Randomized Algorithms for Matrices and Data

To Xiomara

Full text available at: http://dx.doi.org/10.1561/2200000035

Randomized Algorithms for Matrices and Data

Michael W. Mahoney

Department of Mathematics
Stanford University
Stanford, CA 94305
USA
mmahoney@cs.stanford.edu

now
the essence of knowledge

Boston – Delft

**Foundations and Trends® in
Machine Learning**

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is M. W. Mahoney, Randomized Algorithms for Matrices and Data, Foundation and Trends® in Machine Learning, vol 3, no 2, pp 123–224, 2010

ISBN: 978-1-60198-506-4

Foundations and Trends® in Machine Learning

Volume 3 Issue 2, 2010

Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2010, Volume 3, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in
Machine Learning
Vol. 3, No. 2 (2010) 123–224
© 2011 M. W. Mahoney
DOI: 10.1561/2200000035

Randomized Algorithms for Matrices and Data

Michael W. Mahoney

Department of Mathematics, Stanford University, Stanford, CA 94305, USA, mmahoney@cs.stanford.edu

Abstract

Randomized algorithms for very large matrix problems have received a great deal of attention in recent years. Much of this work was motivated by problems in large-scale data analysis, largely since matrices are popular structures with which to model data drawn from a wide range of application domains, and this work was performed by individuals from many different research communities. While the most obvious benefit of randomization is that it can lead to faster algorithms, either in worst-case asymptotic theory and/or numerical implementation, there are numerous other benefits that are at least as important. For example, the use of randomization can lead to simpler algorithms that are easier to analyze or reason about when applied in counterintuitive settings; it can lead to algorithms with more interpretable output, which is of interest in applications where analyst time rather than just computational time is of interest; it can lead implicitly to regularization and more robust output; and randomized algorithms can often be organized to exploit modern computational architectures better than classical numerical methods.

This monograph will provide a detailed overview of recent work on the theory of randomized matrix algorithms as well as the application of those ideas to the solution of practical problems in large-scale data analysis. Throughout this review, an emphasis will be placed on a few simple core ideas that underlie not only recent theoretical advances but also the usefulness of these tools in large-scale data applications. Crucial in this context is the connection with the concept of statistical leverage. This concept has long been used in statistical regression diagnostics to identify outliers; and it has recently proved crucial in the development of improved worst-case matrix algorithms that are also amenable to high-quality numerical implementation and that are useful to domain scientists. This connection arises naturally when one explicitly decouples the effect of randomization in these matrix algorithms from the underlying linear algebraic structure. This decoupling also permits much finer control in the application of randomization, as well as the easier exploitation of domain knowledge.

Most of the review will focus on random sampling algorithms and random projection algorithms for versions of the linear least-squares problem and the low-rank matrix approximation problem. These two problems are fundamental in theory and ubiquitous in practice. Randomized methods solve these problems by constructing and operating on a randomized sketch of the input matrix A — for random sampling methods, the sketch consists of a small number of carefully-sampled and rescaled columns/rows of A, while for random projection methods, the sketch consists of a small number of linear combinations of the columns/rows of A. Depending on the specifics of the situation, when compared with the best previously-existing deterministic algorithms, the resulting randomized algorithms have worst-case running time that is asymptotically faster; their numerical implementations are faster in terms of clock-time; or they can be implemented in parallel computing environments where existing numerical algorithms fail to run at all. Numerous examples illustrating these observations will be described in detail.

Contents

1

Introduction

This monograph will provide a detailed overview of recent work on the theory of *randomized matrix algorithms* as well as the application of those ideas to the solution of practical problems in large-scale data analysis. By "randomized matrix algorithms," we refer to a class of recently-developed random sampling and random projection algorithms for ubiquitous linear algebra problems such as least-squares regression and low-rank matrix approximation. These and related problems are ubiquitous since matrices are fundamental mathematical structures for representing data drawn from a wide range of application domains. Moreover, the widespread interest in randomized algorithms for these problems arose due to the need for principled algorithms to deal with the increasing size and complexity of data that are being generated in many of these application areas.

Not surprisingly, algorithmic procedures for working with matrix-based data have been developed from a range of diverse perspectives by researchers from a wide range of areas — including, e.g., researchers from theoretical computer science (TCS), numerical linear algebra (NLA), statistics, applied mathematics, data analysis, and machine learning, as well as domain scientists in physical and biological sciences — and in many of these cases they have drawn strength from their domain-specific insight. Although this has been great for the development of the area, and for the "technology transfer" of theoretical ideas to practical applications, the technical aspects of dealing with any one of those areas has obscured for many the simplicity and generality of some of the underlying ideas; thus leading researchers to fail to appreciate the underlying connections and the significance of contributions by researchers outside their own area. Thus, rather than focusing on the technical details of proving worst-case bounds or of providing high-quality numerical implementations or of relating to traditional machine learning tools or of using these algorithms in a particular physical or biological domain, in this review we will focus on highlighting for a broad audience the simplicity and generality of some core ideas — ideas that are often obscured but that are fruitful for using these randomized algorithms in large-scale data applications. To do so, we will focus on two fundamental and ubiquitous matrix problems — least-squares approximation and low-rank matrix approximation — that have been at the center of these recent developments.

The work we will review here had its origins within TCS. In this area, one typically considers a particular well-defined problem, and the goal is to prove bounds on the running time and quality-of-approximation guarantees for algorithms for that particular problem that hold for "worst-case" input. That is, the bounds should hold for *any* input matrix, independent of any "niceness" assumptions such as, e.g., that the elements of the matrix satisfy some smoothness or normalization condition or that the spectrum of the matrix satisfies some decay condition. Clearly, the generality of this approach means that the bounds will be suboptimal — and thus can be improved — in any particular application where stronger assumptions can be made about the input. Importantly, though, it also means that the underlying algorithms and techniques will be broadly applicable even in situations where such assumptions do not apply.

