Susbsampling, Spectral Methods and Semidefinite Programming

Alexandre d'Aspremont

Princeton University

Joint work with Noureddine El Karoui, U.C. Berkeley

Support from NSF, DHS and Google.

Introduction

Spectral methods.

- Computing leading eigenvectors using iterative methods costs

$$O\left(\frac{n^2 \log(n/\delta^2)}{\sqrt{\epsilon}}\right)$$

  with probability of failure δ, using the Lanczos method with random start.
- Subsampling methods approximate leading eigenvectors at cost below $O^*(n^2)$.

Averaging.

- **Monte-Carlo:** Averaging approximate eigenvectors to improve precision. . . Run many, cheap independent subsampling approximations.
- Basic CPUs are cheap (Amazon EC2: \$0.10/hour, Google, Yahoo, etc.).
- Clock speed, bandwidth and low latency are very expensive (physical limitations).

When, why does it work?

Introduction

Subsampling procedure from Achlioptas and McSherry (2007).

Given $p \in [0,1]$ and a symmetric matrix $A \in \mathbf{S}_n$, define

$$S_{ij} = \begin{cases} A_{ij}/p & \text{with probability p} \\ 0 & \text{otherwise.} \end{cases}$$

By construction

- $\mathbf{E}[S] = A$.
- S has independent coefficients.
- S is sparse: it has pn^2 nonzero entries on average.

Because of independence, the impact of subsampling on the spectrum is both small and isotropic. . .

Introduction

A few related references.

- Early subsampling results by Groh et al. (1991) and Papadimitriou et al. (2000) who described algorithms based on subsampling and random projections.
- Explicit error estimates for columnwise and elementwise sampling strategies followed in Frieze et al. (2004), Drineas et al. (2006), Achlioptas and McSherry (2007). Survey in Kannan and Vempala (2009).
- More recently, Recht et al. (2007), Candes and Recht (2008), Candes and Tao (2009), Keshavan et al. (2009) focused on low-rank matrix reconstruction.
- Stability results in clustering and ranking by Ng et al. (2001) and Huang et al. (2008) for example.
- Averaging used in Coifman et al. (2008) for Cryo-EM imaging.
- In optimization: Juditsky et al. (2009) and Arora and Kale (2007).

Outline

- Introduction
- **Subsampling & averaging**
- Applications & Numerical Results

Elementwise subsampling

Subsampling: given $A \in \mathbf{S}_n$ and $p \in [0,1]$, define

$$S_{ij} = \begin{cases} A_{ij}/p & \text{with probability p} \\ 0 & \text{otherwise.} \end{cases}$$

(Achlioptas and McSherry, 2007, Th. 1.4)

$$\|A - S\|_2 \leq 4\sqrt{\frac{n}{p}}\ \max_{ij} |A_{ij}|,$$

holds with high probability for n large enough.

- What is the lowest reasonable p here?
- How does $\max_{ij} |A_{ij}| \sqrt{n/p}$ behave when $n \to \infty$?

Matrix Completion

Let's write

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$

- Recent results in Recht et al. (2007), Candes and Recht (2008), Candes and Tao (2009), Keshavan et al. (2009). In particular, when

$$\|u_i\|_\infty^2 \leq \frac{\mu}{n}$$

with $\mu = O(1)$, and we randomly sample more than $O(n \log^4 n)$ coefficients, then

$$\begin{array}{rll} A = & \operatorname{argmin} & \|X\|_* \\ & \text{s.t.} & X_{ij} = S_{ij}, \quad \text{when } S_{ij} \neq 0 \end{array}$$

with high probability.

- All the information we need on A is contained in S (which only has $O(n \log^4 n)$ nonzero coefficients), but is **expensive to extract**.

Elementwise subsampling

With

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$

- Here, we measure the **incoherence** of the spectrum of A by

$$\mu(A, \alpha) = \sum_{i=1}^{n} |\lambda_i| n^{\alpha_i} \|u_i\|_\infty^2$$

  for some $\alpha \in [0, 1]^n$.

- Because $\mathbf{NumCard}(x) = \frac{\|x\|_2^2}{\|x\|_\infty^2}$, the norm $\|u_i\|_\infty$ is a proxy for sparsity.
- When the eigenvector coefficients are uniformly distributed, we have

$$\|u_i\|_\infty \sim n^{-1/2}, \quad \text{when } n \to \infty$$

  we can take $\alpha_i = 1$ and $\mu(A, \alpha)$ will be bounded if $\|X\|_1$ remains bounded.

