Seminar in Data Science

Lecture 5: Generalized Low-Rank Models

Laurent El Ghaoui

Seminar in Data Science and Information Technology, Summer 2020
TBSI – UC Berkeley

7/22/2020

Outline

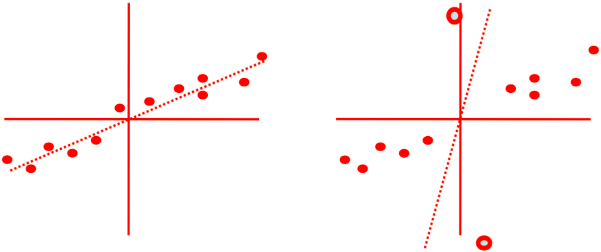
Outline

Motivation

PCA has many restrictions, including

- it works only on fully known matrix (no missing entries);
- it cannot handle different data types, such as Boolean, categorical, non-negative, probability vectors, etc;
- it is very sensitive to outliers, an artefact due to the squared Frobenius (variance) being the underlying metric used.

Gross errors of even one/few points can completely throw off PCA



Reason: Classical PCA minimizes ℓ_2 error, which is susceptible to gross outliers

In this lecture

- Describe a generalization of the low-rank idea, to more general data sets, loss functions, and penalties.
- Examine how the approach can handle missing data, and other extensions.

Outline

Low-rank models and alternate minimization

For $X \in \mathbf{R}^{n \times m}$, ordinary rank-$k$ model solves

$$\min_{L,R} \|X - LR^T\|_F \;:\; L \in \mathbf{R}^{n \times k}, \quad R \in \mathbf{R}^{m \times k},$$

by minimization over L, R alternatively. This is PCA, if we work with a column-centered data matrix.

Note that $(LR^T)_{ij} = l_i^T r_j$, where

$$L = \begin{pmatrix} l_1^T \\ \vdots \\ l_n^T \end{pmatrix}, \quad R = \begin{pmatrix} r_1^T \\ \vdots \\ r_m^T \end{pmatrix},$$

Thus we can write the above problem as

$$\min_{L,R} \sum_{i,j} \mathcal{L}(X_{ij}, l_i^T r_j) \;:\; l_i \in \mathbf{R}^k, \quad i = 1, \ldots, n, \quad r_j \in \mathbf{R}^k, \quad j = 1, \ldots, m,$$

with $\mathcal{L}(a, b) = (a - b)^2$.

Generalization

Generalized low-rank model [2] solves

$$\min_{L,R} \sum_{i,j} \mathcal{L}(X_{ij}, l_i^T r_j) + \sum_i p_i(l_i) + \sum_j q_j(r_j),$$

where $\mathcal{L}$ is convex, and functions p_i, q_j are convex penalties.

- The problem is not convex—but it is with respect to X, R (resp. X, L) when L (resp. R) is fixed.
- We can solve the problem by alternative minimization over L, R.
- In most cases, there is no guarantee of convergence to a global minimum.
- Playing with different losses and penalties we can model a lot of useful situations.

Convex model

An alternative to the alternating minimization method is based on the following idea:

> *In order to minimize the rank of a matrix, we may try to minimize the sum of the singular values.*

This leads to a convex model of the form

$$\min_Z \sum_{i,j} \mathcal{L}(X_{ij}, Z_{ij}) + \lambda \|Z\|_*,$$

where $\|Z\|_*$ is the nuclear norm (sum of the singular values.

Although convex, the problem is challenging due to its size; in practice, alternative minimization is a very good heuristic (when squared regularization is included). For many finance application, the data size is not too big, and the convex model may be a reliable alternative.

Regularized PCA

In regularized PCA we solve the problem

$$\min_{L,R} \|X - LR^T\|_F^2 + \gamma \left(\|L\|_F^2 + \|R\|_F^2\right) \;:\; L \in \mathbf{R}^{n\times k}, \quad R \in \mathbf{R}^{m\times k},$$

with $\gamma > 0$ a regularization parameter.

