Dimension Reduction Techniques

for Efficiently Computing Distances in Massive Data

Workshop on Algorithms for Modern Massive Data Sets

June 22, 2006

Ping Li, Trevor Hastie, and Kenneth Church (MSR)

Department of Statistics

Stanford University

Let's Begin with $\mathbf{A}\mathbf{A}^\mathsf{T}$

The data matrix $\mathbf{A} \in \mathbb{R}^{n\times D}$ consists of n rows (data points) in $\mathbb{R}^D$, D dimensions (features or attributes).

A =

	t_1	t_2	t_3	t_4	...	t_D
u_1	*	*	*	*	...	*
u_2	*	*	*	*	...	*
u_3	*	*	*	*	...	*
u_4	*	*	*	*	...	*
...	...	...	...	...	...	...
u_n	*	*	*	*	...	*

What is the cost of computing $\mathbf{A}\mathbf{A}^\mathsf{T}$? $O(n^2D)$ A big deal ?

What if n = 0.6 million, D = 70 million? $n^2D = 2.5 \times 10^{19}$. Take a while!

Why do we care about $\mathbf{A}\mathbf{A}^\mathsf{T}$? Useful for a lot of things.

- $[\mathbf{A}\mathbf{A}^{\mathsf{T}}]_{1,2} = u_1^{\mathsf{T}} u_2 = \sum_{j=1}^{D} u_{1,j} u_{2,j}$
  is the inner product, an important measure of vector similarity.

- $[\mathbf{A}\mathbf{A}^{\mathsf{T}}]$ is fundamental in distance-based clustering, support vector machine (SVM) kernels, information retrieval, and more.

- An example. Ravichandran *et. al.* (ACL 2005) found the top similar nouns for each of $n = 655,495$ nouns, from a collection of D=70 million Web pages. Brute-force $O(n^2 D) \approx 10^{19}$ may take forever. They used random projections.

Other similarity or dissimilarity measures

- l_2 distance: $\|u_1 - u_2\|_2^2 = \sum_{j=1}^{D} (u_{1,j} - u_{2,j})^2$.
- l_1 distance: $\|u_1 - u_2\|_1 = \sum_{j=1}^{D} |u_{1,j} - u_{2,j}|$
- Multi-way inner product: $\sum_{j=1}^{D} u_{1,j} u_{2,j} u_{3,j}$

Let's Approximate AA^T and Other Distances

Many reasons why approximation is a good idea.

- Exact computation could be practically infeasible.
- Often do not need exact answers. Distances are used by other tasks such as clustering, retrieval, and ranking, which introduce errors.
- An approximate solution may help finding the exact solution more efficiently. Example: Databases query optimization

What Are Real Data Like?: Google Page Hits

	Query	Hits (Google)
	A	22,340,000,000
Function words	The	20,980,000,000
Frequent words	Country	2,290,000,000
	Knuth	5,530,000
Names	”John Nash”	1,090,000
	Kalevala	1,330,000
Rare words	Griseofulvin	423,000

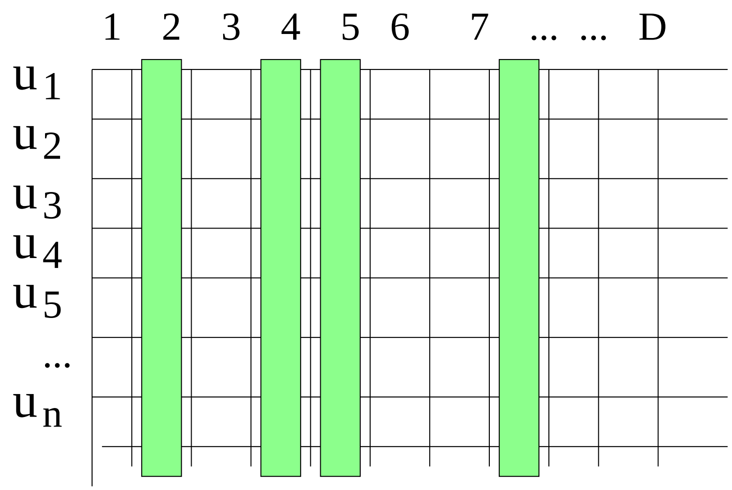
- Term-by-document matrix (n by D) is huge, and highly sparse
  - Approx $n = 10^7$ (interesting) words/items
  - Approx $D = 10^{10}$ Web pages (indexed)
- Lots of large counts (even for so-called rare words)

Outline of the Talk

- Two strategies (besides SVD) for dimension reduction:
  - Sampling
  - Sketching
- Normal random projections (for l_2).
- Cauchy random projections (for l_1). A case study on Microarray Data.
- Conditional Random Sampling (CRS), a new sketching algorithm for sparse data: Sampling + sketching
- Comparisons.

Strategies for Dimension Reduction: Sampling and Sketching

Sampling: Randomly pick k (out of D) columns from the data matrix $\mathbf{A}$.

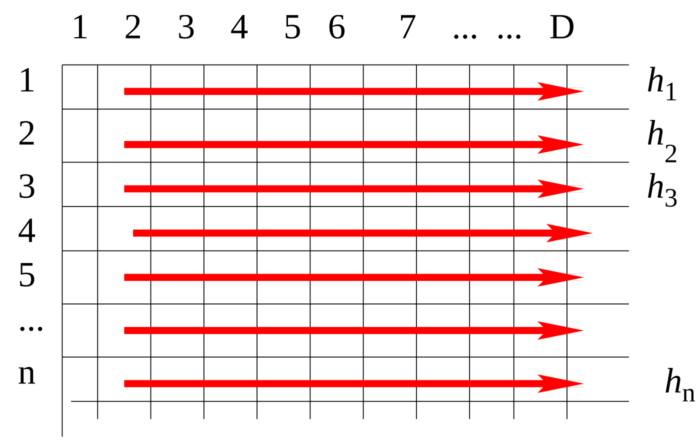


$$\mathbf{A} \in \mathbb{R}^{n\times D} \Longrightarrow \tilde{\mathbf{A}} \in \mathbb{R}^{n\times k}$$

$$\left(u_1^\mathsf{T} u_2 = \sum_{j=1}^{D} u_{1,j}u_{2,j}\right) \approx \left(\tilde{u}_1^\mathsf{T}\tilde{u}_2 = \sum_{j=1}^{k} \tilde{u}_{1,j}\tilde{u}_{2,j}\right) \times \frac{D}{k}$$

- Pros: Simple, popular, generalizes beyond approximating distances
- Cons: No accuracy guarantee. Large errors for worst case (heavy-tailed distributions). Mostly “zeros” in sparse data.

Sketching: Scan the data; compute specific summary statistics; repeat k times.



(Know everything about the margins: means, moments, # of non-zeros)

Two well-known examples of sketching algorithms

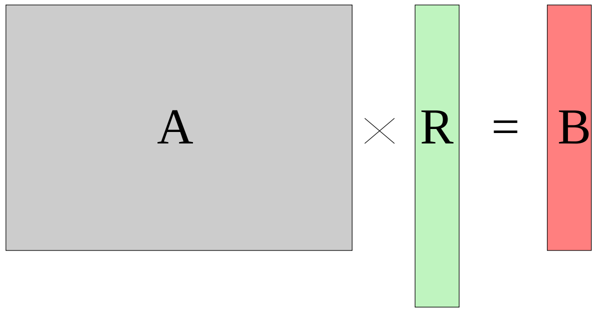
- Random Projections
- Broder's min-wise sketches.

A new algorithm

- Conditional Random Sampling (CRS): sampling + sketching, a hybrid method

Random Projections: A Brief Introduction

Let $\mathbf{B} = \mathbf{AR}$, $\mathbf{A} \in \mathbb{R}^{n\times D}$ is the original data matrix. $\mathbf{R} \in \mathbb{R}^{D\times k}$ is the random projection matrix. $\mathbf{B} \in \mathbb{R}^{n\times k}$ is the projected data.



