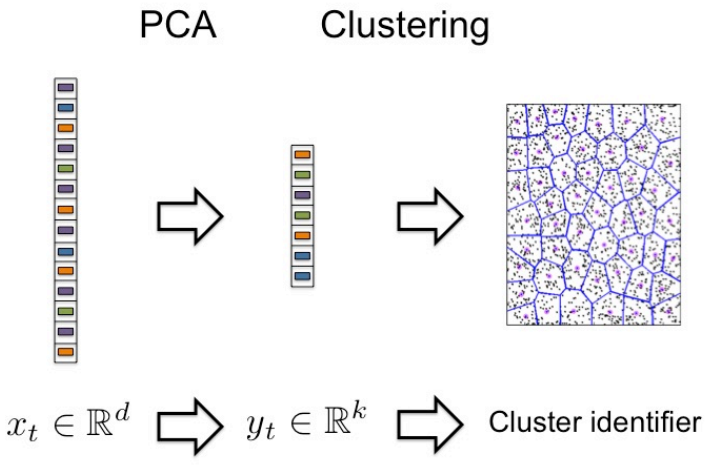
Online Principal Component Analysis

Edo Liberty

PCA Motivation

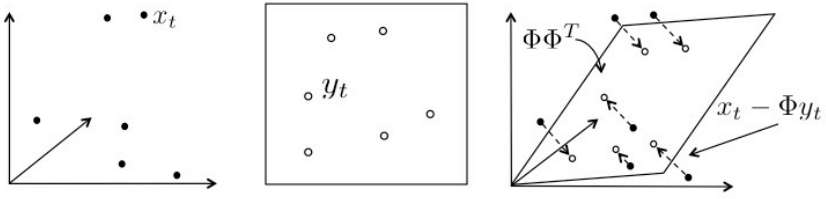


PCA Objective

Given $X \in \mathbb{R}^{d \times n}$ and $k < d$ minimize over $Y \in \mathbb{R}^{k \times n}$

$$\min_{\Phi} \|X - \Phi Y\|_F^2 \quad \text{or} \quad \sum_t \min_{\Phi} \|x_t - \Phi y_t\|^2$$

Think of $X = [x_1, x_2, \ldots]$ and $Y = [y_1, y_2, \ldots]$ as collections of column vectors.



Optimal Offline Solution

Optimal Offline Solution

Let U_k span the top k left singular vectors of X.

- Set $Y = U_k^T X$
- Set $\Phi = U_k$

- Computing U_k is possible offline using the Singular Value Decomposition.
- The optimal reconstruction Φ turns out to be an isometry.

Pass efficient PCA

We can compute U_k from XX^T and

$$XX^T = \sum_t x_t x_t^T .$$

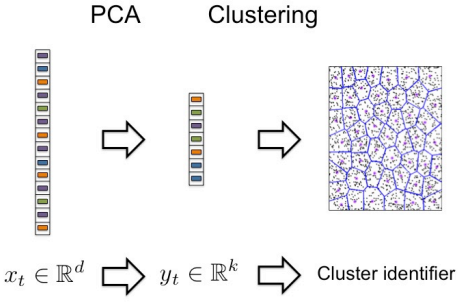
This requires $\Theta(nd^2)$ time (potentially) and $\Theta(d^2)$ space.

Approximating U_k in one pass more efficiently is possible.
[FKV04, DK03, Sar06, DMM08, DRVW06, RV07, WLRT08, CW09, Oli10, CW12, Lib13, GP14, GLPW15]

Nevertheless, a second pass is required to map $x_t \mapsto y_t = U_k^T x_t$.

Online PCA

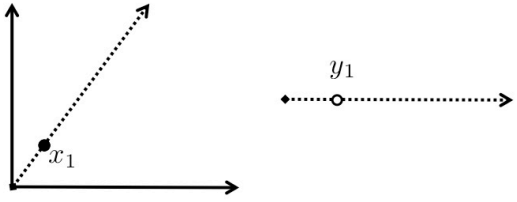
Consider online clustering (e.g. [CCFM97, LSS14]) or online facility location (e.g. [Mey01])



The PCA algorithm must output y_t **before** receiving x_{t+1}.

