Generalized Low Rank Models

Madeleine Udell
Operations Research and Information Engineering
Cornell University
udell@cornell.edu

Corinne Horn
Electrical Engineering
Stanford University
cehorn@stanford.edu

Reza Zadeh
Computational and Mathematical Engineering
Stanford University
rezab@stanford.edu

Stephen Boyd
Electrical Engineering
Stanford University
boyd@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
United States
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. Udell, C. Horn, R. Zadeh and S. Boyd. *Generalized Low Rank Models.* Foundations and Trends® in Machine Learning, vol. 9, no. 1, pp. 1–118, 2016.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-141-2

Foundations and Trends® in Machine Learning
Volume 9, Issue 1, 2016

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including 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, 2016, Volume 9, 6 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. 9, No. 1 (2016) 1–118

DOI: 10.1561/2200000055

Generalized Low Rank Models

Madeleine Udell
Operations Research and Information Engineering
Cornell University
udell@cornell.edu

Corinne Horn
Electrical Engineering
Stanford University
cehorn@stanford.edu

Reza Zadeh
Computational and Mathematical Engineering
Stanford University
rezab@stanford.edu

Stephen Boyd
Electrical Engineering
Stanford University
boyd@stanford.edu

Contents

Abstract

Principal components analysis (PCA) is a well-known technique for approximating a tabular data set by a low rank matrix. Here, we extend the idea of PCA to handle arbitrary data sets consisting of numerical, Boolean, categorical, ordinal, and other data types. This framework encompasses many well known techniques in data analysis, such as nonnegative matrix factorization, matrix completion, sparse and robust PCA, k-means, k-SVD, and maximum margin matrix factorization. The method handles heterogeneous data sets, and leads to coherent schemes for compressing, denoising, and imputing missing entries across all data types simultaneously. It also admits a number of interesting interpretations of the low rank factors, which allow clustering of examples or of features. We propose several parallel algorithms for fitting generalized low rank models, and describe implementations and numerical results.

M. Udell, C. Horn, R. Zadeh and S. Boyd. *Generalized Low Rank Models*. Foundations and Trends® in Machine Learning, vol. 9, no. 1, pp. 1–118, 2016.
DOI: 10.1561/2200000055.

1

Introduction

In applications of machine learning and data mining, one frequently encounters large collections of high dimensional data organized into a table. Each row in the table represents an example, and each column a feature or attribute. These tables may have columns of different (sometimes, non-numeric) types, and often have many missing entries.

For example, in medicine, the table might record patient attributes or lab tests: each row of the table lists test or survey results for a particular patient, and each column corresponds to a distinct test or survey question. The values in the table might be numerical (3.14), Boolean (yes, no), ordinal (never, sometimes, always), or categorical (A, B, O). Tests not administered or questions left blank result in missing entries in the data set. Other examples abound: in finance, the table might record known characteristics of companies or asset classes; in social science settings, it might record survey responses; in marketing, it might record known customer characteristics and purchase history.

Exploratory data analysis can be difficult in this setting. To better understand a complex data set, one would like to be able to visualize archetypical examples, to cluster examples, to find correlated features, to fill in (impute) missing entries, and to remove (or simply identify) spurious, anomalous, or noisy data points. This paper introduces a templated method to enable these analyses even on large data sets with heterogeneous values and with many missing entries. Our approach will be to embed both the rows (examples) and columns (features) of the table into the same low dimensional vector space. These low dimensional vectors can then be plotted, clustered, and used to impute missing entries or identify anomalous ones.

If the data set consists only of numerical (real-valued) data, then a simple and well-known technique to find this embedding is Principal Components Analysis (PCA). PCA finds a low rank matrix that minimizes the approximation error, in the least-squares sense, to the original data set. A factorization of this low rank matrix embeds the original high dimensional features into a low dimensional space. Extensions of PCA can handle missing data values, and can be used to impute missing entries.

Here, we extend PCA to approximate an arbitrary data set by replacing the least-squares error used in PCA with a loss function that is appropriate for the given data type. Another extension beyond PCA is to add regularization on the low dimensional factors to impose or encourage some structure, such as sparsity or nonnegativity, in the low dimensional factors. In this paper we use the term *generalized low rank model* (GLRM) to refer to the problem of approximating a data set as a product of two low dimensional factors by minimizing an objective function. The objective will consist of a loss function on the approximation error together with regularization of the low dimensional factors. With these extensions of PCA, the resulting low rank representation of the data set still produces a low dimensional embedding of the data set, as in PCA.

