Project Report
Low-rank Approximations for Large Incomplete Matrices

Anton Zhevnerchuk
anton.zhevnerchuk@skoltech.ru

https://github.com/zhevnerchuk/matrix-filling

Abstract

There are some challenging industrial problems in which only incomplete data is available and the goal is to "complete" the data. This problem can be stated as an optimization problem in at least two different ways, and a variety of methods can be used for solving the problem in each of the formulations. We implemented three competitive algorithms for matrix-filling problem and compared their performances to each other.

Contents

1 Background

The growth of the Internet influenced the development of recommender systems, data compression and reconstruction. Some of the problems appearing in that fields deal with data, which can be naturally represented as a matrix with some observed values and others values unknown. The goal in such problems is to guess the data at not observed entries. Usually it is assumed that our data has low-rank structure: this assumption follows naturally from the data structure or from the algorithms used for data compression.

2 Problem formulation

We have a matrix $X \in \mathbb{R}^{m \times n}$ with some unknown entries. Let us denote by Ω the set of positions (i, j) for which $X_{i,j}$ is known. We want to construct a low-rank approximation Z of X. In our project we considered two approaches to solving that problem:

- Fix rank of an approximation and minimize error at known entries;
- Fix acceptable error at known entries and minimize rank of an approximation.

For the following we need to introduce some notation. For a $m \times n$ matrix A $P_\Omega(A)$ is a projection of A on the set of known entries:

$$P_\Omega(A)_{i,j} = \begin{cases} A_{i,j}, & (i,j) \in \Omega \\ 0, & (i,j) \notin \Omega \end{cases}$$

When we fix acceptable error at observed entries $\delta > 0$ and minimize rank of an approximation, we have the following problem:

$$\text{minimize} \quad \text{rank}(Z) \qquad \text{subject to} \quad \sum_{(i,j)\in\Omega} (X_{i,j} - Z_{i,j})^2 \leq \delta. \tag{1}$$

Since this optimization problem 2 becomes too hard in general case, the following modification is considered:

$$\text{minimize} \quad \|Z\|_* \qquad \text{subject to} \quad \sum_{(i,j)\in\Omega} (X_{i,j} - Z_{i,j})^2 \leq \delta. \tag{2}$$

Here $\|\cdot\|_*$ is a nuclear norm, which is just a sum of singular values of a matrix. Such a relaxation is a classic approach for matrix-filling problem and is used in many papers (Fazel, 2002; Candes and Recht, 2008; Candes and Tao, 2009; Recht et al., 2007). It occurs that problem 3 is convex, which makes it much easier to solve. Nuclear norm is clearly connected with the rank, so in general a solution for 3 seems to be also a rather good solution for 2. It can be shown that for every $\delta > 0$ there is a $\lambda > 0$ such that 3 is equal to

$$\text{minimize} \quad \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{i,j} - Z_{i,j})^2 + \lambda \|Z\|_*. \tag{3}$$

So, our goal is to solve problem 4. The implemented algorithm (Soft-Input) corresponds to that formulation of a matrix completion problem.

3 Implemented approaches

3.1 Soft-Input

This algorithm is designed for solving matrix-completion problem in the formulation 4. If all matrix X is known, a solution for 4 can be found analytically. If $X = U\Lambda V^T$ is SVD for X, a solution is given by

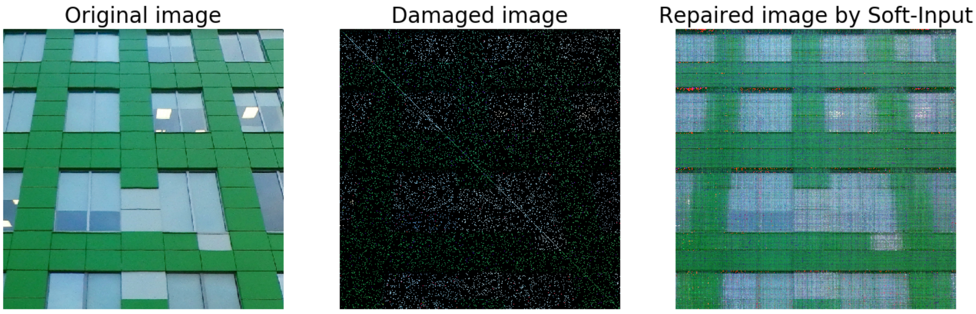
$$Z = U\Lambda_\lambda V^T, \text{ where } \Lambda_\lambda = \text{diag}\Big(\max\{\sigma_1 - \lambda, 0\}, \ldots, \max\{\sigma_{\min\{m,n\}} - \lambda, 0\}\Big).$$

Soft-Input is an iterative method for solving 4 in general setting with very simple main idea: at each iteration we replace unknown entries of X with the values of a current approximation Z_ℓ and solve problem 4 for a matrix with all entries known. In [MHT10] it is shown that this iterative algorithm converges to a solution for 4.

At each iteration we need to find SVD of $m \times n$ matrix $P_\Omega(X) - P_\Omega(Z_\ell) + Z_\ell$. Note that Z_ℓ is a matrix with low-rank, so MATVEC operations can be done in $O((m+n)k)$ flops, where k is rank of Z_ℓ. Matrix $P_\Omega(X) - P_\Omega(Z_\ell)$ is sparse. Therefore, if X has good sparse structure, MATVEC operations with matrix $P_\Omega(X) - P_\Omega(Z_\ell)$ are comparatively cheap as well, so truncated SVD of $P_\Omega(X) - P_\Omega(Z_\ell) + Z_\ell$ can be found much faster than in general case.

4 Demonstration

We download a photo of one of the buildings in Mendeleev Quarter. After that we set only about 5% of pixels to be known and then we try to repair the photo using some of implemented algorithms. We have achieved the following results:



We managed to find pretty completion using Soft-Input.

References

[MHT10] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani, *Spectral Regularization Algorithms for Learning Large Incomplete Matrices*, Journal of Machine Learning (2010), available at `https://web.stanford.edu/%7Ehastie/Papers/mazumder10a.pdf`.