An important feature in the use of randomized algorithms in TCS more generally is that one must identify and then algorithmically deal with relevant "non-uniformity structure" in the data.[1] For the randomized matrix algorithms to be reviewed here and that have proven useful recently in NLA and large-scale data analysis applications, the relevant non-uniformity structure is defined by the so-called *statistical leverage scores*. Defined more precisely below, these leverage scores are basically the diagonal elements of the projection matrix onto the dominant part of the spectrum of the input matrix. As such, they have a long history in statistical data analysis, where they have been used for outlier detection in regression diagnostics. More generally, and very importantly for practical large-scale data applications of recently-developed randomized matrix algorithms, these scores often have a very natural interpretation in terms of the data and processes generating the data. For example, they can be interpreted in terms of the leverage or influence that a given data point has on, say, the best low-rank matrix approximation; and this often has an interpretation in terms of high-degree nodes in data graphs, very small clusters in noisy data, coherence of information, articulation points between clusters, etc.

Historically, although the first generation of randomized matrix algorithms (to be described in Section 3) achieved what is known as additive-error bounds and were extremely fast, requiring just a few passes over the data from external storage, these algorithms did *not* gain a foothold in NLA and only heuristic variants of them were used in machine learning and data analysis applications. In order to "bridge the gap" between NLA, TCS, and data applications, much finer control over the random sampling process was needed. Thus, in the second generation of randomized matrix algorithms (to be described in Sections 4 and 5) that *has* led to high-quality numerical implementations

[1] For example, for those readers familiar with Markov chain-based Monte Carlo algorithms as used in statistical physics, this non-uniformity structure is given by the Boltzmann distribution, in which case the algorithmic question is how to sample efficiently with respect to it as an importance sampling distribution without computing the intractable partition function. Of course, if the data are sufficiently nice (or if they have been sufficiently preprocessed, or if sufficiently strong assumptions are made about them, etc.), then that non-uniformity structure might be uniform, in which case simple methods like uniform sampling might be appropriate — but this is far from true in general, either in worst-case theory or in practical applications.

and useful machine learning and data analysis applications, two key developments were crucial.

- **Decoupling the randomization from the linear algebra.** This was originally implicit within the analysis of the second generation of randomized matrix algorithms, and then it was made explicit. By making this decoupling explicit, not only were improved quality-of-approximation bounds achieved, but also *much* finer control was achieved in the application of randomization. For example, it permitted easier exploitation of domain expertise, in both numerical analysis and data analysis applications.
- **Importance of statistical leverage scores.** Although these scores have been used historically for outlier detection in statistical regression diagnostics, they have also been crucial in the recent development of randomized matrix algorithms. Roughly, the best random sampling algorithms use these scores to construct an importance sampling distribution to sample with respect to; and the best random projection algorithms rotate to a basis where these scores are approximately uniform and thus in which uniform sampling is appropriate.

As will become clear, these two developments are very related. For example, once the randomization was decoupled from the linear algebra, it became nearly obvious that the “right” importance sampling probabilities to use in random sampling algorithms are those given by the statistical leverage scores, and it became clear how to improve the analysis and numerical implementation of random projection algorithms. It is remarkable, though, that statistical leverage scores define the non-uniformity structure that is relevant not only to obtain the strongest worst-case bounds, but also to lead to high-quality numerical implementations (by numerical analysts) as well as algorithms that are useful in downstream scientific applications (by machine learners and data analysts).

Most of this review will focus on random sampling algorithms and random projection algorithms for versions of the linear least-squares problem and the low-rank matrix approximation problem. Here is a brief summary of some of the highlights of what follows.

- **Least-squares approximation.** Given an $m \times n$ matrix A, with $m \gg n$, and an m-dimensional vector b, the overconstrained least-squares approximation problem looks for the vector $x_{opt} = \text{argmin}_x ||Ax - b||_2$. This problem typically arises in statistical models where the rows of A and elements of b correspond to constraints and the columns of A and elements of x correspond to variables. Classical methods, including the Cholesky decomposition, versions of the QR decomposition, and the Singular Value Decomposition, compute a solution in $O(mn^2)$ time. Randomized methods solve this problem by constructing a randomized sketch of the matrix A — for random sampling methods, the sketch consists of a small number of carefully-sampled and rescaled rows of A (and the corresponding elements of b), while for random projection methods, the sketch consists of a small number of linear combinations of the rows of A and elements of b. If one then solves the (still overconstrained) subproblem induced on the sketch, then very fine relative-error approximations to the solution of the original problem are obtained. In addition, for a wide range of values of m and n, the running time is $o(mn^2)$ — for random sampling algorithms, the computational bottleneck is computing appropriate importance sampling probabilities, while for random projection algorithms, the computational bottleneck is implementing the random projection operation. Alternatively, if one uses the sketch to compute a preconditioner for the original problem, then very high-precision approximations can be obtained by then calling classical numerical iterative algorithms. Depending on the specifics of the situation, these numerical implementations run in $o(mn^2)$ time; they are faster in terms of clock-time than the best previously-existing deterministic numerical implementations; or they can be implemented in parallel computing environments where existing numerical algorithms fail to run at all.

- **Low-rank matrix approximation.** Given an $m \times n$ matrix A and a rank parameter k, the low-rank matrix approximation problem is to find a good approximation to A of rank $k \ll \min\{m,n\}$. The Singular Value Decomposition provides the best rank-k approximation to A, in the sense that by projecting A onto its top k left or right singular vectors, then one obtains the best approximation to A with respect to the spectral and Frobenius norms. The running time for classical low-rank matrix approximation algorithms depends strongly on the specifics of the situation — for dense matrices, the running time is typically $O(mnk)$; while for sparse matrices, classical Krylov subspace methods are used. As with the least-squares problem, randomized methods for the low-rank matrix approximation problem construct a randomized sketch — consisting of a small number of either actual columns or linear combinations of columns — of the input A, and then this sketch is manipulated depending on the specifics of the situation. For example, random sampling methods can use the sketch directly to construct relative-error low-rank approximations such as CUR decompositions that approximate A based on a small number of actual columns of the input matrix. Alternatively, random projection methods can improve the running time for dense problems to $O(mn\log k)$; and while they only match the running time for classical methods on sparse matrices, they lead to more robust algorithms that can be reorganized to exploit parallel computing architectures.