Error bounds

Theorem 1

Error bound. *Suppose there is a vector* $\alpha \in [0,1]^n$ *for which*

$$\mu(M,\alpha) \leq \mu \quad \textit{and} \quad \mathbf{Card}(u_i) \leq \frac{\kappa}{2} n^{\alpha_i}, \quad i = 1, \ldots, n$$

as $n \to \infty$, *where* μ *and* κ *are absolute constants. If*

$$\liminf_{n\to\infty} \frac{p\, n^{\alpha_{\min}}}{\log n^{\alpha_{\min}}} = \infty,$$

then

$$\|A - S\|_2 \leq \frac{\kappa\mu}{\sqrt{p n^{\alpha_{\min}}}}$$

almost surely (asymptotically), where $\alpha_{\min} = \min_{i=1,\ldots,n} \alpha_i$.

Error bounds

Proof. (Sketch) Using e.g. (Horn and Johnson, 1991, Th. 5.5.19)

$$\begin{aligned} \|A - S\|_2 &= \sqrt{\frac{1-p}{p}} \left\| \sum_{i=1}^{n} \lambda_i C \circ (u_i u_i^T) \right\|_2 \\ &\leq \sqrt{\frac{1-p}{p}} \sum_{i=1}^{n} |\lambda_i| n^{\alpha_i/2} \|u_i\|_\infty^2 \left\| \frac{C_{\alpha_i}}{n^{\alpha_i/2}} \right\|_2 \end{aligned}$$

where C is i.i.d. Bernoulli with

$$C_{ij} = \begin{cases} \sqrt{(1-p)/p} & \text{with probability } p \\ -\sqrt{p/(1-p)} & \text{otherwise.} \end{cases}$$

with C_{α_i} is a sparse submatrix of C. We first control

$$\left\| \frac{C_{\alpha_i}}{n^{\alpha_i/2}} \right\|_2$$

almost surely (asymptotically), then the max. of these quantities.

Error bounds

A few facts. . .

- The subsampled matrix must have more than

$$O(n^{(2-\alpha_{\min})} \log n)$$

  nonzero coefficients to keep the error small ($\alpha_{\min} = 1$ for incoherent spectrum).

- The **sparsest eigenvector** determines error size through $\alpha_{\min}$.

- The smallest submatrix we can form using this result must have more than $O(n \log n)$ nonzero coefficients.

- **Tightness:** Because $\|C/n^{1/2}\|_2$ blows up if $p \leq (\log n)^{1-\delta}/n$, the best we can hope for here is $O(n \log n)$ nonzero coefficients (Coupon collector effect).

Averaging

- (Kato, 1995, Theorem II.3.9) shows that if

$$\|A - S\|_2 \leq (\lambda_1 - \lambda_2)/2$$

the subsampled matrix can be seen as a small **perturbation** of the original one.

- If the matrix satisfies

$$\frac{\kappa\mu}{\sqrt{pn^{\alpha_{\min}}}} \leq (\lambda_1 - \lambda_2)/2,$$

we get the following expansion for the leading eigenvector v of the subsampled matrix compared to the true vector u

$$v = u - REu + R(E - u^T Eu\mathbf{I})RE + o_P(\|E\|_2^2)$$

where $E = A - S$ and R is the reduced resolvent of A, written

$$R = \sum_{j \neq 1} \frac{1}{\lambda_j - \lambda_1} u_j u_j^T.$$

Averaging

When $\|A - S\|_2 \leq (\lambda_1 - \lambda_2)/2$, the first order term has zero mean

$$\mathbf{E}[R(A - S)u] = 0.$$

Averaging eigenvectors over many subsampled matrices in the perturbative regime means that the residual error will be of order $\|A - S\|_2^2$.

The variance of the first order term is given by

$$\mathbf{E}[\|REu\|_2^2] \leq \frac{1}{(1 - \lambda_2/\lambda_1)^2} \|u_1\|_\infty^2 \frac{\mathbf{NumRank}(A)}{p}$$

so the quality of the eigenvector approximation is a function of

- The **spectral** gap λ_2/λ_1.
- The **numerical sparsity** of u_1, measured by $\|u_1\|_\infty$.
- The **numerical rank** of the matrix A.
- The **sampling probability** p.

Averaging: residual

The first-order term vanishes after averaging. We also control the residual. . .

Theorem 2

Second Order Accuracy. *Suppose the assumptions of the previous result are satisfied. Call u_1 the leading eigenvector of A and v_1 the leading eigenvector of S (such that $u_1^T v_1 \geq 0$), then*

$$\mathbf{E}[\|u_1 - v_1\|_2] = O\left(\frac{1}{(\lambda_1 - \lambda_2)^2} \frac{\mu^2}{pn^{\alpha_{\min}}}\right)$$