Many of the low rank modeling problems we must solve will be familiar. We recover an optimization formulation of nonnegative matrix factorization, matrix completion, sparse and robust PCA, k-means, k-SVD, and maximum margin matrix factorization, to name just a few.The scope of the problems we consider, however, is more broad, encompassing many different combinations of loss function and regularizer. A few of the choices we consider are shown in Tables A.1 and

A.2 of Appendix A for reference; all of these are discussed in detail later in the paper.

These low rank approximation problems are not convex, and in general cannot be solved globally and efficiently. There are a few exceptional problems that are known to have convex relaxations which are tight under certain conditions, and hence are efficiently (globally) solvable under these conditions. However, all of these approximation problems can be heuristically (locally) solved by methods that alternate between updating the two factors in the low rank approximation. Each step involves either a convex problem, or a nonconvex problem that is simple enough that we can solve it exactly. While these alternating methods need not find the globally best low rank approximation, they are often very useful and effective for the original data analysis problem.

1.1 Previous work

Unified views of matrix factorization. We are certainly not the first to note that matrix factorization algorithms may be viewed in a unified framework, parametrized by a small number of modeling decisions. The first instance we find in the literature of this unified view appeared in a paper by Collins, Dasgupta, and Schapire, [29], extending PCA to use loss functions derived from any probabilistic model in the exponential family. Gordon's Generalized^2 Linear^2 models [53] extended the framework to loss functions derived from the generalized Bregman divergence of any convex function, which includes models such as Independent Components Analysis (ICA). Srebro's 2004 PhD thesis [133] extended the framework to other loss functions, including hinge loss and KL-divergence loss, and to other regularizers, including the nuclear norm and max-norm. Similarly, Chapter 8 in Tropp's 2004 PhD thesis [144] explored a number of new regularizers, presenting a range of clustering problems as matrix factorization problems with constraints, and anticipated the k-SVD algorithm [4]. Singh and Gordon [129] offered a complete view of the state of the literature on matrix factorization in Table 1 of their 2008 paper, and noted that by changing the loss function and regularizer, one may recover algorithms including PCA, weighted PCA, k-means, k-medians, ℓ_1 SVD, probabilistic latent semantic indexing (pLSI), nonnegative matrix factorization with ℓ_2 or KL-divergence loss, exponential family PCA, and MMMF. Witten et al. introduced the statistics community to sparsity-inducing matrix factorization in a 2009 paper on penalized matrix decomposition, with applications to sparse PCA and canonical correlation analysis [155]. Recently, Markovsky's monograph on low rank approximation [97] reviewed some of this literature, with a focus on applications in system, control, and signal processing. The GLRMs discussed in this paper include all of these models, and many more.

Heterogeneous data. Many authors have proposed the use of low rank models as a tool for integrating heterogeneous data. The earliest example of this approach is canonical correlation analysis, developed by Hotelling [63] in 1936 to understand the relations between two sets of variates in terms of the eigenvectors of their covariance matrix. This approach was extended by Witten et al. [155] to encourage structured (*e.g.*, sparse) factors. In the 1970s, De Leeuw et al. proposed the use of low rank models to fit data measured in nominal, ordinal and cardinal levels [37]. More recently, Goldberg et al. [52] used a low rank model to perform transduction (*i.e.*, multi-label learning) in the presence of missing data by fitting a low rank model to the features and the labels simultaneously. Low rank models have also been used to embed image, text and video data into a common low dimensional space [54], and have recently come into vogue in the natural language processing community as a means to embed words and documents into a low dimensional vector space [99, 100, 112, 136].

Algorithms. In general, it can be computationally hard to find the global optimum of a generalized low rank model. For example, it is NP-hard to compute an exact solution to k-means [43], nonnegative matrix factorization [149], and weighted PCA and matrix completion [50], all of which are special cases of low rank models.