Estimate original distances from $\mathbf{B}$. (Vempala 2004, Indyk FOCS00,01)

- For l_2 distance, use $\mathbf{R}$ with entries of i.i.d. Normal $N(0,1)$.
- For l_1 distance, use $\mathbf{R}$ with entries of i.i.d. Cauchy $C(0,1)$.

Computational cost: $O(nDk)$ for generating the sketch $\mathbf{B}$.
$O(n^2k)$ for computing all pairwise distances. $k \ll \min(n, D)$.
$O(nDk + n^2k)$ is a huge reduction, from $O(n^2D)$.

Normal Random Projections: l_2 Distance Preserving Properties

Notation: $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{AR}$, $\mathbf{R} = \{r_{ji}\} \in \mathbb{R}^{D\times k}$, r_{ji} i.i.d. $N(0,1)$.

- u_1, $u_2 \in \mathbb{R}^D$, first two rows in $\mathbf{A}$.
- v_1, $v_2 \in \mathbb{R}^k$, first two rows in $\mathbf{B}$.

$\mathbf{BB}^\mathsf{T} \approx \mathbf{AA}^\mathsf{T}$. In fact, $\mathsf{E}\left(\mathbf{BB}^\mathsf{T}\right) = \mathbf{AA}^\mathsf{T}$, in the expectations.

Projected data $(v_{1,i}, v_{2,i})$ (i = 1, 2, ..., k) are i.i.d. samples of a bivariate normal

$$\begin{bmatrix} v_{1,i} \\ v_{2,i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{k}\begin{bmatrix} m_1 & a \\ a & m_2 \end{bmatrix}\right).$$

Margins: $m_1 = \|u_1\|^2$, $m_2 = \|u_2\|^2$,

Inner Product: $a = u_1^\mathsf{T} u_2$,

l_2 distance: $d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a$.

$$\begin{bmatrix} v_{1,i} \\ v_{2,i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{k} \begin{bmatrix} m_1 & a \\ a & m_2 \end{bmatrix} \right).$$

Linear estimators (sample distances are unbiased for original distances)

$$\hat{a} = v_1^{\mathsf{T}} v_2 = \sum_{i=1}^{k} v_{1,i} v_{2,i}, \qquad \mathrm{E}(\hat{a}) = a$$

$$\hat{d} = \|v_1 - v_2\|^2 = \sum_{i=1}^{k} (v_{1,i} - v_{2,i})^2, \qquad \mathrm{E}(\hat{d}) = d$$

However

Marginal norms $m_1 = \|u_1\|^2$, $m_2 = \|u_2\|^2$ can be computed exactly

$\mathbf{B}\mathbf{B}^{\mathsf{T}} \approx \mathbf{A}\mathbf{A}^{\mathsf{T}}$, but at least we can make the diagonals exact (easily).

And off-diagonals can be improved (a little bit more work)

Margin-constrained Normal Random Projections

$$\begin{bmatrix} v_{1,i} \\ v_{2,i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{k} \begin{bmatrix} m_1 & a \\ a & m_2 \end{bmatrix} \right).$$

Linear estimator and its variance

$$\hat{a} = v_1^{\mathsf{T}} v_2, \qquad \text{Var}\left(\hat{a}\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right),$$

If the margins m_1 and m_2 are known; a maximum likelihood estimator, $\hat{a}_{MLE}$, is the solution to a cubic equation:

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

Consequently, an MLE for the distance $\hat{d}_{MLE} = m_1 + m_2 - 2\hat{a}_{MLE}$.

The (asymptotic) variance of the MLE:

$$\text{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{k}\frac{\left(m_1m_2 - a^2\right)^2}{m_1m_2 + a^2} \leq \text{Var}\left(\hat{a}\right) = \frac{1}{k}\left(m_1m_2 + a^2\right)$$

Substantial improvement when the data are strongly correlated ($a^2 \approx m_1m_2$).

But does not help when $a \approx 0$.

Next, Cauchy random projections for l_1 ...

Cauchy Random Projections for l_1

$\mathbf{B} = \mathbf{A}\mathbf{R}$, $\mathbf{R} = \{r_{ji}\} \in \mathbb{R}^{D\times k}$, r_{ji} i.i.d. $C(0,1)$.

- $u_1, u_2 \in \mathbb{R}^D$, first two rows in $\mathbf{A}$.
- $v_1, v_2 \in \mathbb{R}^k$, first two rows in $\mathbf{B}$.

The projected data are Cauchy distributed.

$$v_{1,i} - v_{2,i} = \sum_{j=1}^{D}(u_{1,j} - u_{2,j})r_{ji} \sim C\left(0, \sum_{j=1}^{D}|u_{1,j} - u_{2,j}| = d\right)$$

Linear estimator fails! (Charikar et. al, FOCS03, JACM05)

$\hat{d} = \frac{1}{k}\sum_{i=1}^{k}|v_{1,i} - v_{2,i}|$, does not work. $\mathrm{E}|v_{1,i} - v_{2,i}| = \infty$.

However, if only interested in approximating distances, then ...

Cauchy Random Projections: Our Results

- Many applications (e.g., clustering, SVM kernels) only need the distances, linear or nonlinear estimators do not really matter.
- Statistically, we need to estimate the scale parameter of Cauchy, from k i.i.d. samples of $C(0, d)$: $v_{1,i} - v_{2,i}$, $i = 1, 2, ..., k$.

Two nonlinear estimators:

- A new unbiased estimator is derived, which exhibits exponential tail bounds; (hence an analog of JL bound for l_1 exists, in a sense.)
- The MLE is even better. A highly accurate approximation is proposed for the distribution of the MLE, which does not have closed-from distribution.

Cauchy Random Projections: The Procedure

Estimation Method The original l_1 distance $d = |u_1 - u_2|$ is estimated from the projected data, $v_{1,i} - v_{2,i}$, $i = 1, 2, ..., k$, by

$$\hat{d}_1 = \hat{d}\left(1 - \frac{1}{k}\right),$$

where $\hat{d}$ solves the nonlinear MLE equation

$$-\frac{k}{d} + \sum_{i=1}^{k} \frac{2d}{(v_{1,i} - v_{2,i})^2 + d^2} = 0,$$

by iterative methods, starting with the following initial guess

$$\hat{d}_{gm} = \cos^k\left(\frac{\pi}{2k}\right)\prod_{i=1}^{k}|v_{1,i} - v_{2,i}|^{\frac{1}{k}}.$$

Cauchy Random Projections: An Unbiased Estimator

$$\hat{d}_{gm} = \cos^k\left(\frac{\pi}{2k}\right)\prod_{i=1}^{k}|v_{1,i} - v_{2,i}|^{1/k}, \quad k > 1$$

is unbiased, with the variance (valid when $k > 2$)

$$\text{Var}\left(\hat{d}_{gm}\right) = \frac{\pi^2}{4}\frac{d^2}{k} + O\left(\frac{1}{k^2}\right).$$

The $\frac{\pi^2}{4k} \approx \frac{2.5}{k}$ implies that $\hat{d}_{gm}$ is 80% efficient, as the MLE has variance in terms of $\frac{2.0}{k}$.

Cauchy Random Projections: Tail Bounds

If we restrict that $0 \leq \epsilon < 1$, the following exponential tail bounds hold:

$$\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{8(1+\epsilon)}\right)$$

$$\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{20}\right), \qquad k > \frac{\pi^2}{4\epsilon}$$

An analog of the JL bound follows by restricting $\mathbf{Pr}\left(|\hat{d}_{gm} - d| \geq \epsilon d\right) \leq \xi/\nu$ with $\nu = \frac{n^2}{2}$, (e.g.,) $\xi = 0.05$.

Comments

- These bounds are not tight. (we have more tight bounds)
- Without the restriction $\epsilon < 1$, the exponential bounds do not exist.
- We prefer the exponential bounds of the MLE.

Cauchy Random Projections: MLE

The maximum likelihood estimator $\hat{d}$ is the solution to

$$-\frac{k}{d} + \sum_{i=1}^{k} \frac{2d}{(v_{1,i} - v_{2,i})^2 + d^2} = 0.$$

We suggest the bias-corrected version based on (Bartlett, Biometrika 53):

$$\hat{d}_1 = \hat{d}\left(1 - \frac{1}{k}\right),$$

What about the distribution?