These two problems are the main focus of this review since they are both fundamental in theory and ubiquitous in practice and since in both cases novel theoretical ideas have already yielded practical results. Although not the main focus of this review, other related matrix-based problems to which randomized methods have been applied will be referenced at appropriate points.

Clearly, when a very new paradigm is compared with very well-established methods, a naïve implementation of the new ideas will perform poorly by traditional metrics. Thus, in both data analysis and numerical analysis applications of this randomized matrix algorithm paradigm, the best results have been achieved when coupling closely with more traditional methods. For example, in data analysis applications, this has meant working closely with geneticists and other domain experts to understand how the non-uniformity structure in the data is useful for their downstream applications. Similarly, in scientific computation applications, this has meant coupling with traditional numerical methods for improving quantities like condition numbers and convergence rates. When coupling in this manner, however, qualitatively improved results have *already* been achieved. For example, in their empirical evaluation of the random projection algorithm for the least-squares approximation problem, to be described in Sections 4.4 and 4.5 below, Avron, Maymounkov, and Toledo [9] began by observing that "Randomization is arguably the most exciting and innovative idea to have hit linear algebra in a long time;" and since their implementation "beats LAPACK's[2] direct dense least-squares solver by a large margin on essentially any dense tall matrix," they concluded that their empirical results "show the potential of random sampling algorithms and suggest that random projection algorithms should be incorporated into future versions of LAPACK."

The remainder of this review will cover these topics in greater detail. To do so, we will start in Section 2 with a few motivating applications from one scientific domain where these randomized matrix algorithms have already found application, and we will describe in Section 3 general background on randomized matrix algorithms, including precursors to those that are the main subject of this review. Then, in the next two sections, we will describe randomized matrix algorithms for two fundamental matrix problems: Section 4 will be devoted to describing several related algorithms for the least-squares approximation problem; and Section 5 will be devoted to describing several related algorithms for the problem of low-rank matrix approximation. Then, Section 6 will describe in more detail some of these issues from

[2] LAPACK (short for Linear Algebra PACKage) is a high-quality and widely-used software library of numerical routines for solving a wide range of numerical linear algebra problems.

an empirical perspective, with an emphasis on the ways that statistical leverage scores have been used more generally in large-scale data analysis; Section 7 will provide some more general thought on this successful technology transfer experience; and Section 8 will provide a brief conclusion.

References

[1] D. Achlioptas, “Database-friendly random projections: Johnson-Lindenstrauss with binary coins,” *Journal of Computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[2] D. Achlioptas and F. McSherry, “Fast computation of low-rank matrix approximations,” *Journal of the ACM*, vol. 54, no. 2, p. Article 9, 2007.

[3] N. Ailon and B. Chazelle, “Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform,” in *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pp. 557–563, 2006.

[4] N. Ailon and B. Chazelle, “The fast Johnson-Lindenstrauss transform and approximate nearest neighbors,” *SIAM Journal on Computing*, vol. 39, no. 1, pp. 302–322, 2009.

[5] N. Ailon and B. Chazelle, “Faster dimension reduction,” *Communications of the ACM*, vol. 53, no. 2, pp. 97–104, 2010.

[6] N. Ailon and E. Liberty, “Fast dimension reduction using Rademacher series on dual BCH codes,” in *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1–9, 2008.

[7] N. Ailon and E. Liberty, “An almost optimal unrestricted fast Johnson-Lindenstrauss transform,” in *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 185–191, 2011.

[8] O. Alter, P. O. Brown, and D. Botstein, “Singular value decomposition for genome-wide expression data processing and modeling,” *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 18, pp. 10101–10106, 2000.

[9] H. Avron, P. Maymounkov, and S. Toledo, “Blendenpik: Supercharging LAPACK’s least-squares solver,” *SIAM Journal on Scientific Computing*, vol. 32, pp. 1217–1236, 2010.

[10] M. Baboulin, J. Dongarra, and S. Tomov, "Some issues in dense linear algebra for multicore and special purpose architectures," Technical Report UT-CS-08-200, University of Tennessee, May 2008.

[11] N. M. Ball and R. J. Brunner, "Data mining and machine learning in astronomy," *International Journal of Modern Physics D*, vol. 19, no. 7, pp. 1049–1106, 2010.

[12] N. M. Ball, J. Loveday, M. Fukugita, O. Nakamura, S. Okamura, J. Brinkmann, and R. J. Brunner, "Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks," *Monthly Notices of the Royal Astronomical Society*, vol. 348, no. 3, pp. 1038–1046, 2004.

[13] A. Banerjee and J. Jost, "On the spectrum of the normalized graph Laplacian," *Linear Algebra and its Applications*, vol. 428, no. 11–12, pp. 3015–3022, 2008.

[14] A. Banerjee and J. Jost, "Graph spectra as a systematic tool in computational biology," *Discrete Applied Mathematics*, vol. 157, no. 10, pp. 2425–2431, 2009.

[15] P. Barooah and J. P. Hespanha, "Graph effective resistances and distributed control: Spectral properties and applications," in *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 3479–3485, 2006.

[16] C. Bekas, E. Kokiopoulou, and Y. Saad, "An estimator for the diagonal of a matrix," *Applied Numerical Mathematics*, vol. 57, pp. 1214–1229, 2007.

[17] M.-A. Belabbas and P. J. Wolfe, "Fast low-rank approximation for covariance matrices," in *Second IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pp. 293–296, 2007.

[18] M.-A. Belabbas and P. J. Wolfe, "On sparse representations of linear operators and the approximation of matrix products," in *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*, pp. 258–263, 2008.

