would like to thank Dimitris Achlioptas for reading previous drafts of the material and providing insightful comments. We also thank Xavier Gabaix, Gabor Lugosi, and David Mason for pointers to useful references. Ping Li thanks Amir Dembo, Persi Diaconis, Bradley Efron, Jerome Friedman, Tze Leung Lai, Joseph Romano, Yiyuan She for many helpful conversations or references.



(a) $s = 1$ (b) $s = 3$

(c) $s = 256$ (d) $s = 6000$

Figure 5: Two words "THIS" (u_1) and "HAVE" (u_2) from the MSN Web crawl data are tested. $D = 2^{16}$. Sparse random projections are applied to estimated $a = u_1^{\mathrm{T}} u_2$, with four values of s: 1, 3, 256 = $\sqrt{D}$ and 6000 $\approx \frac{D}{\log D}$, in panels (a), (b), (c) and (d), respectively, presented in terms of the normalized standard error, $\frac{\sqrt{\mathrm{Var}(\hat{a})}}{a}$. There are five curves in each panel. The two labeled as "MF" and "Theor." overlap. "MF" stands for the empirical variance of the "Margin-free" estimator $\hat{a}_{MF}$; while "Theor. MF" for the theoretical variance of $\hat{a}_{MF}$, i.e., (35). The solid curve, labeled as "MLE," presents the empirical variance of $\hat{a}_{MLE}$. There are two curves both labeled as "Theor. ∞," for the asymptotic ($D \to \infty$) theoretical variances of $\hat{a}_{MF}$ (the higher curve) and $\hat{a}_{MLE}$ (the lower curve).

References

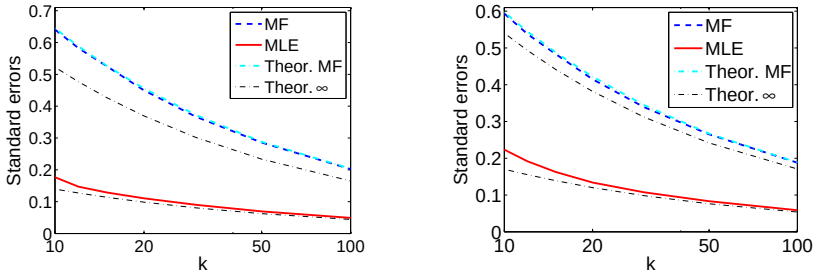
Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

Dimitris Achlioptas, Frank McSherry, and Bernhard Schölkopf. Sampling techniques for kernel methods. In *Proc. of NIPS*, pages 335–342, Vancouver, BC, Canada, 2001.

Shun-Ichi Amari. Differential geometry of curved exponential families-curvatures and information loss. *Annals of Statistics*, 10(2):357–385, 1982.

Sanjeev Arora and Ravi Kannan. Learning mixtures of arbituary gaussians. In *Proc. of STOC*, pages 247–257, Heraklion, Crete, Greece, 2001.

Rosa Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proc. of FOCS (Also to appear in Machine Learning)*, pages 616–623, New York, 1999.



(a) Square root ($s = 6000$)

(b) Logarithmic ($s = 6000$)

Figure 6: After applying term weighting on the original data, *very sparse random projections* are almost as accurate as *normal random projections*, even for $s = 6000 \approx \frac{D}{\log D}$. Note that the legends are the same as in Figure 5.

Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. On kernels, margins, and low-dimensional mappings. In *Proc. of ALT*, pages 194 – 205, Padova, Italy, 2004.

O.E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Chapman & Hall, London, UK, 1994.

V. D. Barnett. Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, 53(1/2):151–165, 1966.

M. S. Bartlett. Approximate confidence intervals, II. *Biometrika*, 40(3/4):306–317, 1953.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proc. of KDD*, pages 245–250, San Francisco, CA, 2001.

Jeremy Buhler and Martin Tompa. Finding motifs using random projections. *Journal of Computational Biology*, 9(2):225–242, 2002.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. of STOC*, pages 380–388, Montreal, Quebec, Canada, 2002.

G. P. Chistyakov and F. Gotze. Limit distributions of studentized means. *The Annals of Probability*, 32(1A): 28–77, 2004.

Francisco Jose De. A. Cysneiros, Sylvio Jose P. dos Santos, and Gass M. Cordeiro. Skewness and kurtosis for maximum likelihood estimator in one-parameter exponential family models. *Brazilian Journal of Probability and Statistics*, 15(1):85–105, 2001.

H. E. Daniels. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistic*, 25(4):631–650, 1954.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *Proc. of FOCS*, pages 634–644, New York, 1999.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60 – 65, 2003.

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

Susan T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236, 1991.

Richard Durrett. *Probability: Theory and Examples*. Duxbury Press, Belmont, CA, second edition, 1995.

Bradley Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, 3(6):1189–1242, 1975.

Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *Proc. of SIGMOD*, pages 251–262, Cambridge,MA, 1999.

William Feller. *An Introduction to Probability Theory and Its Applications (Volume II)*. John Wiley & Sons, New York, NY, second edition, 1971.

Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proc. of ICML*, pages 186–193, Washington, DC, 2003.

Silvia L. P. Ferrari, Denise A. Botter, Gauss M. Cordeiro, and Francisco Cribari-Neto. Second and third order bias reduction for one-parameter family models. *Stat. and Prob. Letters*, 30:339–345, 1996.

Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proc. of KDD*, pages 517–522, Washington, DC, 2003.

P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory A*, 44(3):355–362, 1987.

A. Fuchs, A. Joffe, and J. Teugels. Expectation of the ratio of the sum of squares to the square of the sum: Exact and asymptotic results. *Theory Probab. Appl.*, 46(2):243–255, 2002.

Navin Goel, George Bebis, and Ara Nefian. Face recognition experiments with random projection. In *Proc. of SPIE*, pages 426–437, Bellingham, WA, 2005.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the Association for Computing Machinery*, 42(6):1115–1145, 1995.

Constantino Goutis and George Casella. Explaining the saddlepoint approximation. *The American Statistician*, 53(3):216–224, 1999.

Warren R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proc. of SIGIR*, pages 11–19, Melbourne, Australia, 2003.

P. Hougaard. Survival models for heterogeneous populations. *Biometrika*, 73(2):387–396, 1986.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *FOCS*, pages 189–197, Redondo Beach,CA, 2000.

Piotr Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proc. of FOCS*, pages 10–33, Las Vegas, NV, 2001.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. of STOC*, pages 604–613, Dallas, TX, 1998.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Samuel Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. of IJCNN*, pages 413–418, Piscataway, NJ, 1998.