Outline

- Introduction
- Subsampling
- **Applications & Numerical Results**

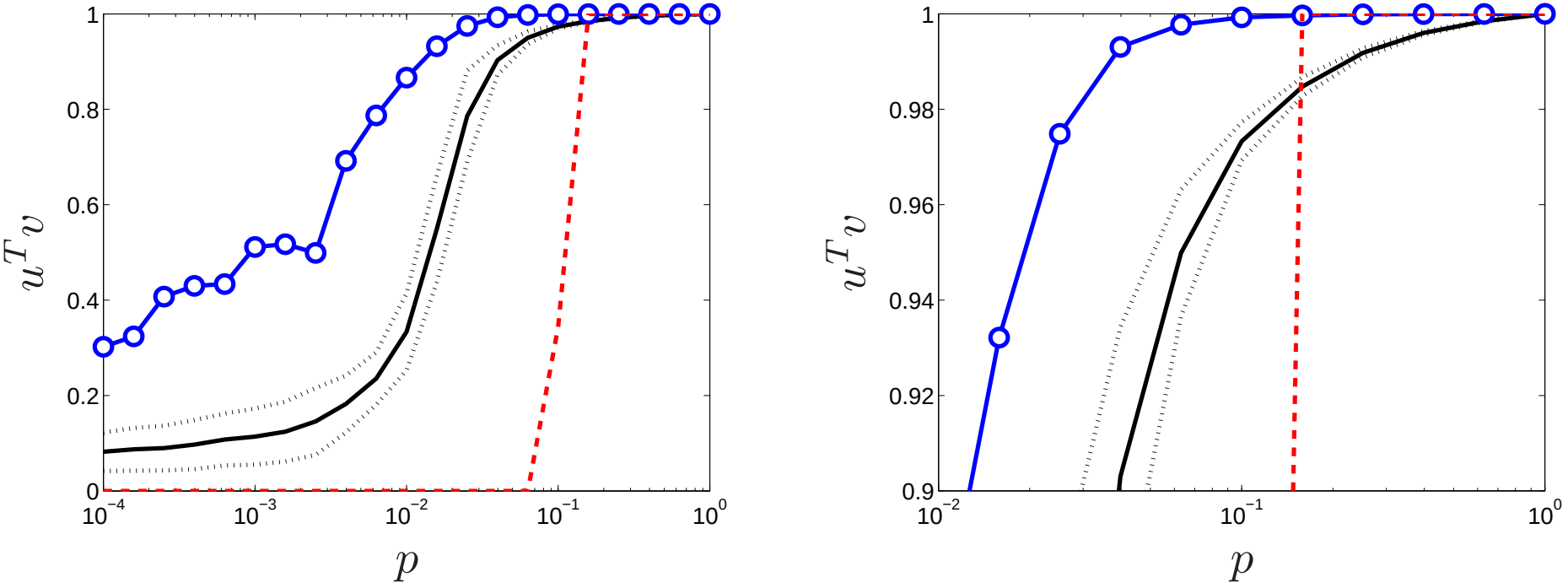
PCA: averaging

A simple experiment.

- Covariance matrix of the 500 most active genes in the cancer data set from Alon et al. (1999).
- Use the subsampling procedure described in the previous slide for various values of the sampling probability p.
- For each sample, measure the **alignment** $u^T v$ between the true eigenvector u and its approximation v.
- Average all subsampled vectors and test quality.

