Robust Principal Component Analysis & Collaborative Filtering

Who? Neil Menghani
Jeffrey Tumminia

When? March 24, 2021

Motivation

Let

$$M \in \mathbb{R}^{n \times m}$$

be a data matrix where n is the number of samples and m is the number of features.

We are interested in the scenario where

$$\left.\begin{array}{lc} i & m \text{ is large} \\ ii & n \approx m \\ iii & \text{noise} \\ iv & \text{outliers/corruption} \\ v & \text{missing entries} \end{array}\right\} \rightarrow \textbf{Collaborative Filtering (CF)}$$

In CF a fraction of entries in M are present, whose locations are $(i, j) \in \Omega$

Motivation

PCA (as matrix separation):

$$\begin{aligned} \min_{L} \quad & ||M - L|| \\ \text{s.t.} \quad & rank(L) \leq k \end{aligned} \tag{1}$$

where $||A|| = max(\sigma_i(A)\ \forall i\,)$

$\hat{L}$ would be the data in the k principal directions with the highest singular values (or variance).

Well-posed when $M = L_0 + N_0$

- L_0 is a low-rank matrix
- N_0 is a matrix of noise

Methodology

Assume

$$M = L_0 + S_0$$

S_0 is sparse corruption/outliers

Robust PCA:

$$\begin{aligned} \min_{L,S} \quad & rank(L) + ||S||_0 \\ \text{s.t.} \quad & M = L + S \end{aligned} \tag{2}$$

where $||A||_0 = \#(i \,|A_i \neq 0)$

Methodology

Through Principal Component Pursuit

$$\begin{aligned} \min_{L,S} \quad & ||L||_* + \lambda ||S||_1 \\ \text{s.t.} \quad & M = L + S \end{aligned} \tag{3}$$

where

- $||A||_* = \sum_i \sigma_i(A)$
- $||A||_1 = \sum_{ij} |A_{ij}|$

$\hat{L} = L_0, \hat{S} = S_0$ when $\lambda = \frac{1}{\sqrt{n}}$ and

- L_0 should not be sparse
- S_0 should not be low-rank

Matrix Completion:

If we assume no outliers in $M \to S_0 = \mathbf{0}$ then we can rewrite RPCA as;

$$\begin{aligned} \min_{L} \quad & ||L||_* \\ \text{s.t.} \quad & M_{ij} = L_{ij}, \ \forall \, (i,j) \in \Omega \end{aligned} \tag{4}$$

Represents the exact matrix completion problem under similar coherence constraints on M as RPCA.

Experiment

- Steam Video Games dataset: take a subset and introduce corruption (random outliers)
- Incorporate varying levels of corruption to the dataset.
- Implement Robust PCA and traditional Matrix Completion techniques on the datasets.

Matrix Completion Experiment

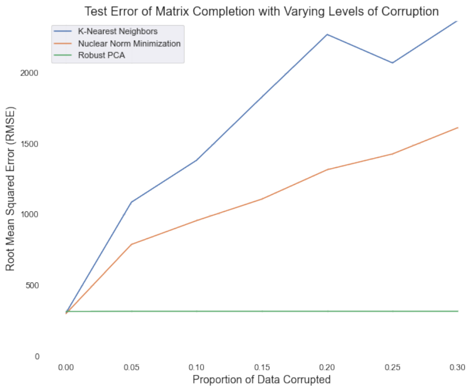
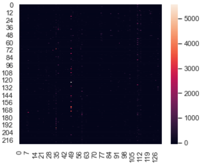
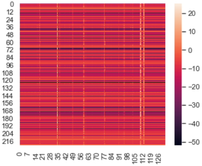


Figure: Plot of Root Mean Squared Error (RMSE) on test set for 3 matrix completion techniques with varying levels of corruption. K-Nearest Neighbors and Nuclear Norm Minimization slightly outperform Robust PCA on dataset with no corruption. However, performance of the former 2 techniques significantly deteriorates as corruption is added, while RPCA performs consistently in the presence of outliers.

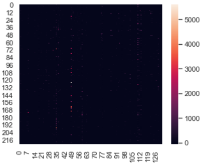
Heatmap Visualization
RPCA No Corruption



(a) Centered Data Matrix M



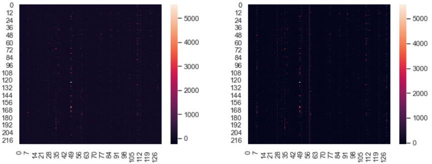
(b) Low-Rank Component L (CF Solution)



(c) Sparse Component S (extracted outliers)

Figure: Heatmaps of Robust PCA output with no corruption. Games displayed on x-axis and users on y-axis. Low-rank and sparse components extracted. Recall that RPCA minimizes objective function subject to $M = L + S$.

Heatmap Visualization
NNM+kNN No Corruption

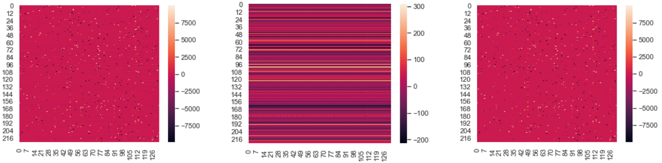


(a) Low-Rank Output L
Nuclear Norm
Minimization

(b) Matrix Completion
Solution
K-Nearest Neighbors

Figure: Heatmaps of traditional matrix completion techniques with no corruption. Games displayed on x-axis and users on y-axis. For certain games, entries in original matrix filled in with values inferred by the 2 techniques.

Heatmap Visualization
RPCA Corruption



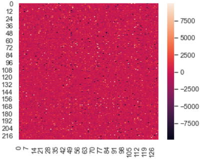
(a) Centered Matrix M with Corruption

(b) Low-Rank Component L (CF Solution)

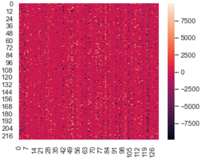
(c) Sparse Component S (extracted outliers)

Figure: Heatmaps of Robust PCA output with 10% corruption. Games displayed on x-axis and users on y-axis. A coherent low-rank component is still extracted, and outliers added by corruption are pulled into the sparse component. As in Figure 3, M=L+S.

Heatmap Visualization
NNM+kNN with Corruption



(a) Low-Rank Output L
Nuclear Norm
Minimization



(b) Matrix Completion
Solution
K-Nearest Neighbors

Figure: Heatmaps of traditional matrix completion techniques with 10% corruption. Games displayed on x-axis and users on y-axis. For certain games, entries in original matrix filled in with values inferred by the 2 techniques.

Closing Remarks

Points of Emphasis:

- Focus was **not** overall performance
- Other RPCA Applications (Video Surveillance)
- Matrix Completion Methods

Helpful Resources:

- Video + TextBook
- Prof. Fernandez-Granda's notes