C. Kraft and L. LeCam. A remark on the roots of the maximum likelihood equation. *The Annals of Mathematical Statistics*, 27(3):1174–1177, 1956.

Man Lan, Chew Lim Tan, Hwee-Boon Low, and Sam Yuan Sung. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Proc. of WWW*, pages 1032–1033, Chiba, Japan, 2005.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, NY, second edition, 1998.

Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. Networking*, 2(1):1–15, 1994.

Edda Leopold and Jorg Kindermann. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3):423–444, 2002.

Henry C.M. Leung, Francis Y.L. Chin, S.M. Yiu, Roni Rosenfeld, and W.W. Tsang. Finding motifs with insufficient number of strong binding sites. *Journal of Computational Biology*, 12(6):686–701, 2005.

Ping Li and Kenneth W. Church. Using sketches to estimate two-way and multi-way associations. Technical Report TR-2005-115, Microsoft Research, Microsoft Corporation, WA, September 2005.

Ping Li, Kenneth W. Church, and Trevor J. Hastie. A sketched-based sampling algorithm on sparse data. Technical report, Department of Statistics, Stanford University, 2006a.

Ping Li, Debashis Paul, Ravi Narasimhan, and John Cioffi. On the distribution of SINR for the MMSE MIMO receiver and performance analysis. *IEEE Trans. Inform. Theory*, 52(1):271–286, 2006b.

Jessica Lin and Dimitrios Gunopulos. Dimensionality reduction by random projection and latent semantic indexing. In *Proc. of SDM*, San Francisco, CA, 2003.

Bing Liu, Yiming Ma, and Philip S. Yu. Discovering unexpected information from your competitors' web sites. In *Proc. of KDD*, pages 144–153, San Francisco, CA, 2001.

Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18 (1):92–106, 2006.

B. F. Logan, C. L. Mallows, S. O. Rice, and L. A. Shepp. Limit distributions of self-normalized sums. *The Annals of Probability*, 1(5):788–809, 1973.

Chris D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, New York, NY, 1979.

M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):232–351, 2005.

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proc. of PODS*, pages 159–168, Seattle,WA, 1998.

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proc. of ACL*, pages 622–629, Ann Arbor, MI, 2005.

Jason D. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proc. of ICML*, pages 616–623, Washington, DC, 2003.

Ozgur D. Sahin, Aziz Gulbeden, Fatih Emekçi, Divyakant Agrawal, and Amr El Abbadi. Prism: indexing multi-dimensional data in p2p networks using reference vectors. In *Proc. of ACM Multimedia*, pages 946–955, Singapore, 2005.

Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

L. R. Shenton and K. Bowman. Higher moments of a maximum-likelihood estimate. *Journal of Royal Statistical Society B*, 25(2):305–317, 1963.

I. S. Shiganov. Refinement of the upper bound of the constant in the central limit theorem. *Journal of Mathematical Sciences*, 35(3):2545–2550, 1986.

Christopher G. Small, Jinfang Wang, and Zejiang Yang. Eliminating multiple root problems in estimation. *Statistical Science*, 15(4):313–341, 2000.

Chunqiang Tang, Sandhya Dwarkadas, and Zhichen Xu. On scaling latent semantic indexing for large peer-to-peer systems. In *Proc. of SIGIR*, pages 112–121, Sheffield, UK, 2004.

Santosh Vempala. Random projection: A new approach to VLSI layout. In *Proc. of FOCS*, pages 389–395, Palo Alto, CA, 1998.

Santosh S. Vempala. *The Random Projection Method*. American Mathematical Society, Providence, RI, 2004.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, NY, fourth edition, 2002.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

Clement T. Yu, K. Lam, and Gerard Salton. Term weighting in information retrieval using the term precision model. *Journal of ACM*, 29(1):152–170, 1982.

Appendix A. Proof of Lemma 2

Recall $u_1, u_2 \in \mathbb{R}^D$, $\mathbf{R} \in \mathbb{R}^{D \times k}$ consists of i.i.d. $N(0,1)$ entries, $v_1 = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}}u_1$, and $v_2 = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}}u_2$. Let $\mathbf{R}_j$ be the j^{th} column of $\mathbf{R}$, $1 \leq j \leq k$. We can write the j^{th} element of v_1 to be $v_{1,j} = \frac{1}{\sqrt{k}}\mathbf{R}_j^{\mathrm{T}}u_1$, which is a weighted sum of normals hence also normal with mean $\mathrm{E}(v_{1,j}) = 0$, and variance $\mathrm{Var}(v_{1,j}) = \frac{1}{k}\|u_1\|^2$. Similarly, $\mathrm{E}(v_{2,j}) = 0$ and $\mathrm{Var}(v_{2,j}) = \frac{1}{k}\|u_2\|^2$.

Since $v_{1,j}v_{2,j} = \frac{1}{k}u_1^{\mathrm{T}}\mathbf{R}_j\mathbf{R}_j^{\mathrm{T}}u_2$, we have $\mathrm{Cov}\left(v_{1,j}, v_{2,j}\right) = \mathrm{E}\left(v_{1,j}v_{2,j}\right) = \frac{1}{k}u_1^{\mathrm{T}}u_2$. Therefore, $(v_{1,j}, v_{2,j})$ are jointly normal with zero mean and covariance Σ

$$\left[\begin{array}{c} v_{1,j} \\ v_{2,j} \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \Sigma = \frac{1}{k}\left[\begin{array}{cc} \|u_1\|^2 & u_1^{\mathrm{T}}u_2 \\ u_1^{\mathrm{T}}u_2 & \|u_2\|^2 \end{array}\right] = \frac{1}{k}\left[\begin{array}{cc} m_1 & a \\ a & m_2 \end{array}\right]\right). \tag{47}$$

Note that $v_1^{\mathrm{T}}v_2 = \sum_{j=1}^{k} v_{1,j}v_{2,j}$ is a sum of i.i.d. samples. It is easier to work with the conditional probability (Mardia et al., 1979, Theorem 3.2.3, 3.2.4):

$$v_{1,j}|v_{2,j} \sim N\left(\frac{a}{m_2}v_{2,j}, \frac{m_1m_2 - a^2}{km_2}\right), \tag{48}$$

from which we can get

$$\begin{aligned} &\mathrm{E}\left(v_{1,j}v_{2,j}\right)^2 = \mathrm{E}\left(\mathrm{E}\left(v_{1,j}^2v_{2,j}^2|v_{2,j}\right)\right) = \mathrm{E}\left(v_{2,j}^2\left(\frac{m_1m_2 - a^2}{km_2} + \left(\frac{a}{m_2}v_{2,j}\right)^2\right)\right) \\ =&\frac{m_2}{k}\frac{m_1m_2 - a^2}{km_2} + \frac{3m_2^2}{k^2}\frac{a^2}{m_2^2} = \frac{1}{k^2}\left(m_1m_2 + 2a^2\right). \end{aligned} \tag{49}$$