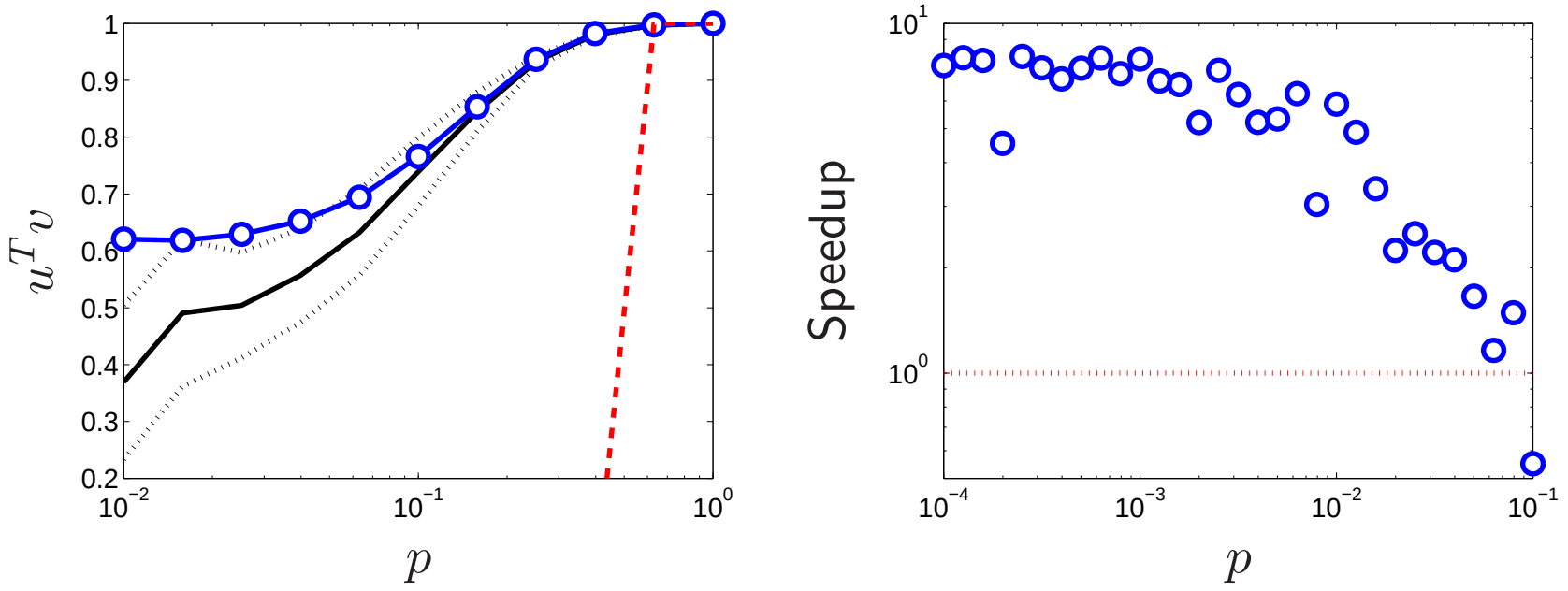
PCA: averaging

Phase transition.



Left: Alignment $u^T v$ between true and normalized average of 1000 subsampled eigenvectors (blue circles), median value of $u^T v$ (solid black line), with dotted lines at plus and minus one stdev, for various values of the sampling probability p on a gene expression covariance matrix.
Right: Zoom on the the interval $p \in [10^{-2}, 1]$.