Online regression

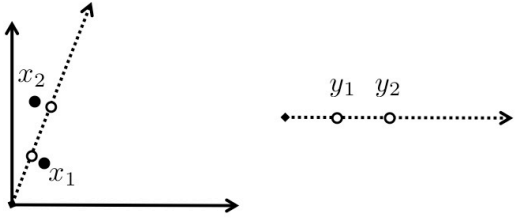
Note that this is non trivial even when $d = 2$ and $k = 1$.



For x_1 there aren't many options...

Online regression

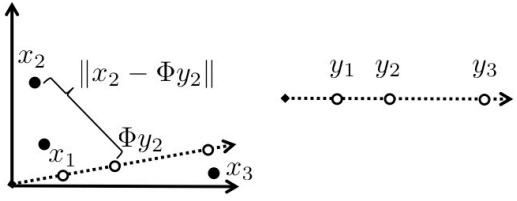
Note that this is non trivial even when $d = 2$ and $k = 1$.



For x_2 this is already a non standard optimization problem

Online regression

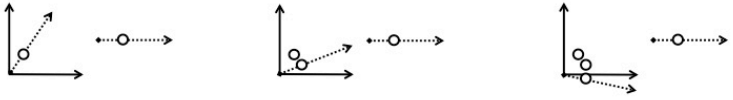
Note that this is non trivial even when $d = 2$ and $k = 1$.



In general, the mapping $x_i \mapsto y_i$ is not necessarily linear.

Online PCA, Possible Problem Definitions

- **Stochastic model:** Bounds $\|X - \Phi Y\|_F^2$ assumes x_t are i.i.d. from an unknown distribution.
  [OK85, ACS13, MCJ13, BDF13]
- **Regret minimization:** Minimizes $\sum_t \|x_t - P_{t-1}x_t\|^2$. Commits to P_{t-1} before observing x_t.
  [WK06, NKW13]



- **Random projection:** can guarantee online that $\|(X - (XY^+)Y\|_F^2$ is small.
  [Sar06, CW09]

Online PCA Problem Definitions

Definition of a (c, ε)-approximation algorithm for Online PCA

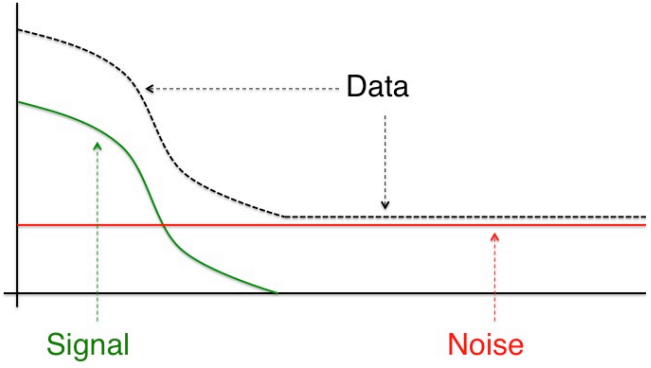
Given $X \in \mathbb{R}^{d \times n}$ as vectors $[x_1, x_2 \ldots]$ and $k < d$ produce $Y = [y_1, y_2, \ldots]$ such that

- y_t is produced before observing x_{t+1}.
- $y_t \in \mathbb{R}^{\ell}$ and $\ell \leq c \cdot k$.
- $\|X - \Phi\, Y\|_F^2 \leq \|X - X_k\|_F^2 + \varepsilon \|X\|_F^2$ for some isometry Φ.