However, there are many (efficient) ways to go about *fitting* a low rank model, by which we mean finding a good model with a small objective value. The resulting model may or may not be the global solution of the low rank optimization problem. We distinguish a model fit in this way from the *solution* to an optimization problem, which always refers to the global solution.

The matrix factorization literature presents a wide variety of methods to fit low rank models in a variety of special cases. For example, there are variants on alternating minimization (with alternating least squares as a special case) [37, 158, 141, 35, 36], alternating Newton methods [53, 129], (stochastic or incremental) gradient descent [75, 88, 104, 119, 10, 159, 118], conjugate gradients [120, 134], expectation minimization (EM) (or "soft-impute") methods [142, 134, 98, 60], multiplicative updates [85], and convex relaxations to semidefinite programs [135, 46, 117, 48].

Generally, expectation minimization, which proceeds by iteratively imputing missing entries in the matrix and solving the fully observed problem, has been found to underperform relative to other methods [129]. However, when used in conjunction with computational tricks exploiting a particular problem structure, such as Gram matrix caching, these methods can still work extremely well [60].

Semidefinite programming becomes computationally intractable for very large (or even just large) scale problems [120]. However, a theoretical analysis of optimality conditions for rank-constrained semidefinite programs [20] has led to a few algorithms for semidefinite programming based on matrix factorization [19, 1, 70] which guarantee global optimality and converge quickly if the global solution to the problem is exactly low rank. Fast approximation algorithms for rank-constrained semidefinite programs have also been developed [127].

Recently, there has been a resurgence of interest in methods based on alternating minimization, as numerous authors have shown that alternating minimization (suitably initialized, and under a few technical assumptions) provably converges to the global minimum for a range of problems including matrix completion [72, 66, 58], robust PCA [103], and dictionary learning [2].

Gradient descent methods are often preferred for extremely large scale problems since these methods parallelize naturally in both shared memory and distributed memory architectures. See [118, 159] and references therein for some recent innovative approaches to speeding up stochastic gradient descent for matrix factorization by eliminating locking and reducing interprocess communication. These stochastic nonlocking methods often run faster than their deterministic counterparts; and for the matrix completion problem in particular, these methods can be shown to provably converge to the global minimum under the same conditions required for alternating minimization [38].

Contributions. The present paper differs from previous work in a number of ways. We are consistently concerned with the *meaning* of applying these different loss functions and regularizers to approximate a data set. The generality of our view allows us to introduce a number of loss functions and regularizers that have not previously been considered. Moreover, our perspective enables us to extend these ideas to arbitrary data sets, rather than just matrices of real numbers.

A number of new considerations emerge when considering the problem so broadly. First, we must face the problem of comparing approximation errors across data of different types. For example, we must choose a scaling to trade off the loss due to a misclassification of a categorical value with an error of 0.1 (say) in predicting a real value.

Second, we require algorithms that can handle the full gamut of losses and regularizers, which may be smooth or nonsmooth, finite or infinite valued, with arbitrary domain. This work is the first to consider these problems in such generality, and therefore also the first to wrestle with the algorithmic consequences. Below, we give a number of algorithms appropriate for this setting, including many that have not been previously proposed in the literature. Our algorithms are all based on alternating minimization and variations on alternating minimization that are more suitable for large scale data and can take advantage of parallel computing resources.

These algorithms for fitting *any* GLRM are particularly useful for interactive data analysis: a practitioner can mix and match different loss functions and regularizers, and test which combinations provide the best fit to the data, without having to identify a different method to fit each particular model. We present a few software packages designed for this purpose, with interfaces in Julia, R, Java, Python, and Scala, in §9.

Finally, we present some new results on some old problems. For example, in Appendix A.1, we derive a formula for the solution to quadratically regularized PCA, and show that quadratically regularized PCA has no local nonglobal minima; and in §7.6 we show how to certify (in some special cases) that a model is a global solution of a GLRM.

1.2 Organization

The organization of this paper is as follows. In §2 we first recall some properties of PCA and its common variations to familiarize the reader with our notation. We then generalize the regularization on the low dimensional factors in §3, and the loss function on the approximation error in §4. Returning to the setting of heterogeneous data, we extend these dimensionality reduction techniques to abstract data types in §5 and to multi-dimensional loss functions in §6. Finally, we address algorithms for fitting GLRMs in §7, discuss a few practical considerations in choosing a GLRM for a particular problem in §8, and describe some implementations of the algorithms that we have developed in §9.