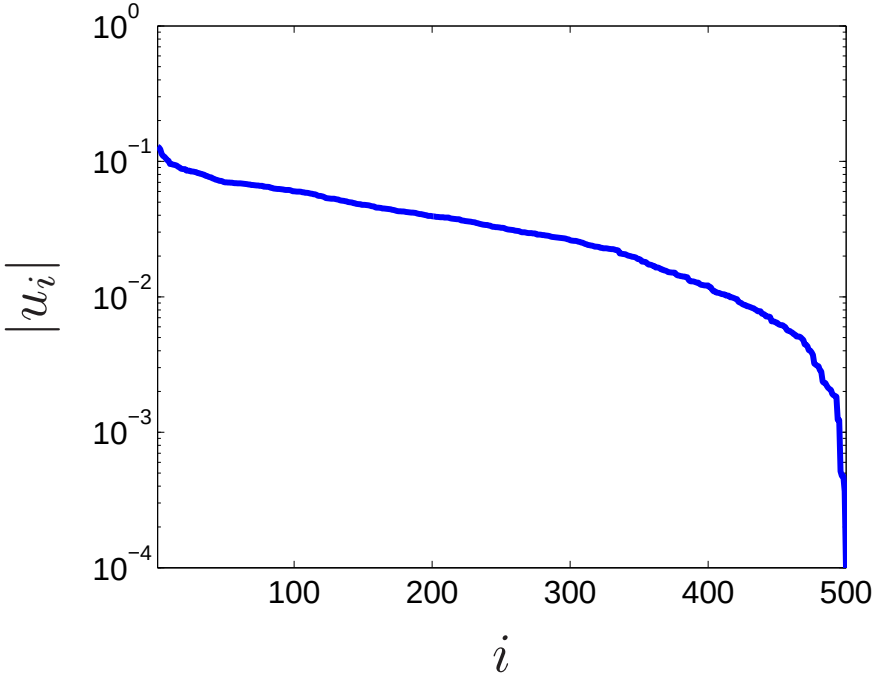
Latent Semantic Indexing: Averaging

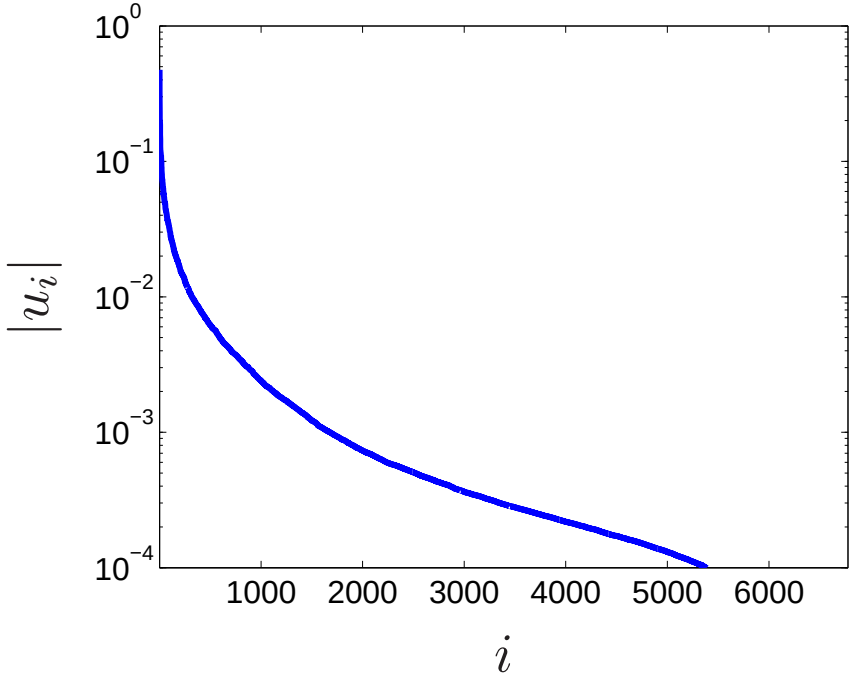


Left: Alignment $u^T v$ between the true and the normalized average of 1000 subsampled left eigenvectors (blue circles), median value (solid black line) and proportion of samples satisfying the perturbation condition (dashed red line), for various values of p on a term document matrices with dimensions 6779×11171.

Right: Speedup in computing leading eigenvector of a gene expression data, including subsampling time, for various values of the sampling probability p. (Memory scales linearly with p).

PCA: Averaging





Magnitude of eigenvector coefficients $|u_i|$ in decreasing order for both the leading eigenvector of the gene expression covariance matrix (left) and the leading left singular vector of the 6779×11171 term document matrix (right).

Numerical results: Ranking

Suppose we are given an adjacency matrix for a **web graph**

$$A_{ij} = 1, \quad \text{if there is a link from } i \text{ to } j$$

with $A \in \mathbf{R}^{n \times n}$. We normalize it into a stochastic matrix

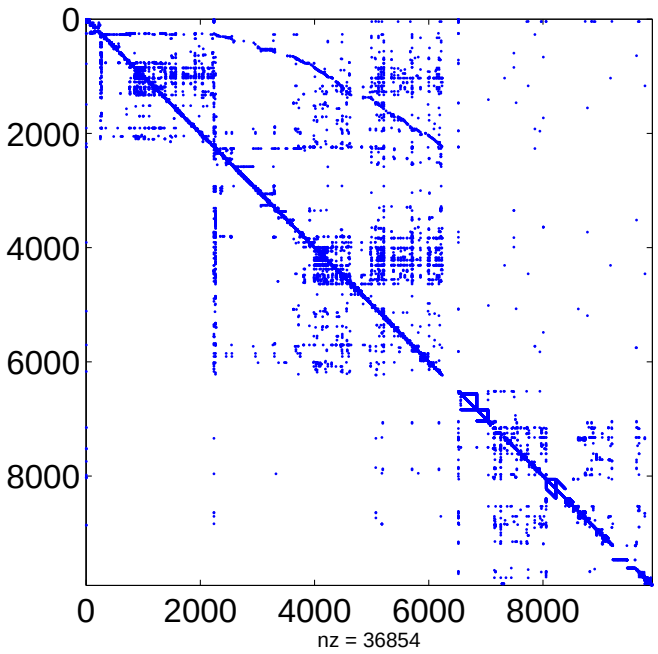
$$P^g_{ij} = \frac{A_{ij}}{\deg_i}$$

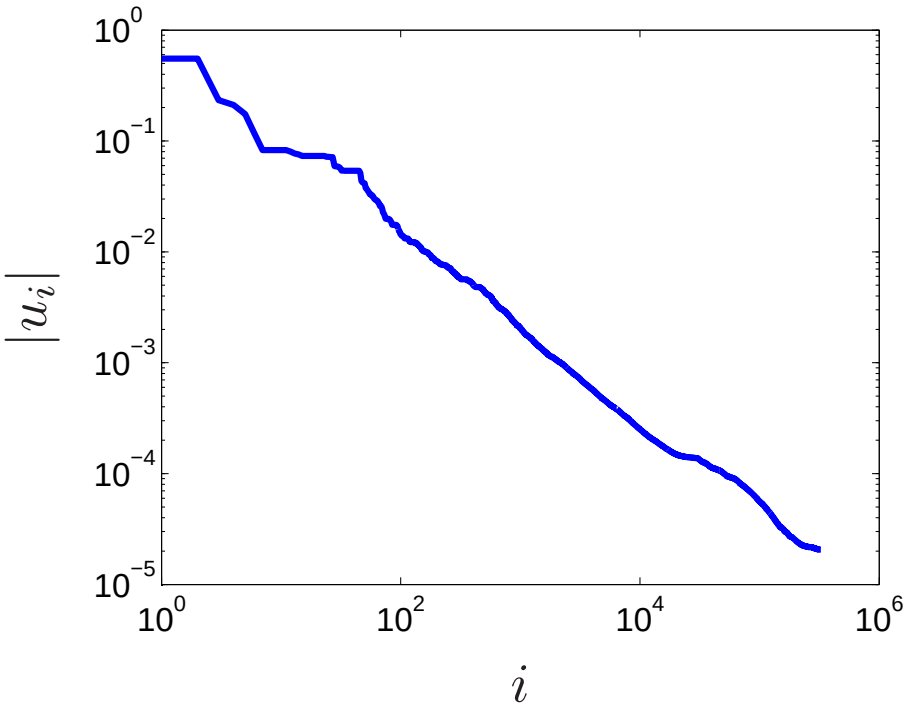
- The matrix P is the transition matrix of a Markov chain on the graph. If we set