Main Contribution [BGKL15]

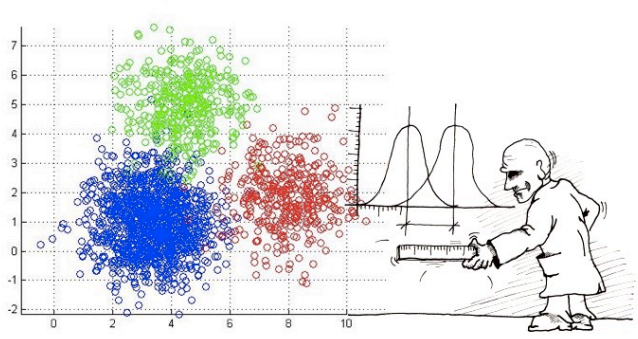
There exists a $(\tilde{O}(\varepsilon^{-2}), \varepsilon)$-approximation algorithm for online PCA.

Noisy Data Spectra



Setting $Y = 0$ gives an $(0, \varepsilon)$ approximation...

Noisy Data Spectra



Sometimes, ”poor reconstruction error” is algorithmically required.

Online PCA Problem Definitions

Setting $Y = U_k^T X$ and $\Phi = U_k$ minimizes

$$\|X - \Phi Y\|_2^2$$

Definition of a (c, ε)-approximation algorithm for Spectral Online PCA

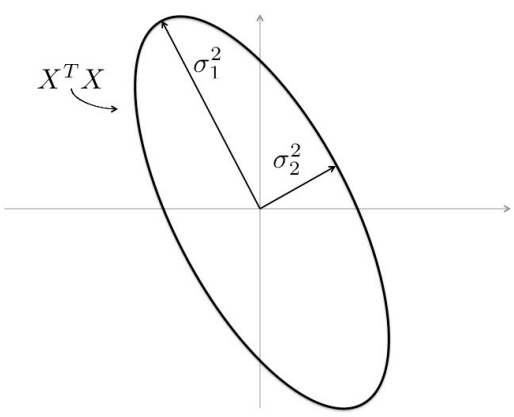
Given $X \in \mathbb{R}^{d \times n}$ as vectors $[x_1, x_2 \ldots]$ and $k < d$ produce $Y = [y_1, y_2, \ldots]$ such that

- y_t is produced before observing x_{t+1}.
- $y_t \in \mathbb{R}^\ell$ and $\ell \leq c \cdot k$.
- $\|X - \Phi Y\|_2^2 \leq \|X - X_k\|_2^2 + \varepsilon \|X\|_2^2$ for some isometry Φ.

Main Contribution [KL15]

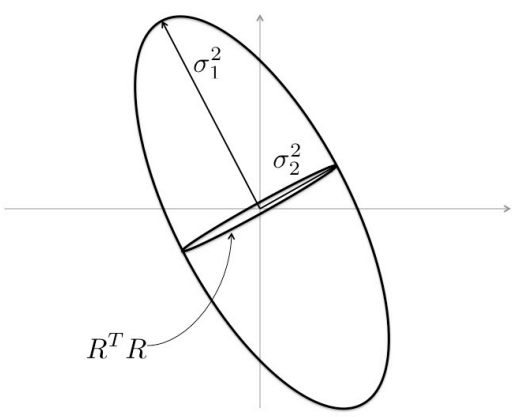
There exists a $(\tilde{O}(\varepsilon^{-2}), \varepsilon)$-approximation algorithm for Spectral Online PCA.

Some Intuition



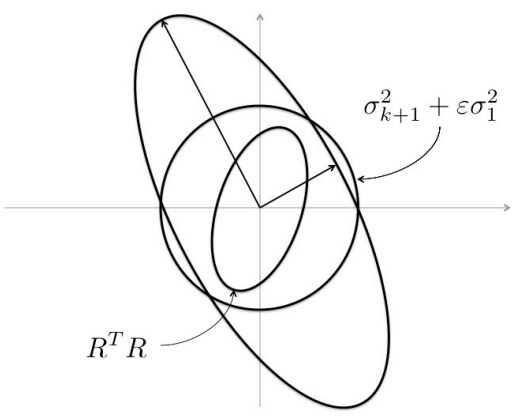
The covariance matrix $X^T X$ visualized as an ellipse.