[19] M.-A. Belabbas and P. J. Wolfe, "On landmark selection and sampling in high-dimensional data analysis," *Philosophical Transactions of the Royal Society, Series A*, vol. 367, pp. 4295–4312, 2009.

[20] M.-A. Belabbas and P. J. Wolfe, "Spectral methods in machine learning and new strategies for very large datasets," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 369–374, 2009.

[21] M.-A. Belabbas and P. Wolfe, "On the approximation of matrix products and positive definite matrices," Technical report. Preprint: arXiv:0707.4448, 2007.

[22] M. W. Berry and M. Browne, "Email surveillance using non-negative matrix factorization," *Computational and Mathematical Organization Theory*, vol. 11, no. 3, pp. 249–264, 2005.

[23] M. W. Berry, S. A. Pulatova, and G. W. Stewart, "Computing sparse reduced-rank approximations to sparse matrices," Technical Report UMIACS TR-2004-32 CMSC TR-4589, University of Maryland, College Park, MD, 2004.

[24] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proceedings of the 7th Annual ACM SIGKDD Conference*, pp. 245–250, 2001.

[25] C. H. Bischof and P. C. Hansen, "Structure-preserving and rank-revealing QR-factorizations," *SIAM Journal on Scientific and Statistical Computing*, vol. 12, no. 6, pp. 1332–1350, 1991.

[26] C. H. Bischof and G. Quintana-Ortí, "Algorithm 782: Codes for rank-revealing QR factorizations of dense matrices," *ACM Transactions on Mathematical Software*, vol. 24, no. 2, pp. 254–257, 1998.

[27] C. H. Bischof and G. Quintana-Ortí, "Computing rank-revealing QR factorizations of dense matrices," *ACM Transactions on Mathematical Software*, vol. 24, no. 2, pp. 226–253, 1998.

[28] P. Bonacich, "Power and centrality: A family of measures," *The American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[29] T. A. Boroson and T. R. Lauer, "Exploring the spectral space of low redshift QSOs," *The Astronomical Journal*, vol. 140, pp. 390–402, 2010.

[30] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," Technical report. Preprint: arXiv:0812.4293v2, 2008.

[31] C. Boutsidis, M. W. Mahoney, and P. Drineas, "Unsupervised feature selection for principal components analysis," in *Proceedings of the 14th Annual ACM SIGKDD Conference*, pp. 61–69, 2008.

[32] C. Boutsidis, M. W. Mahoney, and P. Drineas, "Unsupervised feature selection for the k-means clustering problem," in *Annual Advances in Neural Information Processing Systems 22: Proceedings of the 2009 Conference*, 2009.

[33] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the column subset selection problem," in *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 968–977, 2009.

[34] A. J. Bray and G. J. Rodgers, "Diffusion in a sparsely connected space: A model for glassy relaxation," *Physical Review B*, vol. 38, no. 16, pp. 11461–11470, 1988.

[35] M. E. Broadbent, M. Brown, and K. Penner, "Subset selection algorithms: Randomized vs. deterministic," *SIAM Undergraduate Research Online*, vol. 3, May 13 2010.

[36] R. J. Brunner, S. G. Djorgovski, T. A. Prince, and A. S. Szalay, "Massive datasets in astronomy," in *Handbook of Massive Data Sets*, (J. Abello, P. M. Pardalos, and M. G. C. Resende, eds.), pp. 931–979, Kluwer Academic Publishers, 2002.

[37] T. Budavári, V. Wild, A. S. Szalay, L. Dobos, and C.-W. Yip, "Reliable eigenspectra for new generation surveys," *Monthly Notices of the Royal Astronomical Society*, vol. 394, no. 3, pp. 1496–1502, 2009.

[38] P. Businger and G. H. Golub, "Linear least squares solutions by Householder transformations," *Numerische Mathematik*, vol. 7, pp. 269–276, 1965.

[39] E. Candes, L. Demanet, and L. Ying, "Fast computation of Fourier integral operators," *SIAM Journal on Scientific Computing*, vol. 29, no. 6, pp. 2464–2493, 2007.

[40] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[41] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[42] S. Chaillat and G. Biros, "FaIMS: A fast algorithm for the inverse medium problem with multiple frequencies and multiple sources for the scalar Helmholtz equation," Manuscript, 2010.

[43] D. Chakrabarti and C. Faloutsos, "Graph mining: Laws, generators, and algorithms," *ACM Computing Surveys*, vol. 38, no. 1, p. 2, 2006.

[44] T. F. Chan, "Rank revealing QR factorizations," *Linear Algebra and Its Applications*, vol. 88/89, pp. 67–82, 1987.

[45] T. F. Chan and P. C. Hansen, "Low-rank revealing QR factorizations," *Numerical Linear Algebra with Applications*, vol. 1, pp. 33–44, 1994.

[46] T. F. Chan and P. C. Hansen, "Computing truncated singular value decomposition least squares solutions by rank revealing QR-factorizations," *SIAM Journal on Scientific and Statistical Computing*, vol. 11, pp. 519–530, 1990.

[47] S. Chandrasekaran and I. C. F. Ipsen, "On rank-revealing factorizations," *SIAM Journal on Matrix Analysis and Applications*, vol. 15, pp. 592–622, 1994.

[48] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[49] S. Chatterjee and A. S. Hadi, "Influential observations, high leverage points, and outliers in linear regression," *Statistical Science*, vol. 1, no. 3, pp. 379–393, 1986.

[50] S. Chatterjee and A. S. Hadi, *Sensitivity Analysis in Linear Regression*. New York: John Wiley & Sons, 1988.

[51] S. Chatterjee, A. S. Hadi, and B. Price, *Regression Analysis by Example*. New York: John Wiley & Sons, 2000.

[52] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65–73, 1998.

[53] A. Civril and M. Magdon-Ismail, "Deterministic sparse column based matrix reconstruction via greedy approximation of SVD," in *Proceedings of the 19th Annual International Symposium on Algorithms and Computation*, pp. 414–423, 2008.