Therefore,

$$\mathrm{Var}\left(v_{1,j}v_{2,j}\right) = \frac{1}{k^2}\left(m_1m_2 + a^2\right), \quad \mathrm{Var}\left(v_1^{\mathrm{T}}v_2\right) = \frac{1}{k}\left(m_1m_2 + a^2\right). \tag{50}$$

The third moment can be proved similarly. In fact, one can compute any moments, using the moment generating function:

$$\begin{aligned} &\mathrm{E}\left(\exp(v_{1,j}v_{2,j}t)\right) = \mathrm{E}\left(\mathrm{E}\left(\exp(v_{1,j}v_{2,j}t)\right)|v_{2,j}\right) \\ =&\mathrm{E}\left(\exp\left(\left(\frac{a}{m_2}v_{2,j}\right)v_{2,j}t + \left(\frac{m_1m_2 - a^2}{km_2}\right)\left(v_{2,j}t\right)^2/2\right)\right) \\ =&\mathrm{E}\left(\exp\left(v_{2,j}^2\frac{k}{m_2}\left(\frac{a}{k}t + \frac{1}{k^2}\left(m_1m_2 - a^2\right)\frac{t^2}{2}\right)\right)\right) \\ =&\left(1 - \frac{2a}{k}t - \frac{1}{k^2}\left(m_1m_2 - a^2\right)t^2\right)^{-\frac{1}{2}}, \end{aligned} \tag{51}$$

Here, we use the fact that $\frac{v_{2,j}^2}{m_2/k} \sim \chi_1^2$, a standard chi-squared random variable with one degree of freedom. Note that $\mathrm{E}\left(\exp(Yt)\right) = \exp\left(\mu t + \sigma^2t^2/2\right)$ if $Y \sim N(\mu, \sigma^2)$; and $\mathrm{E}\left(\exp(Yt)\right) = (1-2t)^{-\frac{1}{2}}$ if $Y \sim \chi_1^2$. By independence, we have proved that

$$\mathrm{E}\left(\exp(v_1^{\mathrm{T}}v_2t)\right) = \left(1 - \frac{2}{k}at - \frac{1}{k^2}\left(m_1m_2 - a^2\right)t^2\right)^{-\frac{k}{2}}, \tag{52}$$

where $\frac{-k}{\sqrt{m_1m_2}-a} \leq t \leq \frac{k}{\sqrt{m_1m_2}+a}$. This completes the proof of Lemma 2.

Appendix B. Proof of Lemma 3

Using the results in Appendix A, we can write down the joint likelihood function of $\{v_{1,j}, v_{2,j}\}_{j=1}^k$:

$$\text{lik}\left(\{v_{1,j}, v_{2,j}\}_{j=1}^k\right) \propto |\Sigma|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}\sum_{j=1}^k \left[\begin{array}{cc} v_{1,j} & v_{2,j}\end{array}\right] \Sigma^{-1} \left[\begin{array}{c} v_{1,j} \\ v_{2,j}\end{array}\right]\right), \tag{53}$$

where (assuming $m_1 m_2 \neq a$ to avoid triviality)

$$|\Sigma| = \frac{1}{k^2}(m_1 m_2 - a^2), \qquad \Sigma^{-1} = \frac{k}{m_1 m_2 - a^2}\left[\begin{array}{cc} m_2 & -a \\ -a & m_1\end{array}\right], \tag{54}$$

which allow us to express the $\log$ likelihood function, $l(a)$, to be

$$l(a) = -\frac{k}{2}\log\left(m_1 m_2 - a^2\right) - \frac{k}{2}\frac{1}{m_1 m_2 - a^2}\sum_{j=1}^k \left(v_{1,j}^2 m_2 - 2v_{1,j}v_{2,j}a + v_{2,j}^2 m_1\right), \tag{55}$$

which is a curved exponential model (Efron, 1975) with canonical parameters $\left(\frac{1}{m_1 m_2 - a^2}, \frac{a}{m_1 m_2 - a^2}\right)$ and sufficient statistics $\left(\|v_1\|^2 m_2 + \|v_2\|^2 m_1, v_1^{\mathrm{T}} v_2\right)$. We notice that a special case in which $m_1 = m_2 = 1$ appeared in various places, (e.g.,) (Amari, 1982; Small et al., 2000) (Barndorff-Nielsen and Cox, 1994, Example 9.2.38).

Setting $l'(a)$ to zero, we can get $\hat{a}_{MLE}$, which is the solution to a cubic equation:

$$a^3 - a^2\left(v_1^{\mathrm{T}} v_2\right) + a\left(-m_1 m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1 m_2 v_1^{\mathrm{T}} v_2 = 0. \tag{56}$$

This MLE equation, however, may admit more than one real root and some root(s) may minimize (instead of maximizing) the $\log$ likelihood. By the well-known Cardano condition,

$$\mathbf{Pr}\left(\text{multiple real roots}\right) = \mathbf{Pr}\left(P^2(11 - Q^2/4 - 4Q + P^2) + (Q-1)^3 \leq 0\right), \tag{57}$$

where $P = \frac{v_1^{\mathrm{T}} v_2}{\sqrt{m_1 m_2}}$, $Q = \frac{\|v_1\|^2}{m_1} + \frac{\|v_2\|^2}{m_2}$. We can get a crude upper bound using the fact that $\mathbf{Pr}(A + B \leq 0) \leq \mathbf{Pr}(A \leq 0) + \mathbf{Pr}(B \leq 0)$. That is,

$$\mathbf{Pr}\left(\text{multiple real roots}\right) \leq \mathbf{Pr}\left(11 - Q^2/4 - 4Q \leq 0\right) + \mathbf{Pr}\left(Q - 1 \leq 0\right). \tag{58}$$

We will soon prove the following moment generating function

$$\mathrm{E}\left(\exp(Qt)\right) = \left(1 - \frac{4t}{k} + \frac{4t^2}{k^2}\left(\frac{m_1 m_2 - a^2}{m_1 m_2}\right)\right)^{-\frac{k}{2}}, \tag{59}$$

which enables us to prove the following upper bounds:

$$\mathbf{Pr}\left(Q - 1 \leq 0\right) \leq e^{-0.0966k}, \qquad \mathbf{Pr}\left(11 - Q^2/4 - 4Q \leq 0\right) \leq e^{-0.0085k}, \tag{60}$$

$$\mathbf{Pr}\left(\text{multiple real roots}\right) \leq e^{-0.0966k} + e^{-0.0085k}, \tag{61}$$

using the standard Chernoff inequality, e.g., $\mathbf{Pr}\left(Q > z\right) = \mathbf{Pr}\left(e^{Qt} > e^{zt}\right) \leq \mathrm{E}\left(e^{Qt}\right) e^{-zt}$, choosing t that minimizes the upper bound.