Closed-form solution: Given the SVD of $X = U\Sigma V^T$, we set

$$\tilde{\Sigma}_{ii} = \max(0, \Sigma_{ii} - \gamma), \quad i = 1, \ldots, k$$

and $L = U_k \tilde{\Sigma}^{1/2}$, $R = V_k \tilde{\Sigma}^{1/2}$, with U_k, V_k the first k columns in U, V.

Interpretation: we truncate and threshold the singular values.

Robust PCA

In robust PCA we seek to decompose the input data matrix X into a sum of a sparse and a low-rank component:

$$X = LR^T + S, \quad L \in \mathbf{R}^{n \times k}, \quad R \in \mathbf{R}^{m \times k}, \quad S \text{ sparse.}$$

We can model this with

$$\mathcal{L}(a, b) = |a - b|,$$

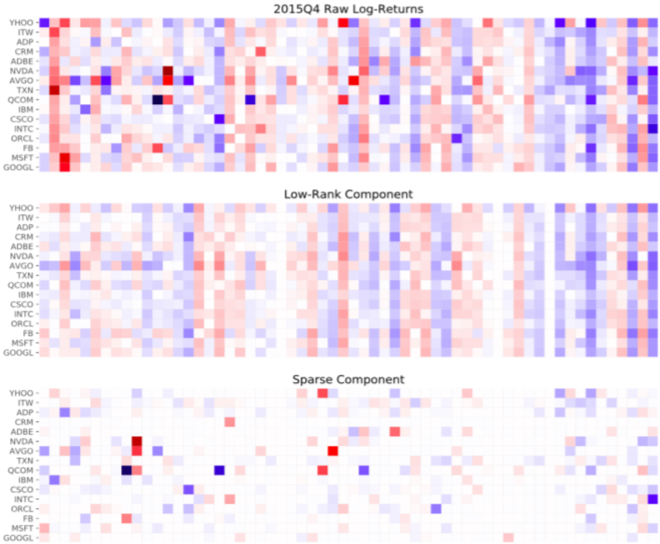
leading to

$$\min_{L,R} \sum_{i,j} |X_{ij} - l_i^T r_j| = \|X - LR^T\|_1,$$

where $\|Z\|_1$ is the sum of the absolute values of the entries of matrix Z.

The l_1-norm is chosen as a heuristic to make the matrix in the norm sparse.

Example



Data: 2015 Q4 raw log-returns for a number of tech companies. For an example in video analytics, see https://www.youtube.com/watch?v=BTrbow8u4Cw

Sparse PCA

In sparse PCA we seek to approximate a matrix by a low-rank one, each factor being sparse:

$$X = LR^T, \text{ with } L, R \text{ sparse.}$$

We can model this with

$$\mathcal{L}(a, b) = |a - b|,$$

leading to

$$\min_{L,R} \sum_{i,j} (X_{ij} - l_i^T r_j)^2 + \|L\|_1 + \|R\|_1.$$

Again the l_1-norm is chosen as a heuristic to make the matrix in the norm sparse.

Non-negative matrix factorization

Non-negative matrix factorization (NNMF) is a variant on PCA where the factors are required to be non-negative:

$$X = LR^T, \text{ with } L \geq 0, \quad R \geq 0,$$

with inequalities understood component-wise. This problem arises when the data matrix is itself non-negative.

We can model this with

$$\min_{L,R} \sum_{i,j} (X_{ij} - l_i^T r_j)^2 \ : \ L \geq 0, \quad R \geq 0,$$

corresponding to penalties p_i, q_j all chosen to be equal to

$$p(z) = \begin{cases} 0 & \text{if } z \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Boolean data

Sometimes entries in the data are Boolean, that is, $X_{ij} \in \{-1, 1\}$. We can model these entries with

$$\mathcal{L}(a, b) = \max(0, 1 - ab) = (1 - ab)_+.$$

Motivation: If $a \in \{-1, 1\}$, $\mathcal{L}(a, b) = 0$ implies b has the same sign as a.