- Need the distribution of $\hat{d}_1$ to select sample size k.
- The distribution of $\hat{d}_1$ can not be characterized exactly,
- We can at least study the asymptotic moments.

Cauchy Random Projections: MLE Moments

The first four (asymptotic) moments of the $\hat{d}_1$ are

$$\mathsf{E}\left(\hat{d}_1 - d\right) = O\left(\frac{1}{k^2}\right)$$

$$\mathsf{Var}\left(\hat{d}_1\right) = \frac{2d^2}{k} + \frac{3d^2}{k^2} + O\left(\frac{1}{k^3}\right)$$

$$\mathsf{E}\left(\hat{d}_1 - \mathsf{E}(\hat{d}_1)\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right)$$

$$\mathsf{E}\left(\hat{d}_1 - \mathsf{E}(\hat{d}_1)\right)^4 = \frac{12d^4}{k^2} + \frac{186d^4}{k^3} + O\left(\frac{1}{k^4}\right)$$

by carrying out the horrible algebra in (Shenton, JORSS 63).

Magic: They match the first four moments of an inverse Gaussian distribution, which has the same support as $\hat{d}_1$, $[0, \infty]$.

Cauchy Random Projections: Inverse Gaussian Approximation

Assume $\hat{d}_1 \sim IG(\alpha, \beta)$, with $\alpha = \frac{1}{\frac{2}{k}+\frac{3}{k^2}}, \qquad \beta = \frac{2d}{k} + \frac{3d}{k^2}$.

The moments

$$\mathrm{E}\left(\hat{d}_1\right) = d, \qquad \mathrm{Var}\left(\hat{d}_1\right) = \frac{2d^2}{k} + \frac{3d^2}{k^2}$$

$$\mathrm{E}\left(\hat{d}_1 - \mathrm{E}(\hat{d}_1)\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right)$$

$$\mathrm{E}\left(\hat{d}_1 - \mathrm{E}(\hat{d}_1)\right)^4 = \frac{12d^4}{k^2} + \frac{156d^4}{k^3} + O\left(\frac{1}{k^4}\right)$$

The exact (asymptotic) fourth moment of $\hat{d}_1 = \frac{12d^4}{k^2} + \frac{186d^4}{k^3} + O\left(\frac{1}{k^4}\right)$

The density

$$\mathbf{Pr}(\hat{d}_1 = y) = \sqrt{\frac{\alpha d}{2\pi}} y^{-\frac{3}{2}} \exp\left(-\frac{(y-d)^2}{2y\beta}\right),$$

The Chernoff bounds

$$\mathbf{Pr}\left(\hat{d}_1 \geq (1+\epsilon)d\right) \leq \exp\left(-\frac{\alpha\epsilon^2}{2(1+\epsilon)}\right), \qquad \epsilon \geq 0$$

$$\mathbf{Pr}\left(\hat{d}_1 \leq (1-\epsilon)d\right) \leq \exp\left(-\frac{\alpha\epsilon^2}{2(1-\epsilon)}\right), \qquad 0 \leq \epsilon < 1.$$

A symmetric bound

$$\mathbf{Pr}\left(|\hat{d}_1 - d| \geq \epsilon d\right) \leq 2\exp\left(-\frac{\alpha\epsilon^2}{2(1+\epsilon)}\right), \qquad 0 \leq \epsilon < 1$$

A JL-type of Bound (Derived by approximation, verified by simulations)

A JL-type of bound follows by letting $\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right) \leq \xi/\nu$,

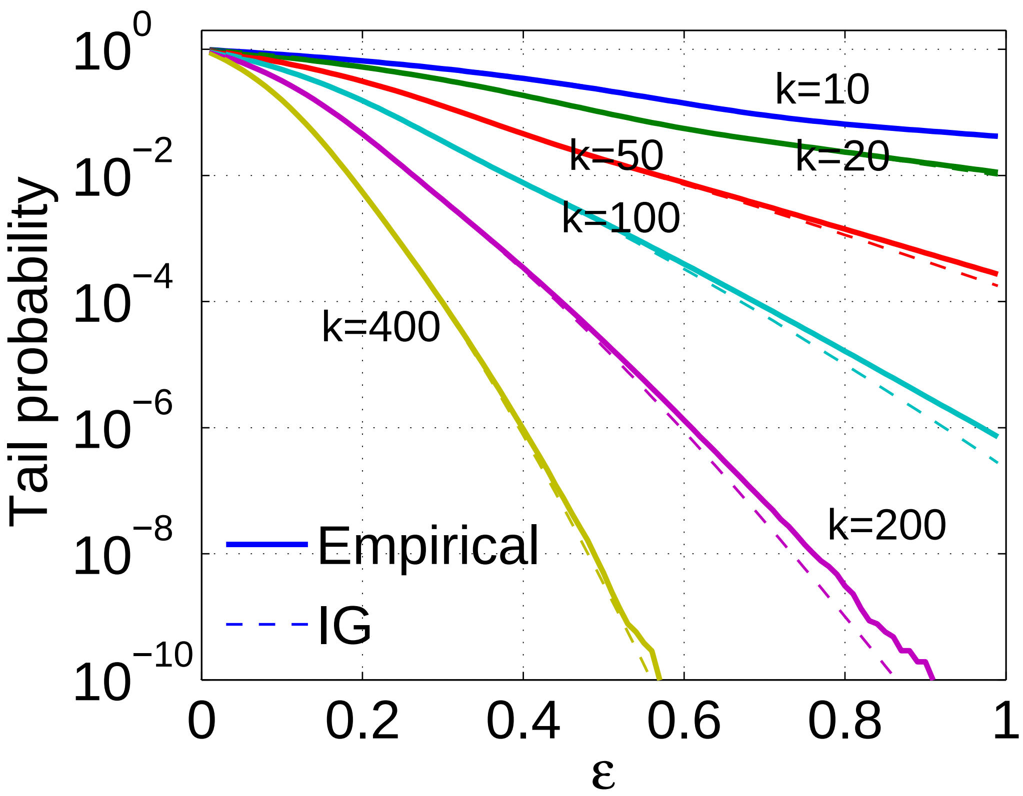
$$k \geq \frac{4.4\left(\log 2\nu - \log \xi\right)}{\epsilon^2/(1+\epsilon)}.$$

This holds at least for $\xi/\nu \geq 10^{-10}$, verified by simulations.

(Why the 95% normal quantile = 1.645?)

Cauchy Random Projections: Simulations

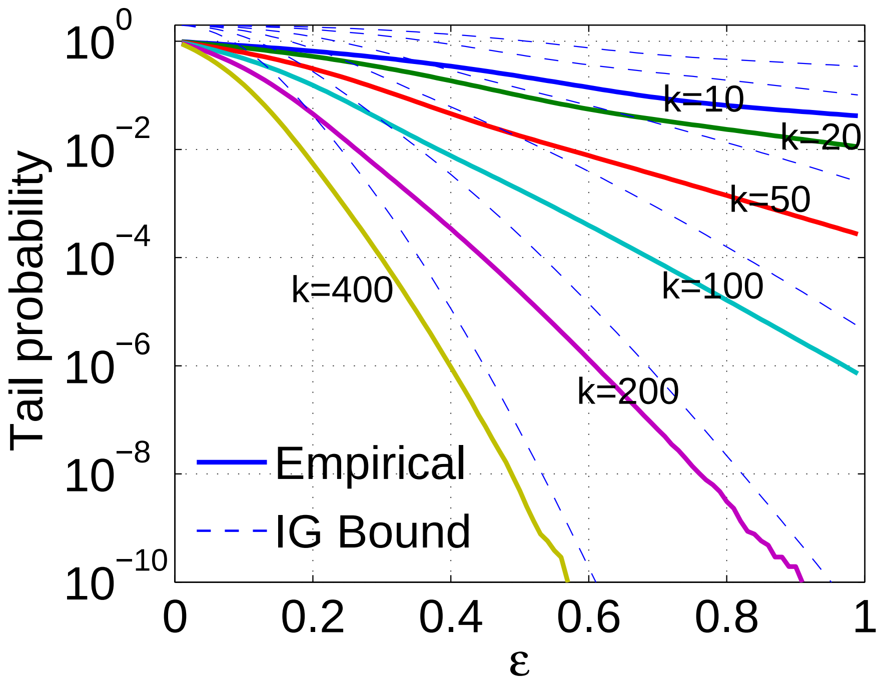
Tail probability $\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right)$



The inverse Gaussian approximation is remarkably accurate.