The upper bound (61) is very crude but nevertheless reveals that the probability of admitting multiple real roots decreases exponentially fast.

It turns out there is a simple exact solution for the special case of $a = m_1 = m_2$, i.e., $Q = 2P = \|v_1\|^2/m_1$, $kP = \frac{k\|v_1\|^2}{m_2} \sim \chi^2_k$, and a (sharp) upper bound:

$$\mathbf{Pr}\left(\text{multiple real roots}\right) = \mathbf{Pr}\left((P-3)^2 \geq 8\right) \leq e^{-1.5328k} + e^{-0.4672k}. \tag{62}$$

To complete the proof of Lemma 3, we need to outline the proof for the moment generating function $\mathrm{E}\left(\exp(Qt)\right)$. Using the conditional probability $v_{1,j}|v_{2,j}$, we know

$$\frac{km_2}{m_1m_2 - a^2} v_{1,j}^2 | v_{2,j} \sim \chi^2_{1,\lambda}, \quad \text{where } \lambda = \frac{ka^2}{m_2(m_1m_2 - a^2)} v_{2,j}^2, \tag{63}$$

$\chi^2_{1,\lambda}$ denotes a non-central chi-squared random variable with one degree of freedom and non-centrality parameter λ. If $Y \sim \chi^2_{1,\lambda}$, then $\mathrm{E}\left(\exp(Yt)\right) = \exp\left(\frac{\lambda t}{1-2t}\right)(1-2t)^{-\frac{1}{2}}$. Because

$$\mathrm{E}\left(\exp(Qt)\right) = \prod_{j=1}^{k} \mathrm{E}\left(\mathrm{E}\left(\exp\left(\frac{v_{1,j}^2}{m_1} + \frac{v_{2,j}^2}{m_2}\right) t \middle| v_{2,j}\right)\right), \tag{64}$$

we can obtain the moment generating function in (59) after some algebra.

Appendix C. Proof of Lemma 4

The large sample theory (Lehmann and Casella, 1998, Theorem 6.3.10) says that $\hat{a}_{MLE}$ is asymptotically unbiased and converges weakly to a normal random variable $N\left(a, \mathrm{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{\mathrm{I}(a)}\right)$, where $\mathrm{I}(a)$, the expected Fisher Information, is $\mathrm{I}(a) = -\mathrm{E}\left(l''(a)\right)$. Recall $l(a)$ is the log likelihood function obtained in Appendix B. Some algebra can show that

$$\mathrm{I}(a) = k\frac{m_1m_2 + a^2}{\left(m_1m_2 - a^2\right)^2}. \qquad \mathrm{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{k}\frac{\left(m_1m_2 - a^2\right)^2}{m_1m_2 + a^2}. \tag{65}$$

Applying the Cauchy-Schwarz inequality a couple of times can prove

$$\mathrm{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{k}\frac{\left(m_1m_2 - a^2\right)^2}{m_1m_2 + a^2} \leq \min\left(\mathrm{Var}\left(\hat{a}_{MF}\right), \mathrm{Var}\left(\hat{a}_{SM}\right)\right) \tag{66}$$

where $\mathrm{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}\left(m_1m_2 + a^2\right)$, $\mathrm{Var}\left(\hat{a}_{SM}\right) = \frac{1}{2k}\left(m_1 + m_2 - 2a\right)^2$.

W also need to analyze the higher-order accuracy of MLE, using stochastic Taylor expansions. We use some formulations appeared in (Bartlett, 1953; Shenton and Bowman, 1963; Ferrari et al., 1996; Cysneiros et al., 2001). The bias

$$\mathrm{E}\left(\hat{a}_{MLE} - a\right) = -\frac{\mathrm{E}\left(l'''(a)\right) + 2\mathrm{I}'(a)}{2\mathrm{I(a)}} + O(k^{-2}), \tag{67}$$

which is often called the "Bartlett correction." Some algebra can show

$$\mathrm{I}'(a) = \frac{2ka(3m_1m_2 + a^2)}{(m_1m_2 - a^2)^3}, \quad \mathrm{E}\left(l'''(a)\right) = -2\mathrm{I}'(a), \quad \mathrm{E}\left(\hat{a}_{MLE} - a\right) = O(k^{-2}). \tag{68}$$

The third central moment

$$\begin{aligned}\mathrm{E}\left(\hat{a}_{MLE}-a\right)^3 &= \frac{-3\mathrm{I}'(a)-\mathrm{E}\left(l'''(a)\right)}{\mathrm{I}^3(a)}+O(k^{-3})\\ &= -\frac{2a(3m_1m_2+a^2)(m_1m_2-a^2)^3}{k^2(m_1m_2+a^2)^3}+O(k^{-3}).\end{aligned} \tag{69}$$

The asymptotic variance $\mathrm{Var}\left(\hat{a}_{MLE}\right)$ is accurate up to the $O(k^{-2})$ terms. The $O(k^{-2})$ term of the variance, denoted by V_2^c, can be written as

$$\begin{aligned}V_2^c &= \frac{1}{\mathrm{I}^3(a)}\left(\mathrm{E}\left(l''(a)\right)^2-\mathrm{I}^2(a)-\frac{\partial\mathrm{E}\left(l'''(a)\right)+2\mathrm{I}'(a)}{\partial a}\right)\\ &\quad+\frac{1}{2\mathrm{I}^4(a)}\left(10\left(\mathrm{I}'(a)\right)^2-\mathrm{E}\left(l'''(a)\right)\left(\mathrm{E}\left(l'''(a)\right)-4\mathrm{I}'(a)\right)\right)\\ &= \frac{\mathrm{E}\left(\left(l''(a)\right)^2\right)-\mathrm{I}^2(a)}{\mathrm{I}^3(a)}-\frac{\left(\mathrm{I}'(a)\right)^2}{\mathrm{I}^4(a)},\quad(\text{as } \mathrm{E}\left(l'''(a)\right)+2\mathrm{I}'(a)=0).\end{aligned} \tag{70}$$