For example, if $X \in \{-1, 1\}^{n \times m}$ is entirely Boolean, we solve

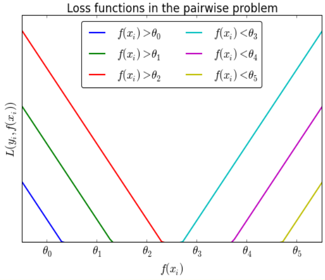
$$\min_{L,R} \sum_{i,j} (1 - X_{ij} l_i^T r_j)_+.$$

Categorical data

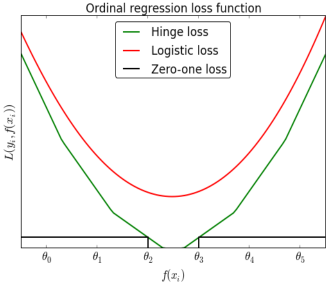
In ordinal PCA we wish to handle data that is categorical, for example stars in ratings, or

Strong Buy, Buy, Hold, Underperform or Sell

We encode all these in a set of thresholds θ_i, $i = 1, \ldots, K - 1$, with K the number of categories; say $\theta_i = i$, $i = 1, \ldots, K$. Each level corresponds to one part of the loss function; the overall loss is a sum of all of these.



Loss functions for each level.



Combining loss functions.

Categorical data: model

We can model categorical data with

$$\mathcal{L}(a,u) = \sum_{b=1}^{a-1}(1-u+b)_+ + \sum_{b=a+1}(1+u-b)_+.$$

Note: This approach assumes that every increment of error is equally bad: for example, that approximating"Strong Buy" by "Buy" is just as bad as approximating "Buy" by "Hold". There is a more flexible approach to this [2].

Outline

Matrix completion problem

Matrix completion is the problem of filling unknown entries of a partially known matrix.

The classical assumption is that the completion should be made so that the completed matrix has the lowest rank prossible.

PCA-based completion

Basic approach based on regularized PCA:

$$\min_{L,R,X\in\mathcal{X}} \|X - LR^T\|_F^2 + \gamma\left(\|L\|_F^2 + \|R\|_F^2\right) \;:\; L \in \mathbf{R}^{n\times k}, \quad R \in \mathbf{R}^{m\times k},$$

with X a variable, and $\mathcal{X}$ the set of $n \times m$ matrices that have the required given entries.

- Alternating minimization works the same! Just add missing entries in X as variables.
- Some theoretical results show that if the locations of missing entries are randomly distributed, convergence to the global minimum is guaranteed [1].
- In practice, for this to work, missing entries should not follow a clear pattern (*e.g.*, they should not all be located at the bottom in a time-series matrix).

Categorical data

With categorical data the filled entries should belong tto the category. To do this, we use

$$\hat{X}_{ij} = \arg\min_a \; L_{ij}(a, l_i^T r_j),$$

with L, R the final values delivered by the algorithm.

For example, with Boolean data, $X_{ij} \in \{-1, 1\}$, and we have

$$L_{ij}(a, b) = \max(1 - ab, 0),$$

so that

$$\hat{X}_{ij} = \arg\min_a \; \max(0, 1 - a l_i^T r_j).$$

Outline

Summary

- Generalized low-rank models offer a very flexible way to model data.
- It is always based on the key low-rank assumption, and generalizes standard PCA in many directions.
- In general, GLRMs are not convex, and convergence is not guaranteed.
- It is always a good idea to add a squared penalty to the loss function.

References

Rong Ge, Jason D Lee, and Tengyu Ma.
Matrix completion has no spurious local minimum.
In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al.
Generalized low rank models.
Foundations and Trends® in Machine Learning, 9(1):1–118, 2016.