Some Intuition



The optimal residual is $R = X - X_k$

Some Intuition



Any residual $R = X - \Phi Y$ such that $\|R^T R\| \leq \sigma_{k+1}^2 + \varepsilon\sigma_1^2$ would work

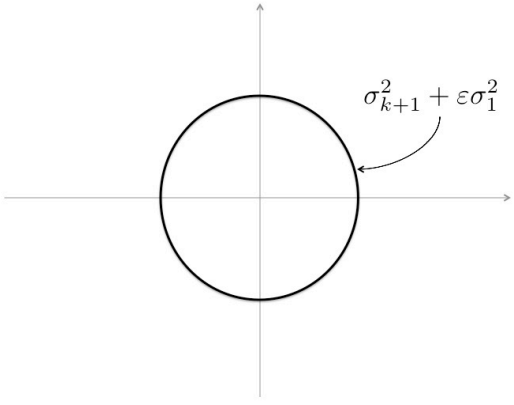
Bad Algorithm, Big Step Forward

$\Delta = \sigma_{k+1}^2 + \varepsilon\sigma_1^2$
$U \leftarrow$ all zeros matrix
for $x_t \in X$ **do**
  if $\|(I - UU^T)X_{1:t}\|^2 \geq \Delta$
    Add the top left singular vector of $(I - UU^T)X_{1:t}$ to U
  yield $y_t = U^T x_t$

Obvious problems with this algorithm (will be fixed later)

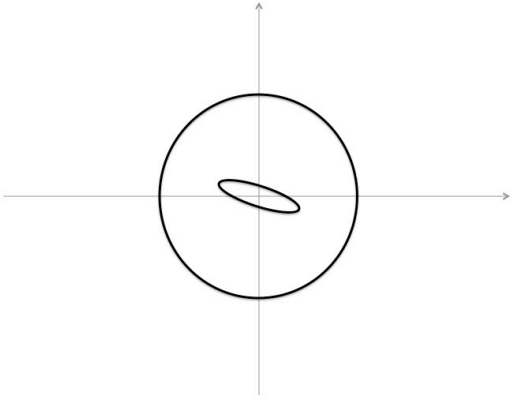
- it must "guess" $\sigma_{k+1}^2 + \varepsilon\sigma_1^2$.
- it stores the entire history $X_{1:t}$
- it computes the top singular value of $(I - UU^T)X_{1:t}$ at every round

Algorithm Intuition



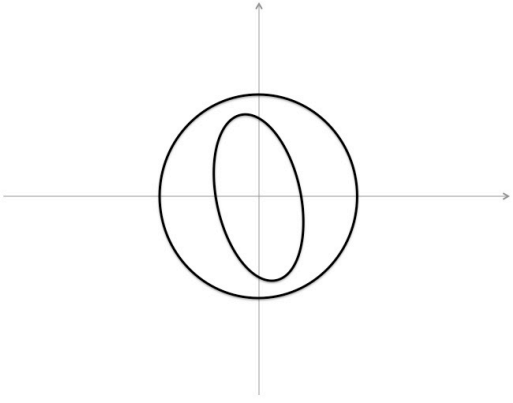
Assume we know $\Delta = \sigma_{k+1}^2 + \varepsilon\sigma_1^2$.

Algorithm Intuition



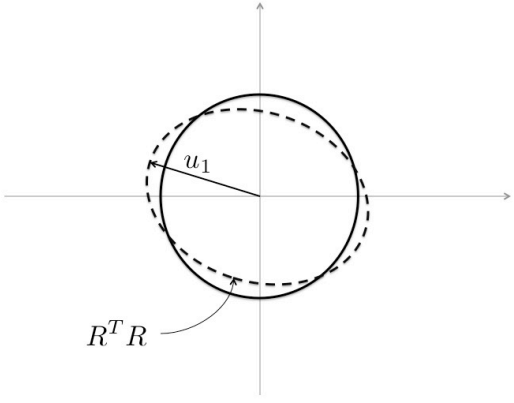
We start with mapping $x_t \mapsto 0$ and $R_{[1:t]} = X_{[1:t]}$