Computing $\mathrm{E}\left(\left(l''(a)\right)^2\right)$ requires some work. We can write

$$l''(a) = -\frac{k}{S^3}\left(T(4a^2+S)-S(m_1m_2+a^2)-4aS(v_1^{\mathrm{T}}v_2)\right), \tag{71}$$

where, for simplicity, we let $S=m_1m_2-a^2$ and $T=\|v_1\|^2m_2+\|v_2\|^2m_1-2v_1^{\mathrm{T}}v_2a$. Expanding $\left(l''(a)\right)^2$ generates terms involving T, T^2, $Tv_1^{\mathrm{T}}v_2$. Rewrite

$$\begin{aligned}T &= \frac{m_1m_2-a^2}{k}\left(\sum_{j=1}^{k}\frac{km_2}{m_1m_2-a^2}\left(v_{1,j}-\frac{a}{m_2}v_{2,j}\right)^2+\sum_{j=1}^{k}v_{2,j}^2\frac{k}{m_2}\right)\\ &= \frac{m_1m_2-a^2}{k}\left(\eta+\zeta\right)\end{aligned} \tag{72}$$

Recall $v_{1,j}|v_{2,j}\sim N\left(\frac{a}{m_2}v_{2,j},\frac{m_1m_2-a^2}{km_2}\right)$, and $v_{2,j}\sim N\left(0,\frac{m_2}{k}\right)$. Then

$$\eta\left|\{v_{1,j}\}_{j=1}^k\right.\sim\chi^2_k,\ (\text{independent of }\ \{v_{1,j}\}_{j=1}^k),\qquad\zeta=\sum_{j=1}^{k}v_{2,j}^2\frac{k}{m_2}\sim\chi^2_k, \tag{73}$$

implying that η and ζ are independent; and $\eta+\zeta\sim\chi^2_{2k}$. Thus,

$$\mathrm{E}(T)=2(m_1m_2-a^2)=2S,\qquad\qquad\mathrm{E}(T^2)=4S^2(1+\frac{1}{k}). \tag{74}$$

We also need to compute $\mathrm{E}\left(Tv_1^{\mathrm{T}}v_2\right)$. Rewrite

$$Tv_1^{\mathrm{T}}v_2=(v_1^{\mathrm{T}}v_2)\|v_1\|^2m_2+(v_1^{\mathrm{T}}v_2)\|v_2\|^2m_1-2\left(v_1^{\mathrm{T}}v_2\right)^2a. \tag{75}$$

Expand $(v_1^{\mathrm{T}} v_2)\|v_1\|^2$

$$(v_1^{\mathrm{T}} v_2)\|v_1\|^2 = \sum_{j=1}^{k} v_{1,j} v_{2,j} \sum_{j=1}^{k} v_{1,j}^2 = \sum_{j=1}^{k} v_{1,j}^3 v_{2,j} + \sum_{i=1}^{k} \left(v_{1,i}^2 \sum_{j \neq i} v_{1,j} v_{2,j} \right). \tag{76}$$

Again, applying the conditional probability argument, we can get $\mathrm{E}\left(v_{1,j}^3 v_{2,j}\right) = \frac{3am_1}{k^2}$, from which it follows that

$$\begin{aligned} \mathrm{E}\left((v_1^{\mathrm{T}} v_2)\|v_1\|^2\right) &= \sum_{j=1}^{k} \mathrm{E}\left(v_{1,j}^3 v_{2,j}\right) + \sum_{i=1}^{k} \left(\mathrm{E}\left(v_{1,i}^2\right) \sum_{j \neq i} \mathrm{E}\left(v_{1,j} v_{2,j}\right) \right) \\ &= \frac{3am_1}{k} + k\frac{m_1}{k} \sum_{j \neq i} \frac{a}{k} = am_1 \left(1 + \frac{2}{k}\right). \end{aligned} \tag{77}$$

To this end, we have all the necessary components for computing $\mathrm{E}\left(\left(l''(a)\right)^2\right)$. After some algebra, we obtain

$$\mathrm{E}\left(\left(l''(a)\right)^2\right) = \frac{k^2}{S^4} \left(\left(m_1 m_2 + a^2\right)^2 + \frac{4}{k} \left(m_1^2 m_2^2 + a^4 + 6a^2 m_1 m_2\right) \right), \tag{78}$$

$$V_2^c = \frac{4}{k^2} \frac{\left(m_1 m_2 - a^2\right)^4}{\left(m_1 m_2 + a^2\right)^4} m_1 m_2. \tag{79}$$

We complete the proof of Lemma 4.

Appendix D. Proof of Lemma 5

To show

$$\begin{aligned} \mathrm{Var}\left(\|v_1\|^2\right) &= \frac{1}{k} \left(2m_1^2 + (\mathrm{E}(r_{ji}^4) - 3) \sum_{j=1}^{D} u_{1,j}^4 \right), \\ \mathrm{Var}\left(\|v_1 - v_2\|^2\right) &= \frac{1}{k} \left(2d^2 + (\mathrm{E}(r_{ji}^4) - 3) \sum_{j=1}^{D} \left(u_{1,j} - u_{2,j}\right)^4 \right), \\ \mathrm{Var}\left(v_1^{\mathrm{T}} v_2\right) &= \frac{1}{k} \left(m_1 m_2 + a^2 + (\mathrm{E}(r_{ji}^4) - 3) \sum_{j=1}^{D} u_{1,j}^2 u_{2,j}^2 \right). \end{aligned}$$

Recall r_{ji} are i.i.d. with zero mean and unit variance.

The following expansions are useful for the proof.

$$m_1 m_2 = \sum_{i=1}^{D} u_{1,i}^2 \sum_{i=1}^{D} u_{2,i}^2 = \sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 + \sum_{i \neq i'}^{D} u_{1,i}^2 u_{2,i'}^2,$$

$$m_1^2 = \left(\sum_{i=1}^{D} u_{1,i}^2 \right)^2 = \sum_{i=1}^{D} u_{1,i}^4 + 2 \sum_{i<i'} u_{1,i}^2 u_{1,i'}^2,$$

$$a^2 = \left(\sum_{i=1}^{D} u_{1,i} u_{2,i} \right)^2 = \sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 + 2 \sum_{i<i'} u_{1,i} u_{2,i} u_{1,i'} u_{2,i'}.$$