Tail bound

$$\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right) \leq \exp\left(-\frac{\alpha\epsilon^2}{2(1+\epsilon)}\right) + \exp\left(-\frac{\alpha\epsilon^2}{2(1-\epsilon)}\right), \qquad 0 \leq \epsilon < 1.$$



The inverse Gaussian Chernoff bound is reliable at least for $\xi/\nu \geq 10^{-10}$.

A Case Study on Microarray Data

Harvard Dataset (PNAS 2001, thank Wing H. Wong): 176 specimen, 3 classes, 12600 genes.

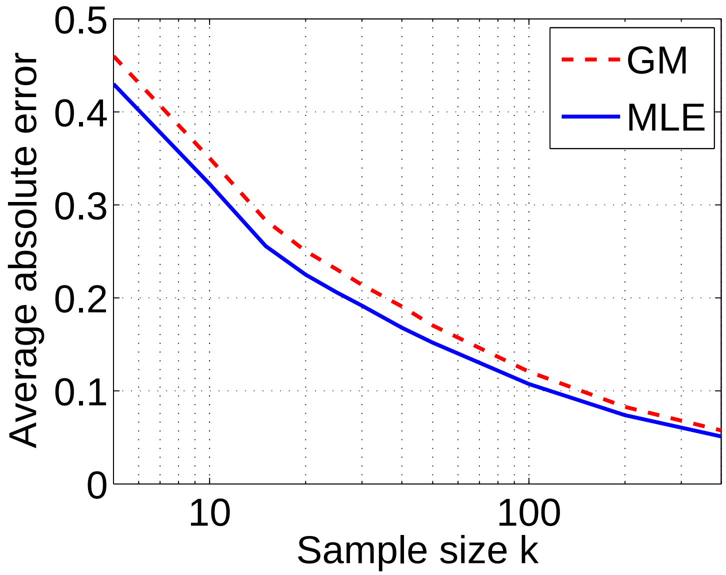
Only 2 (out of 176) specimen were misclassified, by a 5-nearest neighbor classifer using l_1 distances in 12600 dimensions.

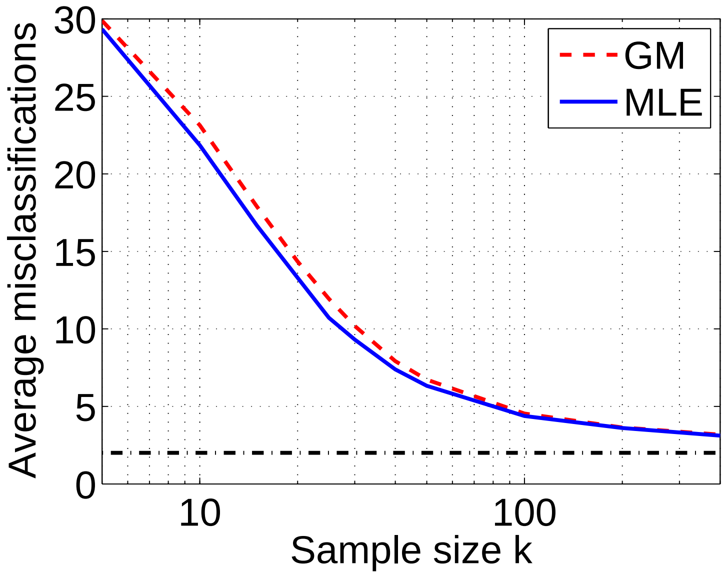
Using Cauchy random projections and both nonlinear estimators, the dimension can be reduced from 12600 to 100, with little loss in accuracy.

Two error measures:

- Median (among $176 \times 175/2 = 15488$ pairs) absolute errors of estimated l_1 distances, normlized by original median l_1 distance.
- Number of misclassifications.

Left: Distance errors   Right: Misclassifications





- When $k = 100$, relative absolute distance error about 10%.
- When $k = 100$, number of misclassifications < 5.
- MLE is about 10% better than GM (unbiased estimator) in distance errors, as expected.
- MLE is about $5\% - 10\%$ better than GM in misclassifications.

Summary for Cauchy Random Projections

- Linear projections + linear estimators do not work well (impossibility results).
- Linear projections + nonlinear estimators are available and suffice for many applications (e.g., clustering, SVM kernels, information retrieval).
- Analog of JL bound in l_1 exists (in a sense), proved using an unbiased nonlinear estimator
- The MLE is even better. Highly accurate and convenient closed-form approximations of the tail bounds are practically useful.

So far so good...

Limitations of Random Projections

- Designed for specific summary statistics (l_1 or l_2)
- Limited to two-way (pairwise) distances

What about sampling?

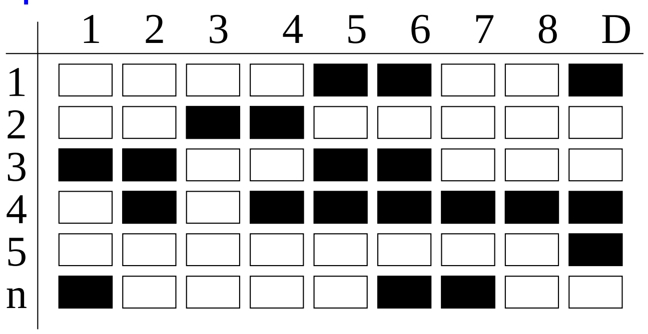
- Suitable for any norm and multi-way
- Most samples are zeros, in sparse data
- Possibly large errors in heavy-tailed data

Conditional Random Sampling (CRS): A sketch-based sampling algorithm.

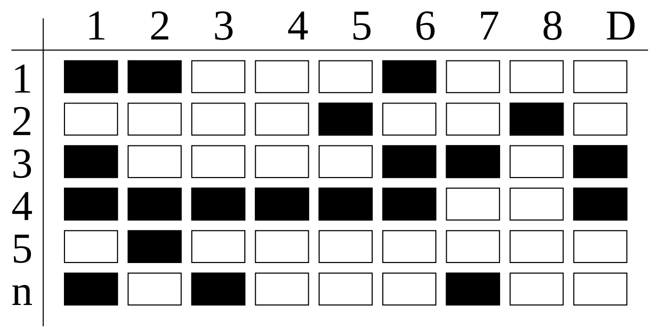
Directly exploit data sparsity

Conditional Random Sampling (CRS): A Global View

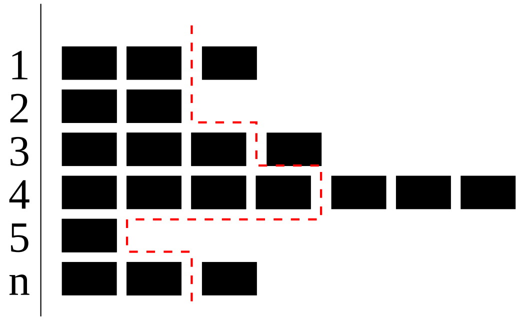
Sparse Data Matrix



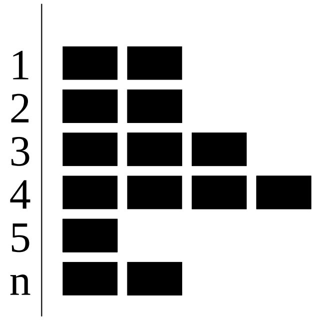
Random Permutation on Columns



Postings (Non-zero Entries)



Sketches (Front of Postings)



Conditional Random Sampling (CRS): An Example

Random Sampling on Data Matrix $\mathbf{A}$: If columns are random, first $D_s = 10$ columns constitute a random sample.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
u_1	0	3	0	2	0	1	0	0	1	2	1	0	1	0	2	0
u_2	1	4	0	0	1	2	0	1	0	0	3	0	0	2	1	1

Postings $\mathbf{P}$: Only store non-zeros, "ID (Value)," sorted ascending by the IDs.

