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

Full text available at: http://dx.doi.org/10.1561/2200000055
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.
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

1 Introduction 2
   1.1 Previous work 4
   1.2 Organization 8

2 PCA and quadratically regularized PCA 9
   2.1 PCA 9
   2.2 Quadratically regularized PCA 11
   2.3 Solution methods 11
   2.4 Missing data and matrix completion 15
   2.5 Interpretations and applications 17
   2.6 Offsets and scaling 20

3 Generalized regularization 21
   3.1 Solution methods 22
   3.2 Examples 23
   3.3 Offsets and scaling 29

4 Generalized loss functions 30
   4.1 Solution methods 30
   4.2 Examples 31
   4.3 Offsets and scaling 35
5 Loss functions for abstract data types 37
  5.1 Solution methods ........................................... 38
  5.2 Examples ...................................................... 38
  5.3 Missing data and data imputation ......................... 41
  5.4 Interpretations and applications ......................... 42
  5.5 Offsets and scaling ........................................ 45
  5.6 Numerical examples ........................................ 45

6 Multi-dimensional loss functions 54
  6.1 Examples ...................................................... 55
  6.2 Offsets and scaling ........................................ 59
  6.3 Numerical examples ........................................ 59

7 Fitting low rank models 61
  7.1 Alternating minimization ................................... 62
  7.2 Early stopping ............................................... 62
  7.3 Quadratic objectives ....................................... 67
  7.4 Convergence ................................................ 68
  7.5 Initialization ............................................... 70
  7.6 Global optimality .......................................... 73

8 Choosing low rank models 78
  8.1 Regularization paths ....................................... 78
  8.2 Choosing model parameters ................................ 80
  8.3 On-line optimization ...................................... 85

9 Implementations 87
  9.1 Python implementation .................................... 88
  9.2 Julia implementation ...................................... 90
  9.3 Spark implementation ..................................... 94

Acknowledgments 98

Appendices 99

A Examples, loss functions, and regularizers 100
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.
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
1.1. Previous work

function and regularizer, one may recover algorithms including PCA, weighted PCA, $k$-means, $k$-medians, $\ell_1$ SVD, probabilistic latent semantic indexing (pLSI), nonnegative matrix factorization with $\ell_2$ 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.
Introduction

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 [132, 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, 11, 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].
1.1. Previous work

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 non-locking 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


References


References


References


References


References