References

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *The Journal of Machine Learning Research*, 10:803–826, 2009.

[2] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013.

[3] P. K. Agarwal and N. H. Mustafa. k-means projective clustering. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 155–165. ACM, 2004.

[4] M. Aharon, M. Elad, and A. Bruckstein. k-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[5] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[6] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

[7] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

[8] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010:1–38, 2011.

[9] J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

[10] V. Bittorf, B. Recht, C. Ré, and J. A. Tropp. Factoring nonnegative matrices with linear programs. *Advances in Neural Information Processing Systems*, 25:1223–1231, 2012.

[11] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, pages 1–36, 2013.

[12] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization: theory and examples*, volume 3. Springer Science & Business Media, 2010.

[13] R. Boyd, B. Drake, D. Kuang, and H. Park. Smallk is a C++/Python high-performance software library for nonnegative matrix factorization (NMF) and hierarchical and flat clustering using the NMF; current version 1.2.0. `http://smallk.github.io/`, June 2014.

[14] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic. Accuracy at the top. In *Advances in Neural Information Processing Systems*, pages 962–970, 2012.

[15] S. Boyd and J. Mattingley. Branch and bound methods. *Lecture notes for EE364b, Stanford University*, 2003.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[17] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[18] S. Boyd, L. Xiao, and A. Mutapcic. Subgradient methods. *Lecture notes for EE364b, Stanford University*, 2003.

[19] S. Burer and R. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[20] S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103, 2005.

[21] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[22] E. Candès and Y. Plan. Matrix completion with noise. *CoRR*, abs/0903.3131, 2009.

[23] E. Candès and B. Recht. Exact matrix completion via convex optimization. *CoRR*, abs/0805.4471, 2008.

[24] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[25] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[26] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2014.

[27] J. Chen and A. Edelman. Parallel prefix polymorphism permits parallelization, presentation & proof. *arXiv preprint arXiv:1410.6449*, 2014.

[28] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[29] M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Processing Systems*, volume 13, page 23, 2001.

[30] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[31] A. Damle and Y. Sun. Random projections for non-negative matrix factorization. *arXiv preprint arXiv:1405.4275*, 2014.

[32] D. Das and S. Das. Quadratic programing solver for non-negative matrix factorization with spark. In *Spark Summit 2014*, 2014.

[33] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *Advances in Neural Information Processing Systems*, volume 16, pages 41–48, 2004.

[34] M. Davenport, Y. Plan, E. Berg, and M. Wootters. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2012.

[35] J. De Leeuw. The Gifi system of nonlinear multivariate analysis. *Data analysis and informatics*, 3:415–424, 1984.

[36] J. De Leeuw and P. Mair. Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, pages 1–30, 2009.

[37] J. De Leeuw, F. Young, and Y. Takane. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4):471–503, 1976.

[38] C. De Sa, K. Olukotun, and C. Ré. Global convergence of stochastic gradient descent for some nonconvex matrix problems. *CoRR*, abs/1411.1134, 2014.

[39] S. Diamond, E. Chu, and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization, version 0.2. `http://cvxpy.org/`, May 2014.

[40] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *CoRR*, cs.AI/9501101, 1995.

[41] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 126–135. ACM, 2006.

[42] A. Dinno. Implementing Horn's parallel analysis for principal component analysis and factor analysis. *Stata Journal*, 9(2):291, 2009.

[43] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.

[44] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[45] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009*, pages 2790–2797. IEEE, 2009.

[46] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *Proceedings of the 2004 American Control Conference (ACC)*, volume 4, pages 3273–3278. IEEE, 2004.

[47] C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[48] W. Fithian and R. Mazumder. Scalable convex methods for flexible low-rank matrix modeling. *arXiv preprint arXiv:1308.4211*, 2013.

[49] N. Gillis. *Nonnegative matrix factorization: Complexity, algorithms and applications*. PhD thesis, UCL, 2011.

[50] N. Gillis and F. Glineur. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1149–1165, 2011.

[51] Nicolas Gillis and François Glineur. A continuous characterization of the maximum-edge biclique problem. *Journal of Global Optimization*, 58(3):439–464, 2014.