P_1: 2 (3) 4 (2) 6 (1) 9 (1) 10 (2) 11 (1) 13 (1) 15 (2)
P_2: 1 (1) 2 (4) 5 (1) 6 (2) 8 (1) 11 (3) 14 (2) 15 (1) 16(1)

Sketches $\mathbf{K}$: A sketch, K_i, of postings P_i, is the first k_i entries of P_i. Suppose $k_1 = 5, k_2 = 6$.

K_1: 2 (3) 4 (2) 6 (1) 9 (1) 10 (2)
K_2: 1 (1) 2 (4) 5 (1) 6 (2) 8 (1) 11 (3)

What if remove the entry 11(3)?... We get random samples.

Exclude all elements of sketches whose IDs are larger than

$$D_s = \min\left(\max(\mathsf{ID}(\mathsf{K}_1)), \max(\mathsf{ID}(\mathsf{K}_2))\right)$$
$$= \min(10, 11) = 10,$$

Obtain exactly the same samples as if directly sampled the first D_s columns.

This converts sketches into random samples by conditioning on D_s, different pairwise (or group-wise), and not known beforehand.

For example, when estimating pairwise distances for all n data points, we will have $\frac{n(n-1)}{2}$ different values of D_s.

Sketch size k_i can be small, but the effective sample D_s could be very large. The more sparse, the better.

Conditional Random Sampling (CRS): Procedure

Our algorithm consists of the following steps:

- A random permutation on the data column IDs to ensure randomness.
- Construct sketches for all data points, i.e. finding k_i entries with the smallest IDs after permutation. Need a linear scan (hence called sketches).
- Construct conditional random samples from sketches online pairwise (or group-wise). Compute D_s. Estimate the original space by scaling ($\frac{D}{D_s}$) any sample distances. (*We can do better than that...*)

Take advantage of the margins for sharper estimates (MLE):

- In 0/1 data, numbers of non-zeros (f_i, document frequency) are known. The MLE amounts to estimating two-way contingency tables with margin constraints. The solution is a cubic equation.
- In general real-valued data, f_i, marginal norms, marginal means are known. The MLE amounts to a cubic equation (assuming normality, works well).

Variances: CRS V.S. Random Projections (RP)

$u_1, u_2 \in \mathbf{R}^D$, Inner Product $a = u_1^\mathsf{T} u_2$, $\hat{a}_{CRS}$ v.s. $\hat{a}_{RP}$ (not using margins).

Var $\left(\hat{a}_{CRS}\right) = \frac{\max(f_1,f_2)}{D}\frac{1}{k}\left(D\sum_{j=1}^D u_{1,j}^2 u_{2,j}^2 - a^2\right)$

Var $\left(\hat{a}_{RP}\right) = \frac{1}{k}\left(\sum_{j=1}^D u_{1,j}^2 \sum_{j=1}^D u_{2,j}^2 + a^2\right)$

Sparsity: f_1 and f_2 are numbers of non-zeros. Often $\frac{\max(f_1,f_2)}{D} < 1\%$

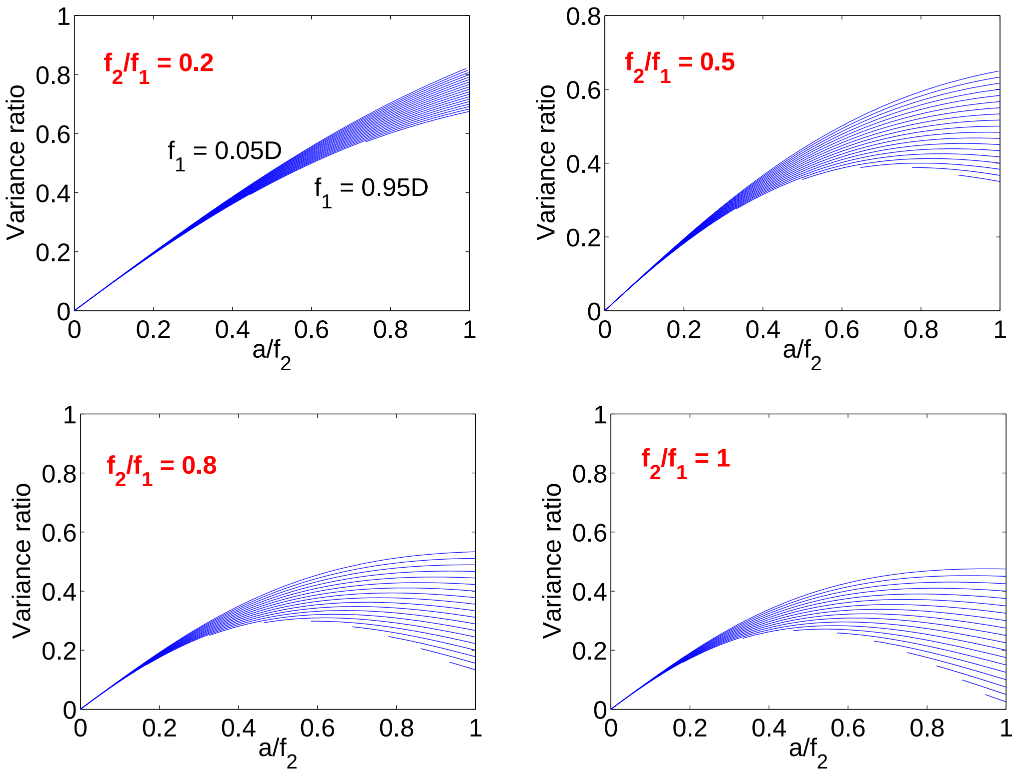
$D\sum_{j=1}^D u_{1,j}^2 u_{2,j}^2 > \sum_{j=1}^D u_{1,j}^2 \sum_{j=1}^D u_{2,j}^2$ usually, $\gg$ in heavy-tailed data.

When u_1 and u_2 are independent, by law of large numbers
$D\sum_{j=1}^D u_{1,j}^2 u_{2,j}^2 \approx \sum_{j=1}^D u_{1,j}^2 \sum_{j=1}^D u_{2,j}^2$,
then Var $\left(\hat{a}_{CRS}\right) <$ Var $\left(\hat{a}_{RP}\right)$, even ignoring sparsity.

In boolean (0/1) data ...