Expand

$$v_{1,j}^2 = \left(\frac{1}{\sqrt{k}} \mathbf{R}_j^{\mathrm{T}} u_1 \right)^2 = \left(\frac{1}{\sqrt{k}} \sum_{i=1}^{D} \left(r_{ji} \right) u_{1,i} \right)^2 = \frac{1}{k} \left(\sum_{i=1}^{D} \left(r_{ji}^2 \right) u_{1,i}^2 + 2 \sum_{i<i'} \left(r_{ji} \right) u_{1,i} \left(r_{ji'} \right) u_{1,i'} \right),$$

from which it follows that

$$\mathrm{E}\left(v_{1,j}^2 \right) = \frac{1}{k} \sum_{i=1}^{D} u_{1,j}^2, \quad \mathrm{E}\left(\|v_1\|^2 \right) = \sum_{i=1}^{D} u_{1,i}^2 = m_1. \tag{80}$$

Expand

$$\begin{aligned} v_{1,j}^4 &= \frac{1}{k^2} \left(\sum_{i=1}^{D} \left(r_{ji}^2 \right) u_{1,i}^2 + 2 \sum_{i<i'} \left(r_{ji} \right) u_{1,i} \left(r_{ji'} \right) u_{1,i'} \right)^2 \\ &= \frac{1}{k^2} \left(\begin{array}{l} \sum_{i=1}^{D} \left(r_{ji}^4 \right) u_{1,i}^4 + 2 \sum_{i<i'} \left(r_{ji}^2 \right) u_{1,i}^2 \left(r_{ji'}^2 \right) u_{1,i'}^2 \\ +4 \left(\sum_{i<i'} \left(r_{ji} \right) u_{1,i} \left(r_{ji'} \right) u_{1,i'} \right)^2 \\ +4 \sum_{i=1}^{D} \left(r_{ji}^2 \right) u_{1,i}^2 \sum_{i<i'} \left(r_{ji} \right) u_{1,i} \left(r_{ji'} \right) u_{1,i'} \end{array} \right), \end{aligned}$$

from which it follows that

$$\mathrm{E}\left(v_{1,j}^4 \right) = \frac{1}{k^2} \left(\mathrm{E}(r_{ji}^4) \sum_{i=1}^{D} u_{1,j}^4 + 6 \sum_{i<i'} u_{1,i}^2 u_{1,i'}^2 \right), \tag{81}$$

$$\begin{aligned} \mathrm{Var}\left(v_{1,j}^2 \right) &= \frac{1}{k^2} \left(\mathrm{E}(r_{ji}^4) \sum_{i=1}^{D} u_{1,j}^4 + 6 \sum_{i<i'} u_{1,i}^2 u_{1,i'}^2 - \left(\sum_{i=1}^{D} u_{1,i}^2 \right)^2 \right) \\ &= \frac{1}{k^2} \left((\mathrm{E}(r_{ji}^4) - 1) \sum_{i=1}^{D} u_{1,j}^4 + 4 \sum_{i<i'} u_{1,i}^2 u_{1,i'}^2 \right) \\ &= \frac{1}{k^2} \left(2m_1^2 + (\mathrm{E}(r_{ji}^4) - 3) \sum_{i=1}^{D} u_{1,i}^4 \right), \end{aligned} \tag{82}$$

Therefore,

$$\text{Var}\left(\|v_1\|^2\right) = \frac{1}{k}\left(2m_1^2 + (\text{E}(r_{ji}^4) - 3)\sum_{j=1}^{D} u_{1,j}^4\right).$$

Similarly,

$$\text{Var}\left(\|v_1 - v_2\|^2\right) = \frac{1}{k}\left(2d^2 + (\text{E}(r_{ji}^4) - 3)\sum_{j=1}^{D} (u_{1,j} - u_{2,j})^4\right).$$

It remains to show Var $\left(v_1^{\text{T}} v_2\right)$. Expand

$$v_{1,j}v_{2,j} = \frac{1}{k}\left(\sum_{i=1}^{D}\left(r_{ji}^2\right)u_{1,i}u_{2,j} + \sum_{i\neq i'}\left(r_{ji}\right)u_{1,i}\left(r_{ji'}\right)u_{2,i'}\right),$$

and

$$\begin{aligned}
&v_{1,j}^2 v_{2,j}^2 \\
=&\frac{1}{k^2}\left(\sum_{i=1}^{D}\left(r_{ji}^2\right)u_{1,i}u_{2,i} + \sum_{i\neq i'}\left(r_{ji}\right)u_{1,i}\left(r_{ji'}\right)u_{2,i'}\right)^2 \\
=&\frac{1}{k^2}\begin{pmatrix}
\sum_{i=1}^{D}\left(r_{ji}^4\right)u_{1,i}^2u_{2,i}^2 + \\
2\sum_{i<i'}\left(r_{ji}^2\right)u_{1,i}u_{2,i}\left(r_{ji'}^2\right)u_{1,i'}u_{2,i'} + \\
\left(\sum_{i\neq i'}\left(r_{ji}\right)u_{1,i}\left(r_{ji'}\right)u_{2,i'}\right)^2 + \\
\sum_{i=1}^{D}\left(r_{ji}^2\right)u_{1,i}u_{2,i}\sum_{i\neq i'}\left(r_{ji}\right)u_{2,i}\left(r_{ji'}\right)u_{1,i'}
\end{pmatrix},
\end{aligned}$$

from which it follows

$$\begin{aligned}
&\text{E}\left(v_{1,j}^2 v_{2,j}^2\right) \\
=&\frac{1}{k^2}\left(\text{E}(r_{ji}^4)\sum_{i=1}^{D}u_{1,j}^2u_{2,j}^2 + 4\sum_{i<i'}u_{1,i}u_{2,i}u_{1,i'}u_{2,i'} + \sum_{i\neq i'}u_{1,i}^2u_{2,i'}^2\right) \\
=&\frac{1}{k^2}\left((\text{E}(r_{ji}^4) - 2)\sum_{i=1}^{D}u_{1,j}^2u_{2,j}^2 + \sum_{i\neq i'}u_{1,i}^2u_{2,i'}^2 + 2a^2\right) \\
=&\frac{1}{k^2}\left(m_1m_2 + (\text{E}(r_{ji}^4) - 3)\sum_{i=1}^{D}u_{1,j}^2u_{2,j}^2 + 2a^2\right), \qquad (83)
\end{aligned}$$

Therefore,

$$\text{Var}\left(v_{1,j}v_{2,j}\right) = \frac{1}{k^2}\left(m_1m_2 + a^2 + (\text{E}(r_{ji}^4) - 3)\sum_{i=1}^{D}u_{1,j}^2u_{2,j}^2\right), \qquad (84)$$

$$\text{Var}\left(v_1^{\text{T}}v_2\right) = \frac{1}{k}\left(m_1m_2 + a^2 + (\text{E}(r_{ji}^4) - 3)\sum_{i=1}^{D}u_{1,j}^2u_{2,j}^2\right). \qquad (85)$$