Algorithm Intuition



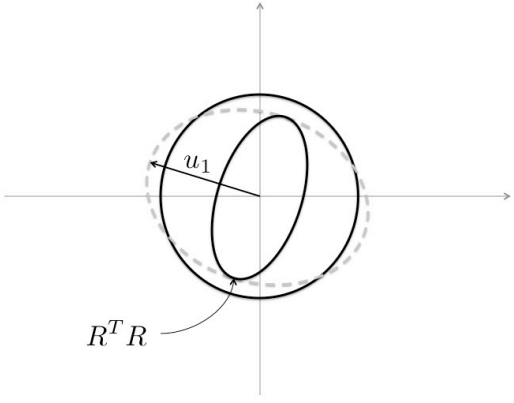
This is continued as long as $\|R^T R\| \leq \Delta$

Algorithm Intuition



When $\|R^T R\| > \Delta$ we commit to a new online PCA direction u_i.

Algorithm Intuition



This prevents $R^T R$ from growing more in the direction u_i.

Algorithm Properties

Theorems 2,5 and 6 in [KL15]

$$\|X - UY\|_2^2 \leq \|R\|_2^2 \leq \sigma_k^2 + \varepsilon\sigma_1^2 + o(\sigma_1^2) \ .$$

"Proof by drawing" above is deceivingly simple. This is the main difficulty!

Theorem 1 in [KL15]

Number of directions added by the algorithm is $\ell \leq k/\varepsilon$.

We sum the inequality $\Delta \leq \|u_i^\top X\|^2$ all added directions $u_1, \ldots, u_\ell$.

$$\ell\Delta \leq \sum_{i=1}^{\ell} \|u_i^\top X\|^2 = \|U_n^\top X\|_F^2 \leq \sum_{i=1}^{\ell} \sigma_i^2 \leq k\sigma_1^2 + (\ell - k)\sigma_{k+1}^2$$

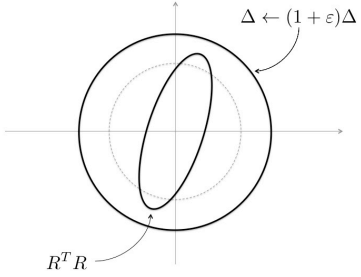
By rearranging we get:

$$\ell \leq (k\sigma_1^2 - k\sigma_{k+1}^2)/(\Delta - \sigma_{k+1}^2)$$

Substituting $\Delta = \sigma_{k+1}^2 + \varepsilon\sigma_1^2$ gives $\ell \leq k/\varepsilon$.

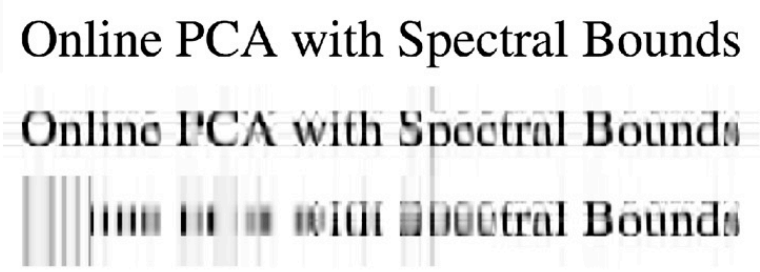
Fixing the Algorithm

- Exponentially search for the right Δ.
  If we added more than k/ε direction to U we can conclude that $\Delta < \sigma_{k+1}^2 + \varepsilon\sigma_1^2$.



- Instead of keeping $X_{1:t}$ use covariance sketching.
  Keep B such that $XX^T \sim BB^T$ and B required $o(d^2)$ to store.
- Only compute the top singular value of $(I - UU^T)X_{1:t}$ "once in a while".

Visual Illustration and Open Problem



- Can we reduce target dimension while keeping the approximation guaranty?
- Would allowing *scaled* isometric registration help reduce the target dimension?
- Can we avoid the exponential search for Δ?
- Is there a simple way to update U that is more accurate than only adding columns?
- Can we reduce the running time of online PCA? Currently the bottleneck is covariance sketching.