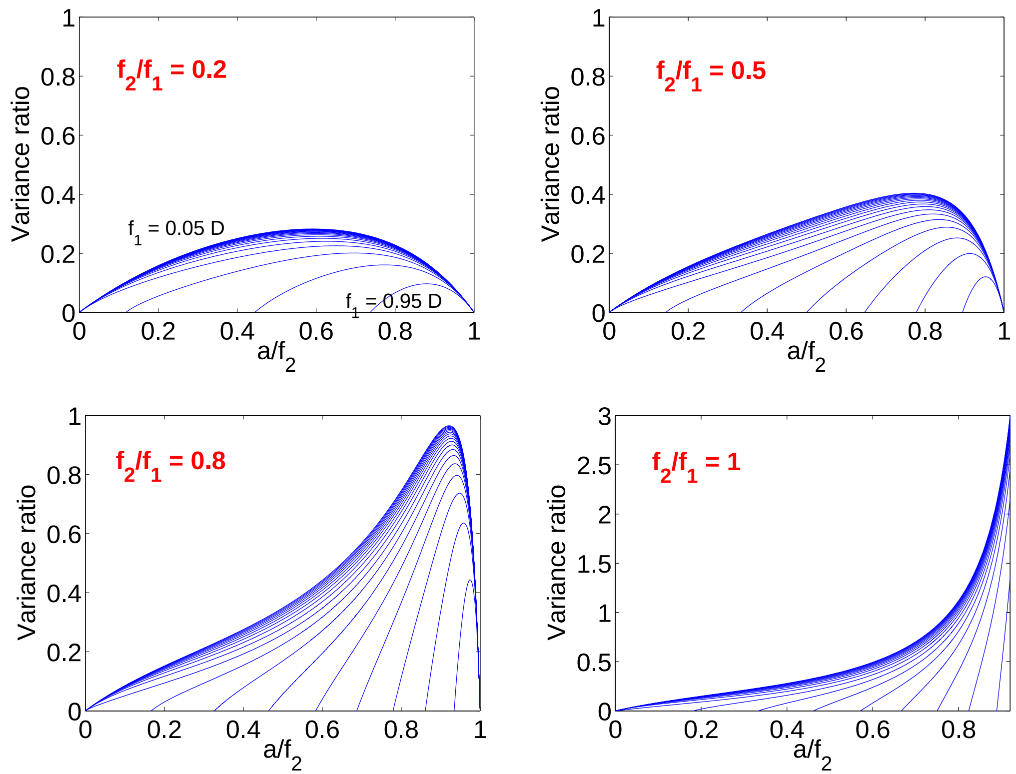
CRS V.S. RP in Boolean Data

CRS are always better in boolean data. The ratio $\frac{\text{Var(CRS)}}{\text{Var(RP)}}$ is always < 1, when both do not use marginal information.



f_1 and f_2 are the numbers of non-zeros in u_1 and u_2.

When both use margins, the ratio $\frac{\text{Var(CRS)}}{\text{Var(RP)}}$ is < 1 almost always, unless u_1 and u_2 are almost identical.



Empirical Evaluations of CRS and RP

Data (Each has total $\frac{n(n-1)}{2}$ pairs of distances)

	n	D	Sparsity	Kurtosis	Skewness
NSF	100	5298	1.09%	349.8	16.3
NEWSGROUP	100	5000	1.01%	352.9	16.5
COREL	80	4096	4.82%	765.9	24.7
MSN (original)	100	65536	3.65%	4161.5	49.6
MSN (square root)	100	65536	3.65%	175.3	10.7
MSN (logarithmic)	100	65536	3.65%	111.8	9.5

- NEWSGROUP and NSF (thank Bingham and Dhillon): document distance
- COREL: Image histogram distance
- MSN : Word distance,
- Median sample kurtosis and skewness, (heavy-tailed, highly-skewed)

Variable sketch size for CRS

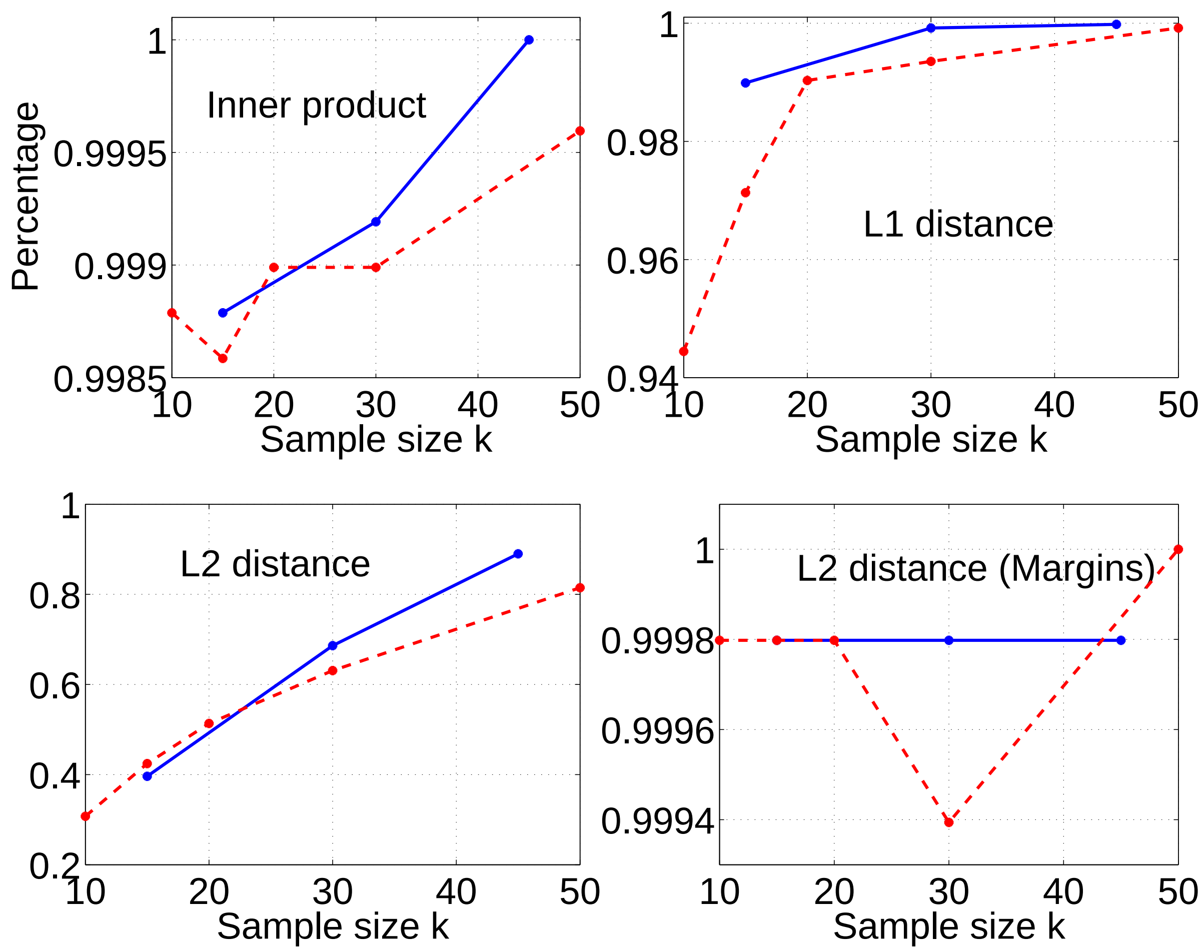
We could adjust sketch sizes according to data sparsity. Sample more from the more frequent ones.

Evaluation metric

Among the $\frac{n(n-1)}{2}$ pairs, the percentage for which CRS does better than random projections. Want > 0.5