$$P = cP^g + (1-c)\mathbf{1}\mathbf{1}^T/n$$

  for $c \in [0, 1]$, this Markov chain will be irreducible.
- The leading (Perron-Frobenius) eigenvector of this matrix is called the **Pagerank** vector (Page et al., 1998).
- The spectral gap is at least c. . .

Numerical results: Ranking





Left: The `wb-cs.stanford` graph (9914 nodes and 36854 edges).
Right: Loglog plot of the Pagerank vector coefficients for the `cnr-2000` graph (325,557 nodes and 3,216,152 egdes).

Webgraph data (Boldi and Vigna, 2004) and `BVGRAPH` by David Gleich.

Numerical results: Ranking

Only the order of the coefficients matters in ranking, so measuring Pearson correlation between pagerank vectors is pointless. For ranking performance, we can use **Spearman's** ρ.

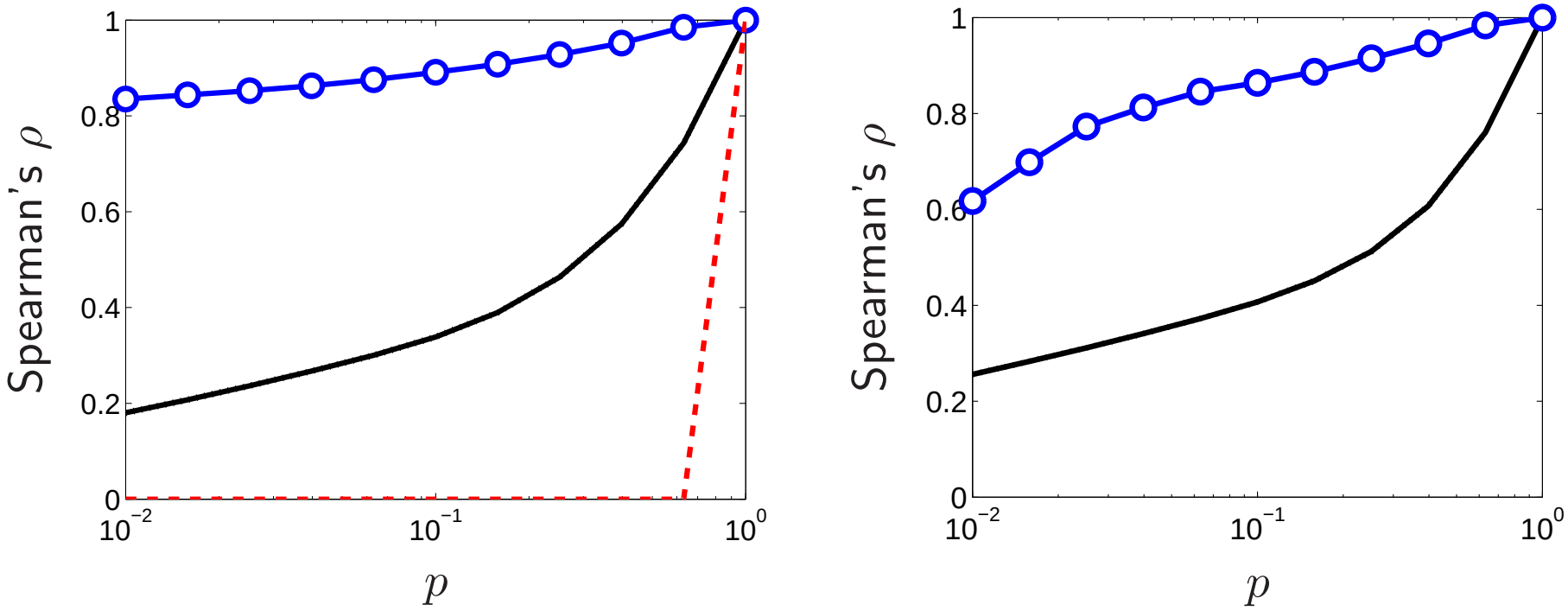
- Suppose we get two pagerank vectors $u, v \in \mathbf{R}^n$, we first convert them into **ranking vectors** $x, y \in [1, n]^n$ such that

$$u_{x_i} \text{ and } v_{y_i} \text{ are decreasing.}$$

- We then compute the Pearson correlation between the ranking vectors

$$\rho(u, v) = \mathbf{corr}(x, y)$$

Ranking



Ranking correlation between true and averaged pagerank vector (blue circles), median value of the correlation over all subsampled matrices (solid black line), proportion of samples satisfying the perturbation condition (dashed red line), for various values of the sampling prob. p.