Thank you

Nir Ailon and Bernard Chazelle.

Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform.
In *Proceedings of the 38st Annual Symposium on the Theory of Compututing (STOC)*, pages 557–563, Seattle, WA, 2006.

Nir Ailon and Bernard Chazelle.

Faster dimension reduction.
Commun. ACM, 53(2):97–104, 2010.

Dimitris Achlioptas.

Database-friendly random projections: Johnson-Lindenstrauss with binary coins.
J. Comput. Syst. Sci., 66(4):671–687, 2003.

Raman Arora, Andy Cotter, and Nati Srebro.

Stochastic optimization of pca with capped msg.
In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1815–1823. 2013.

Nir Ailon, Zohar Shay Karnin, Edo Liberty, and Yoelle Maarek.

Threading machine generated email.
In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 405–414, 2013.

Nir Ailon and Edo Liberty.

Fast dimension reduction using rademacher series on dual bch codes.
Discrete Comput. Geom., 42(4):615–630, 2009.

Nir Ailon and Edo Liberty.

An almost optimal unrestricted fast johnson-lindenstrauss transform.
In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 185–191, 2011.


Nir Ailon and Holger Rauhut.
Fast and rip-optimal transforms.
Discrete & Computational Geometry, 52(4):780–798, 2014.


Rudolf Ahlswede and Andreas Winter.
Strong converse for identification via quantum channels.
IEEE Transactions on Information Theory, 48(3):569–579, 2002.


Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund.
The fast convergence of incremental pca.
In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3174–3182. 2013.


Christos Boutsidis, Dan Garber, Zohar Shay Karnin, and Edo Liberty.
Online principal components analysis.
In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 887–901, 2015.


Moses Charikar, Kevin Chen, and Martin Farach-Colton.
Finding frequent items in data streams.
In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, ICALP '02, pages 693–703, London, UK, UK, 2002. Springer-Verlag.


Kenneth L Clarkson and David P Woodruff.
Numerical linear algebra in the streaming model.
In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009.


Kenneth L. Clarkson and David P. Woodruff.
Low rank approximation and regression in input sparsity time.
CoRR, abs/1207.6365, 2012.

S. DasGupta and A. Gupta.

An elementary proof of the Johnson-Lindenstrauss lemma.
Technical Report, UC Berkeley, 99-006, 1999.

Petros Drineas and Ravi Kannan.

Pass efficient algorithms for approximating large matrices, 2003.

Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós.

A sparse johnson: Lindenstrauss transform.
In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 341–350, 2010.

Erik D. Demaine, Alejandro López-Ortiz, and J. Ian Munro.

Frequency estimation of internet packet streams with limited space.
In *Proceedings of the 10th Annual European Symposium on Algorithms*, ESA '02, pages 348–360, London, UK, UK, 2002. Springer-Verlag.

Alan Frieze, Ravi Kannan, and Santosh Vempala.

Fast monte-carlo algorithms for finding low-rank approximations.
J. ACM, 51(6):1025–1041, November 2004.

P. Frankl and H. Maehara.

The Johnson-Lindenstrauss lemma and the sphericity of some graphs.
Journal of Combinatorial Theory Series A, 44:355–362, 1987.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff.

Frequent directions : Simple and deterministic matrix sketching.
CoRR, abs/1501.01711, 2015.


Mina Ghashami and Jeff M. Phillips.
Relative errors for deterministic low-rank matrix approximations.
In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 707–717, 2014.


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


Zohar Shay Karnin and Edo Liberty.
Online pca with spectral bounds.
In progress, 2015.


Daniel M. Kane and Jelani Nelson.
A derandomized sparse johnson-lindenstrauss transform.
Electronic Colloquium on Computational Complexity (ECCC), 17:98, 2010.


Richard M. Karp, Christos H. Papadimitriou, and Scott Shenker.
A simple algorithm for finding frequent elements in streams and bags.
ACM Transactions on Database Systems, 28:2003, 2003.