[52] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems*, pages 757–765, 2010.

[53] G. J. Gordon. Generalized^2 linear^2 models. In *Advances in Neural Information Processing Systems*, pages 577–584, 2002.

[54] A. Gress and I. Davidson. A flexible framework for projecting heterogeneous data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1169–1178, New York, NY, USA, 2014. ACM.

[55] S. Gunasekar, A. Acharya, N. Gaur, and J. Ghosh. Noisy matrix completion using alternating minimization. In *Machine Learning and Knowledge Discovery in Databases*, pages 194–209. Springer, 2013.

[56] M. Gupta, S. Bengio, and J. Weston. Training highly multiclass classifiers. *The Journal of Machine Learning Research*, 15(1):1461–1492, 2014.

[57] N. Halko, P.-G. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[58] M. Hardt. On the provable convergence of alternating minimization for matrix completion. *arXiv preprint arXiv:1312.0925*, 2013.

[59] M. Hardt and M. Wootters. Fast matrix completion without the condition number. *arXiv preprint arXiv:1407.4070*, 2014.

[60] T. Hastie, R. Mazumder, J. Lee, and R. Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *arXiv*, 2014.

[61] J. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[62] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

[63] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3–4):321–377, 1936.

[64] Z. Huang and M. Ng. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, 7(4):446–452, 1999.

[65] P. Huber. *Robust Statistics*. Wiley, New York, 1981.

[66] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM Symposium on the Theory of Computing*, pages 665–674. ACM, 2013.

[67] I. Jolliffe. *Principal component analysis*. Springer, 1986.

[68] J. Josse and S. Wager. Stable autoencoding: A flexible framework for regularized low-rank matrix estimation. *arXiv preprint arXiv:1410.8275*, 2014.

[69] J. Josse, S. Wager, and F. Husson. Confidence areas for fixed-effects pca. *arXiv preprint arXiv:1407.7614*, 2014.

[70] M. Journée, F. Bach, P. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[71] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[72] R. Keshavan. *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University, 2012.

[73] R. Keshavan and A. Montanari. Regularization for matrix completion. In *2010 IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1503–1507. IEEE, 2010.

[74] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, pages 952–960, 2009.

[75] R. Keshavan and S. Oh. A gradient descent algorithm on the Grassman manifold for matrix completion. *arXiv preprint arXiv:0910.5260*, 2009.

[76] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[77] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[78] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[79] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[80] J. Kim and H. Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *Eighth IEEE International Conference on Data Mining*, pages 353–362. IEEE, 2008.

[81] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[82] R. Koenker. *Quantile regression*. Cambridge University Press, 2005.

[83] R. Koenker and J. G. Bassett. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[84] E. Lawler and D. Wood. Branch-and-bound methods: A survey. *Operations Research*, 14(4):699–719, 1966.

[85] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[86] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

[87] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2006.

[88] J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. Tropp. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, pages 1297–1305, 2010.

[89] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[90] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 1932.

[91] C. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[92] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[93] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[94] L. Mackey. Deflation methods for sparse PCA. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, 2009.

[95] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

[96] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach. Supervised dictionary learning. In *Advances in Neural Information Processing Systems*, pages 1033–1040, 2009.

[97] I. Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications*. Communications and Control Engineering. Springer, 2012.

[98] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[99] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[100] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[101] T. Minka. Automatic choice of dimensionality for pca. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, pages 598–604. MIT Press, 2001.

[102] K. Mohan and M. Fazel. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the 2010 American Control Conference (ACC)*, pages 2953–2959. IEEE, 2010.

[103] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Provable non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[104] F. Niu, B. Recht, C. Ré, and S. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2011.

[105] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[106] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[107] S. Osnaga. *Low Rank Representations of Matrices using Nuclear Norm Heuristics*. PhD thesis, Colorado State University, 2014.

[108] A. Owen and P. Perry. Bi-cross-validation of the svd and the non-negative matrix factorization. *The Annals of Applied Statistics*, pages 564–594, 2009.

[109] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[110] H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, 36(2, Part 2):3336–3341, 2009.

[111] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[112] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12, 2014.

[113] P. Perry. Cross-validation for unsupervised learning. *arXiv preprint arXiv:0909.3052*, 2009.