Left: On the `cnr-2000` graph. 325,557 nodes and 3,216,152 edges, 1000 samples.
Right: On the `UK-2002` graph. 1.8×10^7 nodes and 3×10^8 edges, 100 samples.

Conclusion & Open Questions

- The perturbation regime often holds for surprisingly low sampling rates.
- Averaging produces second-order accurate eigenvector approximations.

What next?

- Applications in optimization (semidefinite programming).
- Explain performance on webgraph matrices?
- Volume sampling produces **relative accuracy** sampling bounds

$$\|A - S_k\|_F^2 \leq (1+\epsilon)\|A - A_k\|_F^2$$

  But computing the sampling probabilities is computationally intensive, which defeats our purpose here.

Slides, source code, binaries: `www.princeton.edu/~aspremon`

*

References

D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM*, 54(2), 2007.

A. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology*, 96:6745–6750, 1999.

S. Arora and S. Kale. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 227–236, 2007.

Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference (WWW 2004)*, pages 595–601, Manhattan, USA, 2004. ACM Press.

E.J. Candes and B. Recht. Exact matrix completion via convex optimization. *preprint*, 2008.

E.J. Candes and T. Tao. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *arXiv:0903.1476*, 2009.

R. Coifman, Y. Shkolnisky, F.J. Sigworth, and A. Singer. Cryo-EM structure determination through eigenvectors of sparse matrices. *working paper*, 2008.

P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix. *SIAM Journal on Computing*, 36:158, 2006.

A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6): 1025–1041, 2004.

G.H. Golub and C.F. Van Loan. Matrix computation. *North Oxford Academic*, 1990.

D. J. Groh, R. A. Marshall, A. B. Kunz, and C. R. Givens. An approximation method for eigenvectors of very large matrices. *Journal of Scientific Computing*, 6(3):251–267, 1991.

R.A. Horn and C.R. Johnson. *Topics in matrix analysis*. Cambridge university press, 1991.

L. Huang, D. Yan, M.I. Jordan, and N. Taft. Spectral Clustering with Perturbed Data. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

R. Kannan and S. Vempala. *Spectral algorithms*. 2009. URL `http://www.cc.gatech.edu/$\sim$vempala/spectralbook.html`.

T. Kato. *Perturbation theory for linear operators*. Springer, 1995.

R.H. Keshavan, A. Montanari, and S. Oh. Matrix Completion from a Few Entries. *arXiv:0901.3150*, 2009.

J. Kuczynski and H. Wozniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM J. Matrix Anal. Appl*, 13(4):1094–1122, 1992.

A.Y. Ng, A.X. Zheng, and M.I. Jordan. Stable algorithms for link analysis. In *ACM SIGIR*, pages 258–266. ACM New York, NY, USA, 2001.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford CS Technical Report*, 1998.

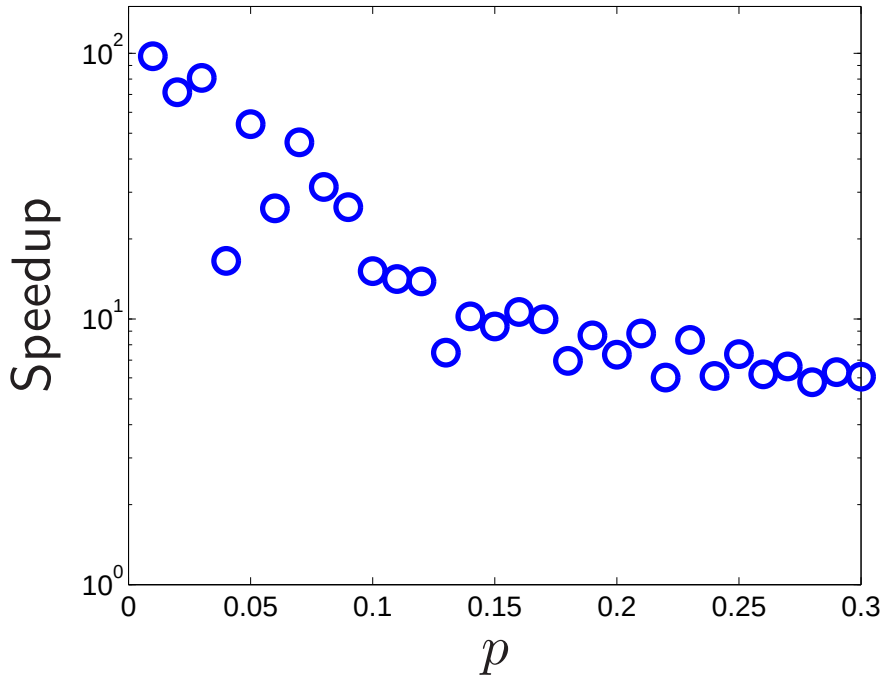
C.H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: a probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *Arxiv preprint arXiv:0706.4138*, 2007.