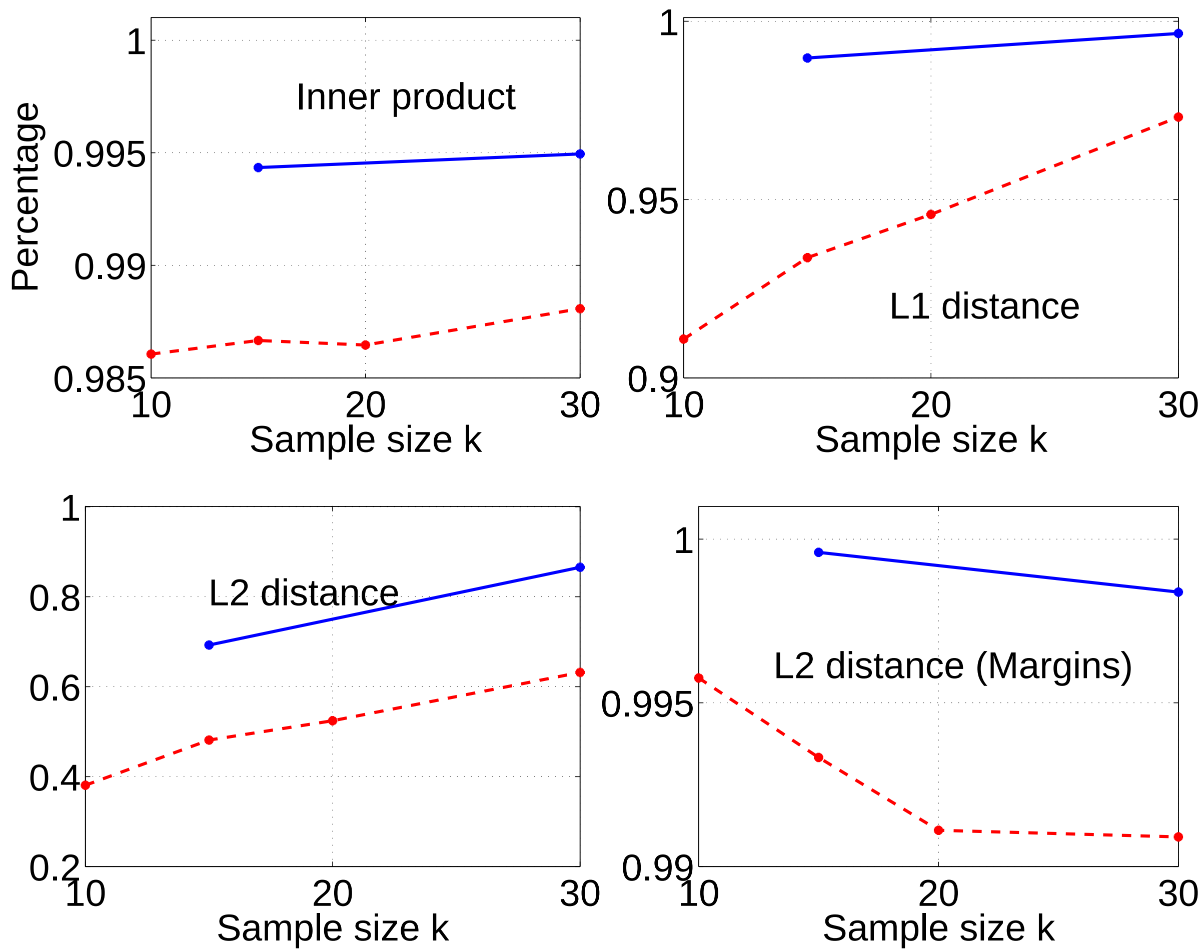
Results...

NSF Data: Conditional Random Sampling (CRS) is overwhelmingly better than Random Projections (RP).

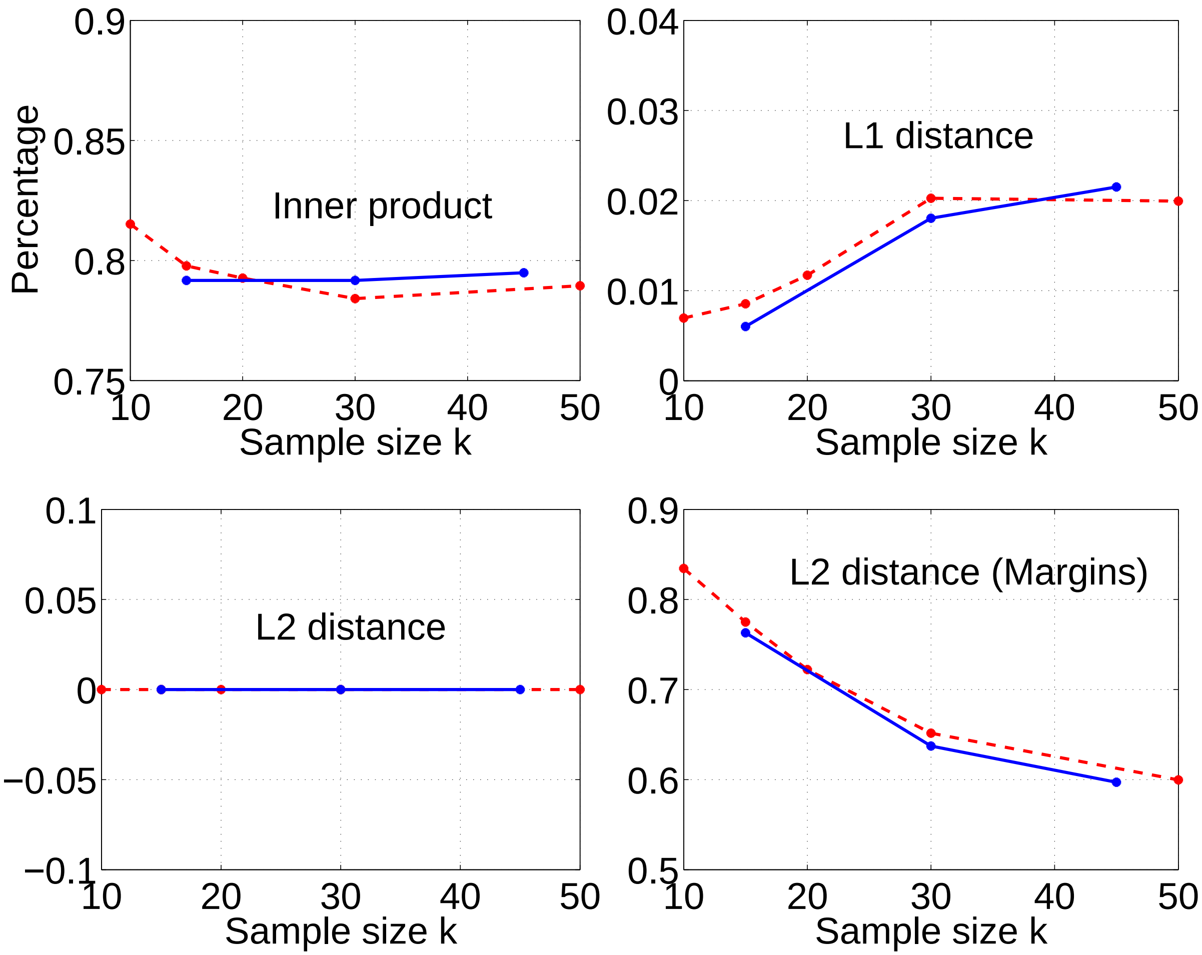


Dashed: Fixed sample size, Solid: Variable sketch size

NEWSGROUP Data: CRS is overwhelmingly better than RP.