[54] A. Civril and M. Magdon-Ismail, "On selecting a maximum volume sub-matrix of a matrix and related problems," *Theoretical Computer Science*, vol. 410, pp. 4801–4811, 2009.

[55] K. L. Clarkson and D. P. Woodruff, "Numerical linear algebra in the streaming model," in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pp. 205–214, 2009.

[56] E. S. Coakley, V. Rokhlin, and M. Tygert, "A fast randomized algorithm for orthogonal projection," *SIAM Journal on Scientific Computing*, vol. 33, no. 2, pp. 849–868, 2011.

[57] A. J. Connolly and A. S. Szalay, "A robust classification of galaxy spectra: Dealing with noisy and incomplete data," *The Astronomical Journal*, vol. 117, no. 5, pp. 2052–2062, 1999.

[58] A. J. Connolly, A. S. Szalay, M. A. Bershady, A. L. Kinney, and D. Calzetti, "Spectral classification of galaxies: an orthogonal approach," *The Astronomical Journal*, vol. 110, no. 3, pp. 1071–1082, 1995.

[59] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[60] A. Dasgupta, R. Kumar, and T. Sarlós, "A sparse Johnson-Lindenstrauss transform," in *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pp. 341–350, 2010.

[61] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures and Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[62] A. d'Aspremont, "Subsampling algorithms for semidefinite programming," Technical Report. Preprint: arXiv:0803.1990, 2008.

[63] S. T. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[64] A. Deshpande and S. Vempala, "Adaptive sampling and fast low-rank matrix approximation," Technical Report TR06-042, Electronic Colloquium on Computational Complexity, March 2006.

[65] A. Deshpande and S. Vempala, "Adaptive sampling and fast low-rank matrix approximation," in *Proceedings of the 10th International Workshop on Randomization and Computation*, pp. 292–303, 2006.

[66] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, and A. N. Samukhin, "Spectra of complex networks," *Physical Review E*, vol. 68, p. 046109, 2003.

[67] N. R. Draper and D. M. Stoneman, "Testing for the inclusion of variables in linear regression by a randomisation technique," *Technometrics*, vol. 8, no. 4, pp. 695–699, 1966.

[68] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering large graphs via the singular value decomposition," *Machine Learning*, vol. 56, no. 1–3, pp. 9–33, 2004.

[69] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication," *SIAM Journal on Computing*, vol. 36, pp. 132–157, 2006.

[70] 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*, vol. 36, pp. 158–183, 2006.

[71] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition," *SIAM Journal on Computing*, vol. 36, pp. 184–206, 2006.

[72] P. Drineas, J. Lewis, and P. Paschou, "Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers," *PLoS ONE*, vol. 5, no. 8, no. 8, p. e11892, 2010.

[73] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, "Fast approximation of matrix coherence and statistical leverage," Technical report. Preprint: arXiv:1109.3843, 2011.

[74] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *Journal of Machine Learning Research*, vol. 6, pp. 2153–2175, 2005.

[75] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Sampling algorithms for ℓ_2 regression and applications," in *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136, 2006.

[76] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error CUR matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, pp. 844–881, 2008.

[77] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, "Faster least squares approximation," *Numerische Mathematik*, vol. 117, no. 2, pp. 219–249, 2010.

[78] B. Engquist and O. Runborg, "Wavelet-based numerical homogenization with applications," in *Multiscale and Multiresolution Methods: Theory and Applications*, LNCSE, (T. J. Barth, T. F. Chan, and R. Haimes, eds.), pp. 97–148, Springer, 2001.

[79] B. Engquist and L. Ying, "Fast directional multilevel algorithms for oscillatory kernels," *SIAM Journal on Scientific Computing*, vol. 29, pp. 1710–1737, 2007.

[80] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hungar. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[81] S. Eriksson-Bique, M. Solbrig, M. Stefanelli, S. Warkentin, R. Abbey, and I. C. F. Ipsen, "Importance sampling for a Monte Carlo matrix multiplication algorithm, with application to information retrieval," *SIAM Journal on Scientific Computing*, vol. 33, no. 4, pp. 1689–1706, 2011.

[82] S. N. Evangelou, "A numerical study of sparse random matrices," *Journal of Statistical Physics*, vol. 69, no. 1-2, pp. 361–383, 1992.

[83] I. J. Farkas, I. Derényi, A.-L. Barabási, and T. Vicsek, "Spectra of "real-world" graphs: Beyond the semicircle law," *Physical Review E*, vol. 64, p. 026704, 2001.

[84] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proceedings of the 20th International Conference on Machine Learning*, pp. 186–193, 2003.

[85] S. R. Folkes, O. Lahav, and S. J. Maddox, "An artificial neural network approach to the classification of galaxy spectra," *Mon. Not. R. Astron. Soc.*, vol. 283, no. 2, pp. 651–665, 1996.

[86] L. V. Foster, "Rank and null space calculations using matrix decomposition without column interchanges," *Linear Algebra and Its Applications*, vol. 74, pp. 47–71, 1986.

[87] D. Fradkin and D. Madigan, "Experiments with random projections for machine learning," in *Proceedings of the 9th Annual ACM SIGKDD Conference*, pp. 517–522, 2003.

[88] P. Frankl and H. Maehara, "The Johnson-Lindenstrauss lemma and the sphericity of some graphs," *Journal of Combinatorial Theory Series A*, vol. 44, no. 3, pp. 355–362, 1987.

[89] A. Frieze and R. Kannan, "Quick approximation to matrices and applications," *Combinatorica*, vol. 19, no. 2, pp. 175–220, 1999.

[90] A. Frieze, R. Kannan, and S. Vempala, "Fast Monte-Carlo algorithms for finding low-rank approximations," *Journal of the ACM*, vol. 51, no. 6, pp. 1025–1041, 2004.