[114] J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems*, pages 547–553, 1999.

[115] K. Preacher and R. MacCallum. Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 2(1):13–43, 2003.

[116] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, pages 759–766. ACM, 2007.

[117] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, August 2010.

[118] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[119] B. Recht, C. Ré, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[120] J. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 713–719. ACM, 2005.

[121] W. Revelle and K. Anderson. Personality, motivation and cognitive performance: Final report to the army research institute on contract MDA 903-93-K-0008. Technical report, 1998.

[122] P. Richtárik, M. Takáč, and S. Ahipaşaoğlu. Alternating maximization: Unifying framework for 8 sparse PCA formulations and efficient parallel codes. *arXiv preprint arXiv:1212.4137*, 2012.

[123] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

[124] A. Schein, L. Saul, and L. Ungar. A generalized linear model for principal component analysis of binary data. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume 38, page 46, 2003.

[125] S. Schelter, V. Satuluri, and R. Zadeh. Factorbird — a parameter server approach to distributed matrix factorization. *NIPS 2014 Workshop on Distributed Machine Learning and Matrix Computations*, 2014.

[126] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[127] S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-scale convex minimization with a low-rank constraint. *arXiv preprint arXiv:1106.1622*, 2011.

[128] H. Shen and J. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

[129] A. Singh and G. Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases*, pages 358–373. Springer, 2008.

[130] R. Smith. Nuclear norm minimization methods for frequency domain subspace identification. In *Proceedings of the 2010 American Control Conference (ACC)*, pages 2689–2694. IEEE, 2012.

[131] M. Soltanolkotabi and E. Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[132] M. Soltanolkotabi, E. Elhamifar, and E. Candes. Robust subspace clustering. *arXiv preprint arXiv:1301.2603*, 2013.

[133] N. Srebro. *Learning with Matrix Factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

[134] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *ICML*, volume 3, pages 720–727, 2003.

[135] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, volume 17, pages 1329–1336, 2004.

[136] V. Srikumar and C. Manning. Learning distributed representations for structured output prediction. In *Advances in Neural Information Processing Systems*, pages 3266–3274, 2014.

[137] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[138] H. Steck. Hinge rank loss and the area under the ROC curve. In J. N. Kok, J. Koronacki, R. L. Mantaras, S. Matwin, D. Mladenič, and A. Skowron, editors, *Machine Learning: ECML 2007*, volume 4701 of *Lecture Notes in Computer Science*, pages 347–358. Springer Berlin Heidelberg, 2007.

[139] D. L. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[140] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via nonconvex factorization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 270–289. IEEE, 2015.

[141] Y. Takane, F. Young, and J. De Leeuw. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, 42(1):7–67, 1977.

[142] M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[143] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[144] J. Tropp. *Topics in Sparse Approximation*. PhD thesis, The University of Texas at Austin, 2004.

[145] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[146] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

[147] M. Tweedie. An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, pages 579–604, 1984.

[148] N. Usunier, D. Buffoni, and P. Gallinari. Ranking with ordered weighted pairwise classification. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 1057–1064. ACM, 2009.

[149] S. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

[150] R. Vidal. A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2010.

[151] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

[152] V. Vu, J. Cho, J. Lei, and K. Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2670–2678. Curran Associates, Inc., 2013.

[153] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35, 2010.

[154] J. Weston, H. Yee, and R. J. Weiss. Learning to rank recommendations with the k-order statistic loss. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 245–248, New York, NY, USA, 2013. ACM.

[155] D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[156] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. In *Advances in Neural Information Processing Systems*, volume 3, 2009.

[157] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

[158] F. Young, J. De Leeuw, and Y. Takane. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4):505–529, 1976.

[159] H. Yun, H.-F. Yu, C.-J. Hsieh, S. V. N. Vishwanathan, and I. Dhillon. NOMAD: Non-locking, stOchastic Multi-machine algorithm for Asynchronous and Decentralized matrix completion. *arXiv preprint arXiv:1312.0193*, 2013.

[160] M. Zaharia, M. Chowdhury, M. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on hot topics in cloud computing*, page 10, 2010.

[161] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

[162] W. Zwick and W. Velicer. Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432, 1986.