Appendix E. Proof of Lemma 6

To show that, when r_{ji} has zero mean and unit variance and satisfies $\mathrm{E}\left(r_{ji}^p\right) \leq \mathrm{E}\left(Z^p\right)$ for any positive integer p, where $Z \sim N(0,1)$, we have for all $0 \leq t < \frac{1}{2}$,

$$\mathrm{E}\left(\exp\left(\frac{\|v_1\|^2}{m_1/k}t\right)\right) \leq (1-2t)^{-\frac{k}{2}} \tag{86}$$

$$\mathrm{E}\left(\exp\left(\frac{\|v_1-v_2\|^2}{d/k}t\right)\right) \leq (1-2t)^{-\frac{k}{2}} \tag{87}$$

Recall $v_{1,j}^2 = \left(\frac{1}{\sqrt{k}}\sum_{i=1}^D r_{ji}u_{1,i}\right)^2$. Note that

$$\mathrm{E}\left(\exp(v_{1,j}^2 t)\right) = \prod_{p=0}^{\infty}\mathrm{E}\left(\frac{v_{1,j}^{2p}t^p}{p!}\right). \tag{88}$$

If we expand $\mathrm{E}\left(v_{1,j}^{2p}\right) = \mathrm{E}\left(\left(\frac{1}{\sqrt{k}}\sum_{i=1}^D r_{ji}u_{1,i}\right)^{2p}\right)$, all (odd) terms involving $(r_{ji}u_{1,i})^{2l+1}$ will vanish, because r_{ji} has zero mean and i.i.d. Therefore, we can replace r_{ji} by i.i.d. $Z_i \sim N(0,1)$ to bound

$$\mathrm{E}\left(\exp(v_{1,j}^2 t)\right) \leq \mathrm{E}\left(\exp\left(\frac{1}{\sqrt{k}}\sum_{i=1}^D Z_i u_{1,i}\right)^2 t\right) = \exp\left(m_1/k\right)(1-2t)^{-\frac{1}{2}}, \tag{89}$$

from which it follows

$$\mathrm{E}\left(\exp\left(\frac{\|v_1\|^2}{m_1/k}t\right)\right) = \prod_{j=1}^{k}\mathrm{E}\left(\exp\left(\frac{v_{1,j}^2}{m_1/k}t\right)\right) \leq (1-2t)^{-\frac{k}{2}}. \tag{90}$$

Similarly, we can show

$$\mathrm{E}\left(\exp\left(\frac{\|v_1-v_2\|^2}{d/k}t\right)\right) \leq (1-2t)^{-\frac{k}{2}}. \tag{91}$$

Appendix F. Proofs of Lemmas 7, 8, and 9

Restate Lemma 7: Assuming $\mathrm{E}\left(|u_{1,j}|^3\right) < \infty$, as $D \to \infty$,

$$\frac{v_{1,j}}{\sqrt{m_1/k}} \overset{\mathcal{L}}{\Longrightarrow} N(0,1), \qquad \frac{\|v_1\|^2}{m_1/k} \overset{\mathcal{L}}{\Longrightarrow} \chi_k^2,$$

with the rate of convergence

$$|F_{v_{1,j}}(y) - \Phi(y)| \leq 0.8\sqrt{s}\frac{\sum_{i=1}^D |u_{1,i}|^3}{m_1^{3/2}} \to 0.8\sqrt{\frac{s}{D}}\frac{\mathrm{E}|u_{1,i}|^3}{\left(\mathrm{E}\left(u_{1,i}^2\right)\right)^{3/2}} \to 0,$$

where $\stackrel{\mathcal{L}}{\Longrightarrow}$ denotes "convergence in distribution," $F_{v_{1,j}}(y)$ is the empirical cumulative density function (CDF) of $v_{1,j}$ and $\Phi(y)$ is the standard normal $N(0,1)$ CDF.

Proof of Lemma 7 The Lindeberg central limit theorem (CLT) and the Berry-Esseen theorem are needed for the proof (Feller, 1971, Theorems VIII.4.3 and XVI.5.2).[12]

Write $v_{1,j} = \frac{1}{\sqrt{k}}\mathbf{R}_j^{\mathrm{T}} u_1 = \sum_{i=1}^{D} \frac{1}{\sqrt{k}}\,(r_{ji})\,u_{1,i} = \sum_{i=1}^{D} z_i$, with $z_i = \frac{1}{\sqrt{k}}\,(r_{ji})\,u_{1,i}$. Then

$$\mathrm{E}(z_i) = 0, \quad \mathrm{Var}(z_i) = \frac{u_{1,i}^2}{k}, \quad \mathrm{E}(|z_i|^{2+\delta}) = s^{\frac{\delta}{2}}\frac{|u_{1,i}|^{2+\delta}}{k^{(2+\delta)/2}}, \forall \delta > 0.$$

Let $s_D^2 = \sum_{i=1}^{D} \mathrm{Var}(z_i) = \frac{\sum_{i=1}^{D} u_{1,i}^2}{k} = \frac{m_1}{k}$. Assume the Lindeberg condition

$$\frac{1}{s_D^2}\sum_{i=1}^{D} \mathrm{E}\left(z_i^2; |z_i| \geq \epsilon s_D\right) \to 0, \text{ for any } \epsilon > 0.$$

Then

$$\frac{\sum_{i=1}^{D} z_i}{s_D} = \frac{v_{1,j}}{\sqrt{m_1/k}} \stackrel{\mathcal{L}}{\Longrightarrow} N(0,1),$$

which immediately leads to

$$\frac{v_{1,j}^2}{m_1/k} \stackrel{\mathcal{L}}{\Longrightarrow} \chi_1^2, \qquad \frac{\|v_1\|^2}{m_1/k} = \sum_{j=1}^{k}\left(\frac{v_{1,j}^2}{m_1/k}\right) \stackrel{\mathcal{L}}{\Longrightarrow} \chi_k^2.$$

We need to go back and check the Lindeberg condition.