Felix Krahmer and Rachel Ward.
New and improved johnson-lindenstrauss embeddings via the restricted isometry property.
SIAM J. Math. Analysis, 43(3):1269–1281, 2011.


Edo Liberty, Nir Ailon, and Amit Singer.
Dense fast random projections and lean walsh transforms.
Discrete & Computational Geometry, 45(1):34–44, 2011.

Edo Liberty.

Simple and deterministic matrix sketching.
In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 581–588, New York, NY, USA, 2013. ACM.

Edo Liberty, Ram Sriharsha, and Maxim Sviridenko.

An algorithm for online k-means clustering.
CoRR, abs/1412.5721, 2014.

Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi.

Efficient computation of frequent and top-k elements in data streams.
In Thomas Eiter and Leonid Libkin, editors, *Database Theory - ICDT 2005*, volume 3363 of *Lecture Notes in Computer Science*, pages 398–412. Springer Berlin Heidelberg, 2005.

Jirí Matousek.

On variants of the johnson-lindenstrauss lemma.
Random Struct. Algorithms, 33(2):142–156, 2008.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain.

Memory limited, streaming pca.
In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2886–2894. 2013.

Adam Meyerson.

Online facility location.
In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 426–431, 2001.

Jayadev Misra and David Gries.

Finding repeated elements.
Technical report, Ithaca, NY, USA, 1982.

Jiazhong Nie, Wojciech Kotlowski, and Manfred K. Warmuth.

Online pca with optimal regrets.
In *ALT*, pages 98–112, 2013.

Jelani Nelson and Huy L. Nguyen.

Sparsity lower bounds for dimensionality reducing maps.
In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 101–110, 2013.

Jelani Nelson and Huy L. Nguyên.

Lower bounds for oblivious subspace embeddings.
In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pages 883–894, 2014.

Jelani Nelson, Huy L. Nguyên, and David P. Woodruff.

On deterministic sketching and streaming for sparse recovery and norm estimation.
In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 627–638, 2012.

Erkki Oja and Juha Karhunen.

On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix.
Journal of Mathematical Analysis and Applications, 106(1):69 – 84, 1985.

Roberto Imbuzeiro Oliveira.

Sums of random hermitian matrices and an inequality by rudelson.
arXiv:1004.3821v1, April 2010.

Mark Rudelson and Roman Vershynin.

Sampling from large matrices: An approach through geometric functional analysis.
J. ACM, 54(4), July 2007.

Tamas Sarlos.

Improved approximation algorithms for large matrices via random projections.
In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, Washington, DC, USA, 2006.

Roman Vershynin.

A note on sums of independent random matrices after ahlswede-winter.
Lecture Notes.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg.

Feature hashing for large scale multitask learning.
In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1113–1120, New York, NY, USA, 2009. ACM.

Manfred K. Warmuth and Dima Kuzmin.

Randomized PCA algorithms with regret bounds that are logarithmic in the dimension.
In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1481–1488, 2006.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert.

A fast randomized algorithm for the approximation of matrices.
Applied and Computational Harmonic Analysis, 25(3):335 – 366, 2008.

David P. Woodruff.

Low rank approximation lower bounds in row-update streams.
In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1781–1789, 2014.

Amit Deshpande, Luis Rademacher, Santosh Vempala, Grant Wang

Matrix Approximation and Projective Clustering via Volume Sampling
in *Theory of Computing* pages 225–247, 2006

P. Drineas, M. W. Mahoney, and S. Muthukrishnan

Relative-error cur matrix decompositions
in *SIAM Journal Matrix Analysis and Applications*, 30(2):844?881, 2008.

Moses Charikar, Chandra Chekuri, Tomás Feder, Rajeev Motwani

Incremental Clustering and Dynamic Information Retrieval
Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing, El Paso, Texas, USA, May 4-6, 1997

Raman Arora, Andrew Cotter, Nati Srebro

Stochastic Optimization of PCA with Capped MSG

Advances in Neural Information Processing Systems, 2013, pp 1815–1823