V.H. Vu. Spectral norm of random matrices. *Combinatorica*, 27(6):721–736, 2007.

Raw CPU gain from sparsity

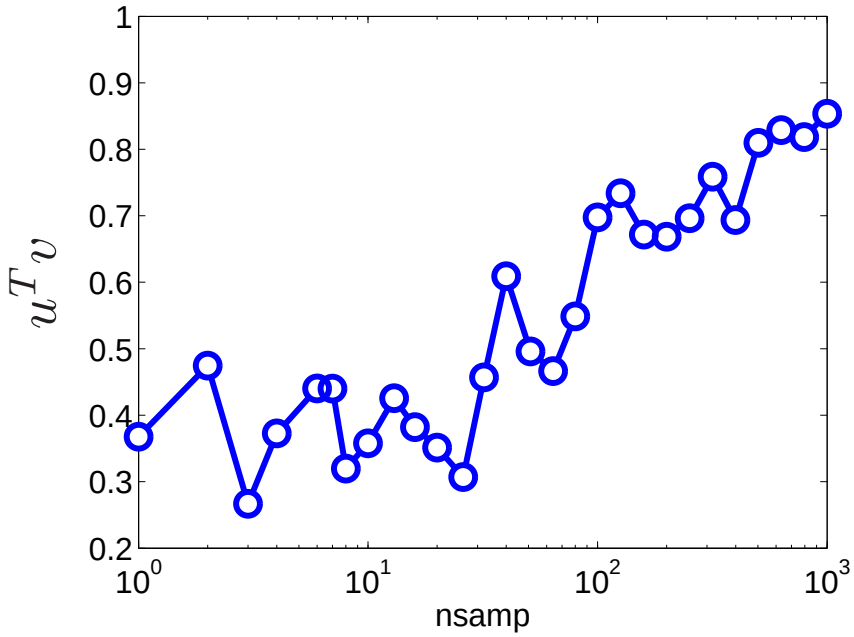
The cost of computing a few leading eigenvectors of a sparse matrix using the power or Lanczos methods (see (Golub and Van Loan, 1990, Chap. 8-9) for example) is proportional to the number of nonzero coefficients.

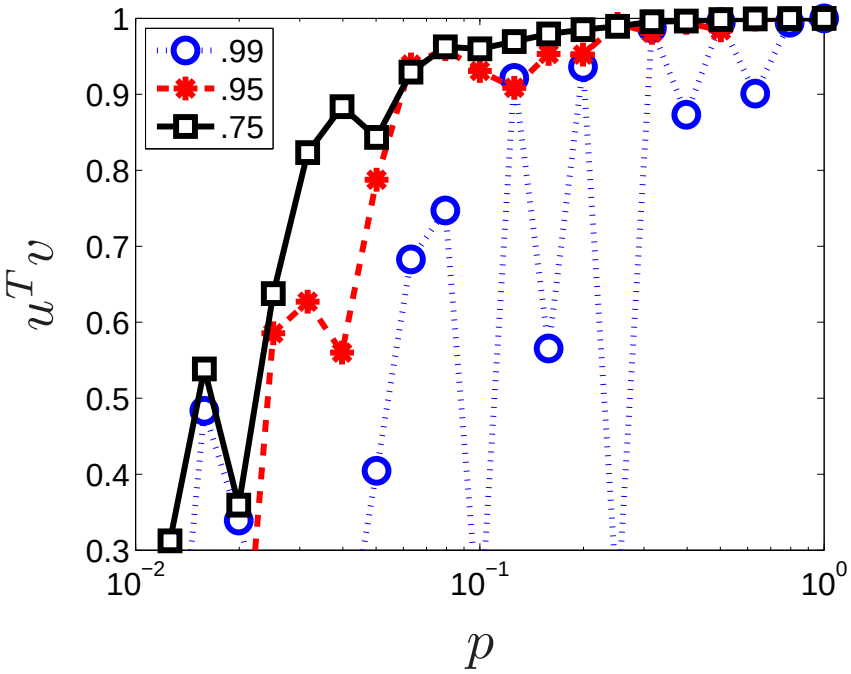


Ratio of average CPU time for computing the leading eigenvalue of a (sparse) subsampled matrix using ARPACK (directly) over average CPU time for the original (dense) matrix, versus sampling probability p, with $n = 2000$.

The main benefit is **lower memory** usage however.

Numerical results: Averaging





Left: Alignment $u^T v$ between the true leading eigenvector u and the normalized average leading eigenvector versus number of samples, on the gene expression covariance matrix with subsampling probability $p = 10^{-2}$.

Right: Alignment $u^T v$ for various values of the spectral gap $\frac{\lambda_2}{\lambda_1} = 0.75, 0.95, 0.99$.