[91] Z. Füredi and J. Komlós, "The eigenvalues of random symmetric matrices," *Combinatorica*, vol. 1, no. 3, pp. 233–241, 1981.

[92] Y. V. Fyodorov and A. D. Mirlin, "Localization in ensemble of sparse random matrices," *Physical Review Letters*, vol. 67, pp. 2049–2052, 1991.

[93] S. Georgiev and S. Mukherjee, Unpublished results. 2011.

[94] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 937–947, 2010.

[95] N. Goel, G. Bebis, and A. Nefian, "Face recognition experiments with random projection," *Proceedings of the SPIE*, vol. 5779, pp. 426–437, 2005.

[96] K.-I. Goh, B. Kahng, and D. Kim, "Spectra and eigenvectors of scale-free networks," *Physical Review E*, vol. 64, p. 051903, 2001.

[97] G. H. Golub, M. W. Mahoney, P. Drineas, and L.-H. Lim, "Bridging the gap between numerical linear algebra, theoretical computer science, and data applications," *SIAM News*, vol. 39, no. 8, October 2006.

[98] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, 1996.

[99] S. A. Goreinov and E. E. Tyrtyshnikov, "The maximum-volume concept in approximation by low-rank matrices," *Contemporary Mathematics*, vol. 280, pp. 47–51, 2001.

[100] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin, "A theory of pseudoskeleton approximations," *Linear Algebra and Its Applications*, vol. 261, pp. 1–21, 1997.

[101] S. J. Gould, *The Mismeasure of Man*. New York: W. W. Norton and Company, 1996.

[102] L. Grasedyck and W. Hackbusch, "Construction and arithmetics of H-matrices," *Computing*, vol. 70, no. 4, pp. 295–334, 2003.

[103] L. Greengard and V. Rokhlin, "A fast algorithm for particle simulations," *Journal of Computational Physics*, vol. 73, no. 2, pp. 325–348, 1987.

[104] L. Greengard and V. Rokhlin, "A new version of the fast multipole method for the Laplace equation in three dimensions," *Acta Numerica*, vol. 6, pp. 229–269, 1997.

[105] M. Gu and S. C. Eisenstat, "Efficient algorithms for computing a strong rank-revealing QR factorization," *SIAM Journal on Scientific Computing*, vol. 17, pp. 848–869, 1996.

[106] N. Halko, P.-G. Martinsson, Y. Shkolnisky, and M. Tygert, "An algorithm for the principal component analysis of large data sets," Technical report. Preprint: arXiv:1007.5510, 2010.

[107] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, no. 2, pp. 217–288, 2011.

[108] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*. London and New York: Chapman and Hall, 1964.

[109] S. Har-Peled, "Low rank matrix approximation in linear time," Manuscript, January 2006.

[110] D. C. Hoaglin and R. E. Welsch, "The hat matrix in regression and ANOVA," *The American Statistician*, vol. 32, no. 1, pp. 17–22, 1978.

[111] Y. P. Hong and C. T. Pan, "Rank-revealing QR factorizations and the singular value decomposition," *Mathematics of Computation*, vol. 58, pp. 213–232, 1992.

[112] B. D. Horne and N. J. Camp, "Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation," *Genetic Epidemiology*, vol. 26, no. 1, pp. 11–21, 2004.

[113] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pp. 604–613, 1998.

[114] A. Javed, P. Drineas, M. W. Mahoney, and P. Paschou, "Efficient genomewide selection of PCA-correlated tSNPs for genotype imputation," *Annals of Human Genetics*, vol. 75, no. 6, pp. 707–722, 2011.

[115] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipshitz mapping into Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[116] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001.

[117] E. A. Jonckheere, M. Lou, J. Hespanha, and P. Barooah, "Effective resistance of Gromov-hyperbolic graphs: Application to asymptotic sensor network problems," in *Proceedings of the 46th IEEE Conference on Decision and Control*, pp. 1453–1458, 2007.

[118] D. M. Kane and J. Nelson, "A derandomized sparse Johnson-Lindenstrauss transform," Technical Report. Preprint: arXiv:1006.3585, 2010.

[119] D. M. Kane and J. Nelson, "Sparser Johnson-Lindenstrauss transforms," Technical Report. Preprint: arXiv:1012.1577, 2010.

[120] S. Kaski, "Dimensionality reduction by random mapping: fast similarity computation for clustering," in *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks*, pp. 413–418, 1998.

[121] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[122] R. Kühn, "Spectra of sparse random matrices," *J. of Physics A: Math. and Theor*, vol. 41, p. 295002, 2008.

[123] S. Kumar, M. Mohri, and A. Talwalkar, "On sampling-based approximate spectral decomposition," in *Proceedings of the 26th International Conference on Machine Learning*, pp. 553–560, 2009.

[124] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling techniques for the Nyström method," in *Proceedings of the 12th Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 304–311, 2009.

[125] F. G. Kuruvilla, P. J. Park, and S. L. Schreiber, "Vector algebra in the analysis of genome-wide expression data," *Genome Biology*, vol. 3, no. 3, pp. research0011.1–0011.11, 2002.

[126] M. Li, J. T. Kwok, and B.-L. Lu, "Making large-scale Nyström approximation possible," in *Proceedings of the 27th International Conference on Machine Learning*, pp. 631–638, 2010.

[127] E. Liberty, N. Ailon, and A. Singer, "Dense fast random projections and lean Walsh transforms," in *Proceedings of the 12th International Workshop on Randomization and Computation*, pp. 512–522, 2008.

[128] E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert, "Randomized algorithms for the low-rank approximation of matrices," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 51, pp. 20167–20172, 2007.

[129] L. Lin, J. Lu, and L. Ying, "Fast construction of hierarchical matrix representation from matrix-vector multiplication," *Journal of Computational Physics*, vol. 230, pp. 4071–4087, 2011.