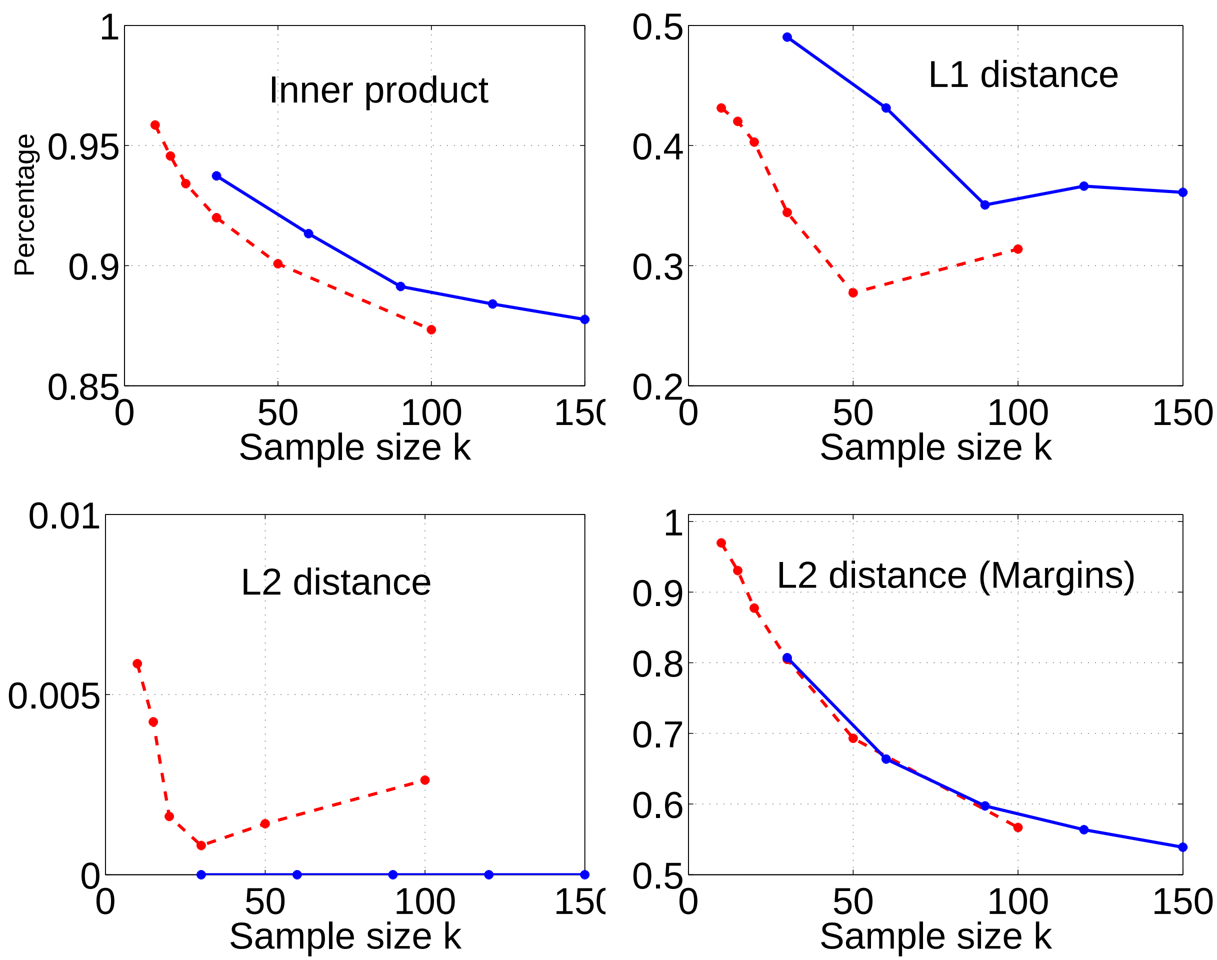


COREL Image Data: CRS are still better than RP for inner product and l_2 distance (using margins)

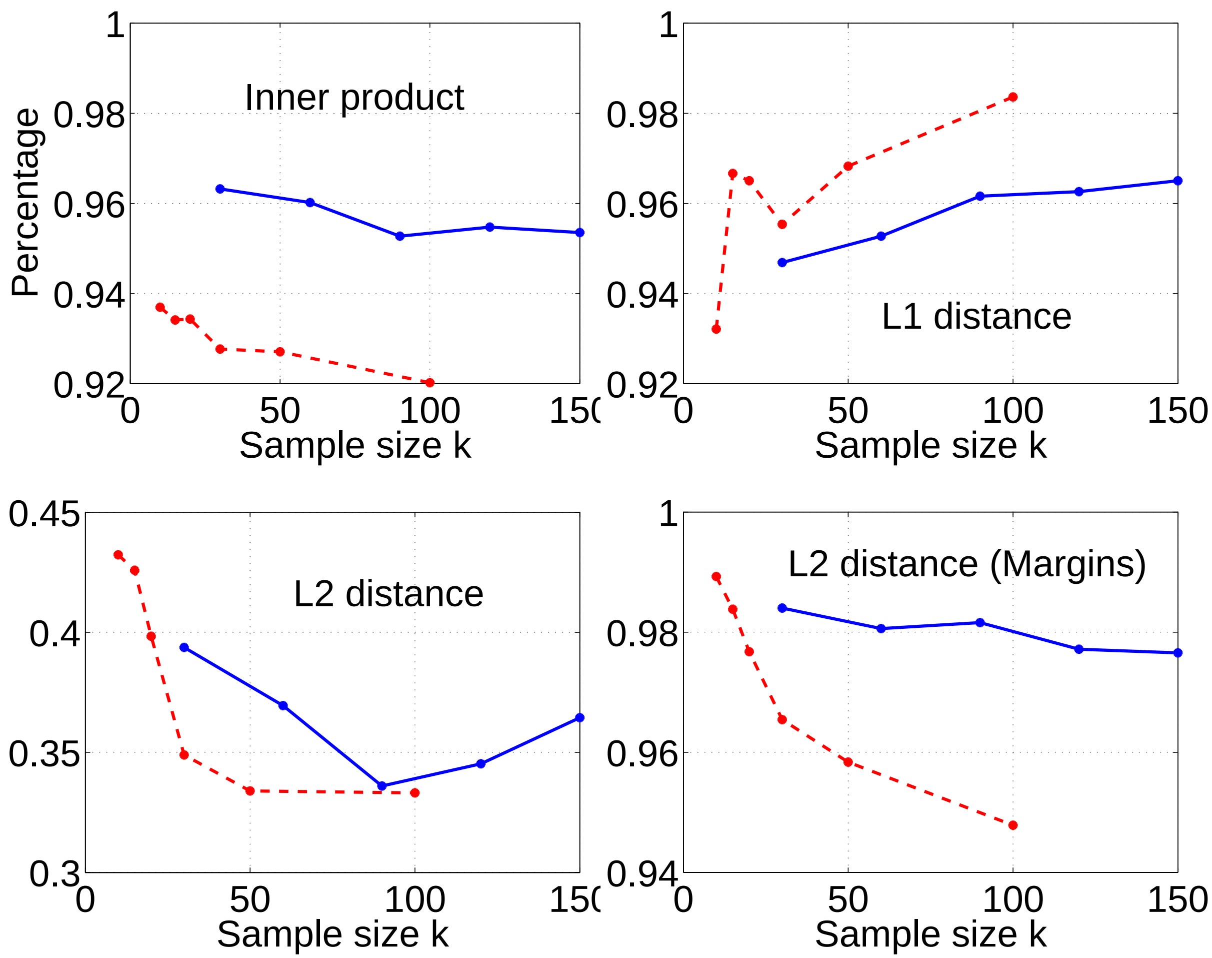


	n	D	Sparsity	Kurtosis	Skewness
NSF	100	5298	1.09%	349.8	16.3
NEWSGROUP	100	5000	1.01%	352.9	16.5
COREL	80	4096	4.82%	765.9	24.7
MSN (original)	100	65536	3.65%	4161.5	49.6
MSN (square root)	100	65536	3.65%	175.3	10.7
MSN (logarithmic)	100	65536	3.65%	111.8	9.5

MSN Data (original): CRS do better than RP in inner product and l_2 distance (using margins)



MSN Data (square root): After transformation (as in practice), CRS do better than RP in inner product, l_1 and l_2 (using margins)



Summary of the Empirical Comparisons

Conditional Random Sampling (CRS) v.s. Random Projections (RP)

- CRS are particularly well-suited for inner products.
- CRS are often comparable to Cauchy random projections for l_1 distances.
- Using the margins, CRS are also effectively for l_2 distances.
- Can adjust the sketch size according to the data sparsity, which in general improves the overall performance.
  - Using a fixed sketch size, then the less freqent (but often more interesting) items are emphasized.

Conclusions

- Too much data (although never enough)
  - Compact data representations
  - Accurate approximation algorithms (estimators)
- Dimension Reduction Techniques (in addition to SVD)
  - Random sampling
  - Sketching (e.g., normal and Cauchy random projections)
  - Conditional Random Sampling (sampling + sketching)
- Improve normal random projection (for l_2) using margins by nonlinear MLE.
- Propose nonlinear estimators for Cauchy random projections for l_1.
- Conditional Random Sampling (CRS), for sparse data and 0/1 data
  - Flexible (can adjust sample size according to sparsity)
  - Good for estimating inner products
  - Easy to take advantage of margins.

References

Ping Li, Trevor Hastie, and Kenneth Church,

Practical Procedurs for Dimension Reduction in l_1,

Tech. report, Stanford Statistics, 2006

http://www.stanford.edu/˜pingli98/publications/cauchy_rp_tr.pdf

Ping Li, Kenneth Church, and Trevor Hastie,

Conditional Random Sampling: A Sketch-based Sampling Technique for Sparse Data,

Tech. report, Stanford Statistics, 2006

http://www.stanford.edu/˜pingli98/publications/CRS_tr.pdf