$$\begin{aligned}
&\frac{1}{s_D^2}\sum_{i=1}^{D} \mathrm{E}\left(z_i^2; |z_i| \geq \epsilon s_D\right) \leq \frac{1}{s_D^2}\sum_{i=1}^{D} \mathrm{E}\left(\frac{|z_i|^{2+\delta}}{(\epsilon s_D)^{\delta}}\right) \\
&= \left(\frac{s}{D}\right)^{\frac{\delta}{2}} \frac{1}{\epsilon^{\delta}} \frac{\sum_{i=1}^{D} |u_{1,i}|^{2+\delta}/D}{\left(\sum_{i=1}^{D} u_{1,i}^2/D\right)^{(2+\delta)/2}} \\
&\to \left(\frac{o(D)}{D}\right)^{\frac{\delta}{2}} \frac{1}{\epsilon^{\delta}} \frac{\mathrm{E}|u_{1,j}|^{2+\delta}}{\left(\mathrm{E}(u_{1,j}^2)\right)^{(2+\delta)/2}} \to 0,
\end{aligned}$$

provided $\mathrm{E}|u_{1,j}|^{2+\delta} < \infty$, for some $\delta > 0$, which is much weaker than our assumption that $\mathrm{E}(|u_{1,j}|^3) < \infty$. It remains to show the rate of convergence using the Berry-Esseen theorem. Let $\rho_D = \sum_{i=1}^{D} \mathrm{E}|z_i|^3 = \frac{c^{1/2}}{k^{3/2}} \sum_{i=1}^{D} |u_{1,i}|^3$, then

$$|F_{v_{1,j}}(y) - \Phi(y)| \leq 0.8\frac{\rho_D}{s_D^3} = 0.8\sqrt{s}\frac{\sum_{i=1}^{D} |u_{1,i}|^3}{m_1^{3/2}} \to 0.8\sqrt{\frac{s}{D}}\frac{\mathrm{E}|u_{1,i}|^3}{\left(\mathrm{E}\left(u_{1,i}^2\right)\right)^{3/2}} \to 0.$$

The proof for Lemma 8 is exactly analogous to the proof of Lemma 7.

Lemma 9 concerns the joint distribution of $(v_{1,j}, v_{2,j})$. The proof is also straightforward since we assume all bounded third moments.

12. The best Berry-Essen constant 0.7915 (≈ 0.8) is from (Shiganov, 1986).

Appendix G. Very Sparse Random Projections In Heavy-tailed Data

We illustrate that *very sparse random projections* are fairly robust against heavy-tailed data, by a Pareto distribution.

The assumption of finite moments has simplified the analysis of convergence a great deal. For example, assuming finite $(2+\delta)$th moments, $0\delta > 0$, and $s = o(D)$, we have

$$(s)^{\delta/2}\frac{\sum_{i=1}^{D}|u_{1,i}|^{2+\delta}}{\left(\sum_{i=1}^{D}(u_{1,i}^2)\right)^{1+\delta/2}} = \left(\frac{s}{D}\right)^{\delta/2}\frac{\sum_{i=1}^{D}|u_{1,i}|^{2+\delta}/D}{\left(\sum_{i=1}^{D}(u_{1,i}^2)/D\right)^{1+\delta/2}}$$
$$\rightarrow \left(\frac{s}{D}\right)^{\delta/2}\frac{\mathrm{E}\left(u_{1,j}^{2+\delta}\right)}{\left(\mathrm{E}\left(u_{1,j}^2\right)\right)^{1+\delta/2}} \rightarrow 0. \quad (92)$$

Note that $\delta = 2$ corresponds to the rate of convergence for the variances of the projected data, and $\delta = 1$ corresponds to the rate of convergence for asymptotic normality in Lemma 7. From the proof of Lemma 7, we can see that the convergence of (92) (to zero) with any $\delta > 0$ suffices for achieving asymptotic normality.

The most common model for heavy-tailed data is the Pareto distribution with the density function[13] $f(x;\alpha) = \frac{\alpha}{x^{\alpha+1}}$, whose mth moment = $\frac{\alpha}{\alpha-m}$, only defined if $\alpha > m$. The measurements of α for many types of data are available in Newman (2005). For example, $\alpha = 1.2$ for the word frequency, $\alpha = 2.04$ for the citations to papers, $\alpha = 2.51$ for the copies of books sold in the US.

For simplicity, we consider $2 < \alpha \leq 2+\delta \leq 4$. Under this assumption, the asymptotic normality of the projected data is guaranteed and it remains to show the rate of convergence of variances and distributions. In this case, the second moment $\mathrm{E}\left(u_{1,j}^2\right)$ exists. The sum $\sum_{i=1}^{D}|u_{1,i}|^{2+\delta}$ grows as $O\left(D^{(2+\delta)/\alpha}\right)$ as shown in (Durrett, 1995, Example 2.7.4).[14] Thus, we can write

$$s^{\delta/2}\frac{\sum_{i=1}^{D}|u_{1,i}|^{2+\delta}}{\left(\sum_{i=1}^{D}(u_{1,i}^2)\right)^{1+\delta/2}} = O\left(\frac{s^{\delta/2}}{D^{1+\delta/2-\frac{2+\delta}{\alpha}}}\right) = \begin{cases} O\left(\frac{s}{D^{2-4/\alpha}}\right) & \delta = 2 \\ O\left(\frac{s}{D^{3-6/\alpha}}\right)^{1/2} & \delta = 1 \end{cases}, \quad (93)$$

from which we can choose s using prior knowledge of α.

For example, suppose $\alpha = 3$ and $s = \sqrt{D}$. (93) indicates that the rate of convergence for variances would be $O(D^{1/12})$ in terms of the standard error. (93) also verifies that the rate of convergence to normality is $O(D^{1/4})$, as expected.

When $2 < \alpha < 3$, we suggest $s = D^{1-2/\alpha}$, achieving a convergence rate of $O\left(\sqrt{D^{1-2/\alpha}}\right)$ for the standard error, and a rate of $O\left(D^{1-2/\alpha}\right)$ for the convergence to asymptotic normality.

What if $\alpha < 2$? The second moment no longer exists. The analysis may involve the so-called *self-normalizing sums* (Logan et al., 1973; Fuchs et al., 2002; Chistyakov and Gotze, 2004). Intuitively, in (92), we can let both denominator and numerator grow together and the overall growth may be still (hopefully) under control. We will not delve into this topic. Instead, we suggest using $s \leq 3$, or any other subgaussian projection distributions.

13. Note that in general, a Pareto distribution has an addition parameter x_{min}, and $f(x;\alpha,x_{min}) = \frac{\alpha x_{min}}{x^{\alpha+1}}$ with $x \geq x_{min}$. Since we are only interested in the relative ratio of moments, we can without loss of generality assume $x_{min} = 1$. Also note that in Newman (2005), their "α" is equal to our $\alpha + 1$.

14. Note that if $x \sim$ Pareto(α), then $x^t \sim$ Pareto(α/t).