[130] L. Lin, C. Yang, J. C. Meza, J. Lu, L. Ying, and W. E. SelInv, "An algorithm for selected inversion of a sparse symmetric matrix," *ACM Transactions on Mathematical Software*, vol. 37, no. 4, p. 40, 2011.

[131] Z. Lin and R. B. Altman, "Finding haplotype tagging SNPs by use of principal components analysis," *American Journal of Human Genetics*, vol. 75, pp. 850–861, 2004.

[132] L. Mackey, A. Talwalkar, and M. I. Jordan, "Divide-and-conquer matrix factorization," Technical report. Preprint: arXiv:1107.0789, 2011.

[133] D. Madgwick, O. Lahav, K. Taylor, and the 2dFGRS Team, "Parameterisation of galaxy spectra in the 2dF galaxy redshift survey," in *Mining the Sky: Proceedings of the MPA/ESO/MPE Workshop, ESO Astrophysics Symposia*, pp. 331–336, 2001.

[134] M. Magdon-Ismail, "Row sampling for matrix algorithms via a non-commutative Bernstein bound," Technical report. Preprint: arXiv:1008.0587, 2010.

[135] A. Magen and A. Zouzias, "Low rank matrix-valued Chernoff bounds and approximate matrix multiplication," in *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1422–1436, 2011.

[136] M. W. Mahoney, "Computation in large-scale scientific and Internet data applications is a focus of MMDS 2010," Technical report. Preprint: arXiv:1012.4231, 2010.

[137] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 697–702, 2009.

[138] M. W. Mahoney, L.-H. Lim, and G. E. Carlsson, "Algorithmic and statistical challenges in modern large-scale data analysis are the focus of MMDS 2008," Technical report. Preprint: arXiv:0812.3702, 2008.

[139] M. Mahoney, M. Maggioni, and P. Drineas, "Tensor-CUR decompositions for tensor-based data," in *Proceedings of the 12th Annual ACM SIGKDD Conference*, pp. 327–336, 2006.

[140] P.-G. Martinsson, "Rapid factorization of structured matrices via randomized sampling," Technical Report. Preprint: arXiv:0806.2339, 2008.

[141] P.-G. Martinsson, V. Rokhlin, and M. Tygert, "A randomized algorithm for the decomposition of matrices," *Applied and Computational Harmonic Analysis*, vol. 30, pp. 47–68, 2011.

[142] J. Matousek, "On variants of the Johnson–Lindenstrauss lemma," *Random Structures and Algorithms*, vol. 33, no. 2, pp. 142–156, 2008.

[143] R. C. McGurk, A. E. Kimball, and Z. Ivezić, "Principal component analysis of SDSS stellar spectra," *The Astronomical Journal*, vol. 139, pp. 1261–1268, 2010.

[144] X. Meng, M. A. Saunders, and M. W. Mahoney, "LSRN: A parallel iterative solver for strongly over- or under-determined systems," Technical report. Preprint: arXiv:arXiv:1109.5981, 2011.

[145] Z. Meng, D. V. Zaykin, C. F. Xu, M. Wagner, and M. G. Ehm, "Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes," *American Journal of Human Genetics*, vol. 73, no. 1, pp. 115–130, 2003.

[146] P. Menozzi, A. Piazza, and L. Cavalli-Sforza, "Synthetic maps of human gene frequencies in Europeans," *Science*, vol. 201, no. 4358, pp. 786–792, 1978.

[147] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, vol. 21, pp. 1087–1092, 1953.

[148] A. D. Mirlin and Y. V. Fyodorov, "Universality of level correlation function of sparse random matrices," *J. Phys. A: Math. Gen*, vol. 24, pp. 2273–2286, 1991.

[149] M. Mitrović and B. Tadić, "Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities," *Physical Review E*, vol. 80, p. 026123, 2009.

[150] R. Motwani and P. Raghavan, *Randomized Algorithms.* New York: Cambridge University Press, 1995.

[151] S. Muthukrishnan, *Data Streams: Algorithms and Applications.* Boston: Foundations and Trends in Theoretical Computer Science. Now Publishers Inc, 2005.

[152] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, pp. 39–54, 2005.

[153] N. H. Nguyen, T. T. Do, and T. D. Tran, "A fast and efficient algorithm for low-rank approximation of a matrix," in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pp. 215–224, 2009.

[154] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante, "Genes mirror geography within Europe," *Nature*, vol. 456, pp. 98–101, 2008.

[155] C. C. Paige and M. A. Saunders, "Algorithm 583: LSQR: Sparse linear equations and least-squares problems," *ACM Transactions on Mathematical Software*, vol. 8, no. 2, pp. 195–209, 1982.

[156] C.-T. Pan, "On the existence and computation of rank-revealing LU factorizations," *Linear Algebra and Its Applications*, vol. 316, pp. 199–222, 2000.

[157] C. T. Pan and P. T. P. Tang, "Bounds on singular values revealed by QR factorizations," *BIT Numerical Mathematics*, vol. 39, pp. 740–756, 1999.

[158] F. Pan, X. Zhang, and W. Wang, "CRD: Fast co-clustering on large datasets utilizing sampling-based matrix decomposition," in *Proceedings of the 34th SIGMOD international conference on Management of data*, pp. 173–184, 2008.

[159] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, 2000.

[160] P. Parker, P. J. Wolfe, and V. Tarok, "A signal processing application of randomized low-rank approximations," in *Proceedings of the 13th IEEE Workshop on Statistical Signal Processing*, pp. 345–350, 2005.

[161] P. Paschou, P. Drineas, J. Lewis, C. M. N. D. A. Nickerson, J. D. Smith, P. M. Ridker, D. I. Chasman, R. M. Krauss, and E. Ziv, "Tracing sub-structure in the European American population with PCA-informative markers," *PLoS Genetics*, vol. 4, no. 7, p. e1000114, 2008.

[162] P. Paschou, J. Lewis, A. Javed, and P. Drineas, "Ancestry informative markers for fine-scale individual assignment to worldwide populations," *Journal of Medical Genetics*, 2010. doi:10.1136/jmg.2010.078212.

[163] P. Paschou, M. W. Mahoney, A. Javed, J. R. Kidd, A. J. Pakstis, S. Gu, K. K. Kidd, and P. Drineas, "Intra- and interpopulation genotype reconstruction from tagging SNPs," *Genome Research*, vol. 17, no. 1, pp. 96–107, 2007.

[164] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas, "PCA-correlated SNPs for structure identification in worldwide human populations," *PLoS Genetics*, vol. 3, pp. 1672–1686, 2007.

[165] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genetics*, vol. 2, no. 12, pp. 2074–2093, 2006.

[166] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the 8th Annual ACM SIGKDD Conference*, pp. 61–70, 2002.

[167] G. J. Rodgers and A. J. Bray, "Density of states of a sparse random matrix," *Physical Review B*, vol. 37, no. 7, pp. 3557–3562, 1988.

[168] V. Rokhlin, A. Szlam, and M. Tygert, "A randomized algorithm for principal component analysis," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1100–1124, 2009.

[169] V. Rokhlin and M. Tygert, "A fast randomized algorithm for overdetermined linear least-squares regression," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 36, pp. 13212–13217, 2008.

[170] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by local linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[171] M. Rudelson, "Random vectors in the isotropic position," *Journal of Functional Analysis*, vol. 164, no. 1, pp. 60–72, 1999.

[172] M. Rudelson and R. Vershynin, "Sampling from large matrices: An approach through geometric functional analysis," *Journal of the ACM*, vol. 54, no. 4, p. Article 21, 2007.

[173] Y. Saad, J. R. Chelikowsky, and S. M. Shontz, "Numerical methods for electronic structure calculations of materials," *SIAM Review*, vol. 52, no. 1, pp. 3–54, 2010.

[174] T. Sarlós, "Improved approximation algorithms for large matrices via random projections," in *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 143–152, 2006.

[175] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee, "Spectral methods for dimensionality reduction," in *Semisupervised Learning*, (O. Chapelle, B. Schoelkopf, and A. Zien, eds.), pp. 293–308, MIT Press, 2006.

[176] B. Savas and I. Dhillon, "Clustered low rank approximation of graphs in information science applications," in *Proceedings of the 11th SIAM International Conference on Data Mining*, 2011.

[177] D. N. Spendley and P. J. Wolfe, "Adaptive beamforming using fast low-rank covariance matrix approximations," in *Proceedings of the IEEE Radar Conference*, pp. 1–5, 2008.

[178] D. A. Spielman and N. Srivastava, "Graph sparsification by effective resistances," in *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 563–568, 2008.

[179] G. Stewart, "Four algorithms for the efficient computation of truncated QR approximations to a sparse matrix," *Numerische Mathematik*, vol. 83, pp. 313–323, 1999.

[180] G. Strang, *Linear Algebra and Its Applications*. Harcourth Brace Jovanovich, 1988.

[181] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, "Less is more: Compact matrix decomposition for large sparse graphs," in *Proceedings of the 7th SIAM International Conference on Data Mining*, 2007.

[182] A. Talwalkar, S. Kumar, and H. Rowley, "Large-scale manifold learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[183] A. Talwalkar and A. Rostamizadeh, "Matrix coherence and the Nyström method," in *Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence*, 2010.

[184] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[185] The International HapMap Consortium, "The International HapMap Project," *Nature*, vol. 426, pp. 789–796, 2003.

[186] The International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, pp. 1299–1320, 2005.

[187] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, 2007.

[188] H. Tong, S. Papadimitriou, J. Sun, P. S. Yu, and C. Faloutsos, "Colibri: Fast mining of large static and dynamic graphs," in *Proceedings of the 14th Annual ACM SIGKDD Conference*, pp. 686–694, 2008.

[189] P. F. Velleman and R. E. Welsch, "Efficient computing of regression diagnostics," *The American Statistician*, vol. 35, no. 4, pp. 234–242, 1981.

[190] S. Venkatasubramanian and Q. Wang, "The Johnson-Lindenstrauss transform: An empirical study," in *ALENEX11: Workshop on Algorithms Engineering and Experimentation*, pp. 164–173, 2011.

[191] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray*

Data Analysis, (D. P. Berrar, W. Dubitzky, and M. Granzow, eds.), pp. 91–109, Kluwer, 2003.

[192] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Annual Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pp. 682–688, 2001.

[193] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, "A fast randomized algorithm for the approximation of matrices," *Applied and Computational Harmonic Analysis*, vol. 25, no. 3, pp. 335–366, 2008.

[194] C. W. Yip, A. J. Connolly, A. S. Szalay, T. Budavári, M. SubbaRao, J. A. Frieman, R. C. Nichol, A. M. Hopkins, D. G. York, S. Okamura, J. Brinkmann, I. Csabai, A. R. Thakar, M. Fukugita, and Z. Ivezić, "Distributions of galaxy spectral types in the Sloan Digital Sky Survey," *The Astronomical Journal*, vol. 128, no. 2, pp. 585–609, 2004.

[195] C. W. Yip, A. J. Connolly, D. E. Vanden Berk, Z. Ma, J. A. Frieman, M. SubbaRao, A. S. Szalay, G. T. Richards, P. B. Hall, D. P. Schneider, A. M. Hopkins, J. Trump, and J. Brinkmann, "Spectral classification of quasars in the Sloan Digital Sky Survey: Eigenspectra, redshift, and luminosity effects," *The Astronomical Journal*, vol. 128, no. 6, pp. 2603–2630, 2004.

[196] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[197] K. Zhang and J. T. Kwok, "Clustered Nyström method for large scale manifold learning and dimension reduction," *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1576–1587, 2010.

[198] K. Zhang, I. W. Tsang, and J. T. Kwok, "Improved Nyström low-rank approximation and error analysis," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 1232–1239, 2008.