CSE 8803RS: Recommendation Systems

Lecture 12: Matrix Completion from a Few Entries

Liangda Li

School of Computational Science & Engineering
College of Computing
Georgia Institute of Technology

Problem Definition for Matrix Completion

- Imagine that each of m customers watches and rates a subset of the n movies available through a movie rental service. This yields a dataset of customer-moive pairs $(i, j) \in E \subseteq [m] \times [n]$, and for each such pair, a rating $M_{ij} \in \mathcal{R}$.
- The object is to predict the rating for the missing item of given matrice.
- The paper is interested in very large data sets, thus focus on the limit $m, n \to \infty$ with $m/n = \alpha$ bounded away from 0 and ∞.

The General Question to Address

- Under which conditions do the known ratings provide sufficient information to infer unknown ones?
- Can this inference problem be solved efficiently?

Model Definition

- Assume the matrix of ratings has rank $r \ll m, n$. More precisely, we denote by M the matrix whose entry $(i, j) \in [m] \times [n]$ corresponds to the rating user i would assign to movie j.
- We assume that there exist matrices U, of dimensions $m \times r$, and V, of dimensions $n \times r$, and a diagonal matrix Σ, of dimensions $r \times r$ such that

$$M = U \Sigma V^T$$

- Let M^E be the $m \times n$ matrix that contains the revealed entries of M, and is filled with 0's in the other positions

$$M^E_{i,j} = \begin{cases} M_{i,j} & \text{if}(i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Assumptions

- The subset of observed entries E is uniformly random.
- The factors U, V are unstructured. This notion is formalized by the incoherence condition introduced by Candes and Recht [1].

Incoherence property

- **A1.** There exist a constant $\mu_0 > 0$ such that for all $i \in [m]$, $j \in [n]$, we have $\sum_{k=1}^{r} U_{i,k}^2 \leq \mu_0 r$, $\sum_{k=1}^{r} V_{i,k}^2 \leq \mu_0 r$.
- **A2.** There exist μ_1 such that $|\sum_{k=1}^{r} U_{i,k}\Sigma_k V_{j,k}| \leq \mu_1 r^{1/2}$.
- In particular, the incoherence condition is satisfied with high probability if U and V are uniformly random matrices with $U^T U = m\mathbf{1}$ and $V^T V = n\mathbf{1}$.

A Naive Algorithm

- **Projection**. Compute the singular value decomposition (SVD) of M^E (with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$)

$$M^E = \sum_{i=1}^{min(m,n)} \sigma_i x_i y_i^T$$

  and return the matrix $T_r(M^E) = (mn/|E|) \sum_{i=1}^{r} \sigma_i x_i y_i^T$ obtained by setting to 0 all but the r largest singular values.
- Notice that, apart from the rescaling factor $(mn/|E|)$, $T_r(M^E)$ is the orthogonal projection of M^E onto the set of rank-r matrices. The rescaling factor compensates the smaller average size of the entries of M^E with respect to M.

Weakness in Naive Algorithm

- If $|E| = \Theta(n)$, this algorithm performs very poorly.
- M^E contains columns and rows with $\Theta(logn/loglogn)$ non-zero (revealed) entries. The largest singular values of M^E are of order $\Theta(\sqrt{logn/loglogn})$. The corresponding singular vectors are highly concentrated on high-weight column or row indices.
- Such singular vectors are an artifact of the high-weight columns/rows and do not provide useful information about the hidden entries of M.

Trimming

set to zero all columns in M^E with degree larger that $2|E|/n$. Set to zero all rows with degree larger than $2|E|/m$ (degree of a column of a row is the number of its revealed entries).

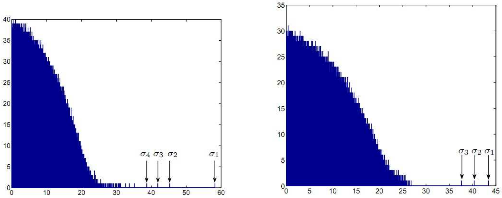

Figure 1: Histogram of the singular values of a partially revealed matrix M^E before trimming (left) and after trimming (right) for $10^4 \times 10^4$ random rank-3 matrix M with $\epsilon = 30$ and $\Sigma = \text{diag}(1, 1.1, 1.2)$. After trimming the underlying rank-3 structure becomes clear. Here the number of revealed entries per row follows a heavy tail distribution with $\mathbb{P}\{N = k\} = \text{const.}/k^3$.

The above figure shows that trimming makes the underlying rank-3 structure much more apparent. This effect becomes even more important when the number of revealed entries per row/column follows a heavy tail distribution, as for real data.

Proposed Algorithm

Algorithm 1 Spectral Matrix Completion

Trim M^E, and let $\tilde{M}^E$ be the output;
Project $\tilde{M}^E$ to $T_r(\tilde{M}^E)$;
Clean residual errors by minimizing the descrepancy $F(X, Y)$.

Cleaning

- The last step of aboving algorithm allows to reduce small discrepancies between $T_r(\tilde{M}^E)$ and M.
- **Cleaning**. Given $X \in \mathcal{R}^{m \times r}$, $Y \in \mathcal{R}^{n \times r}$ with $X^T X = m\mathbf{1}$ and $Y^T Y = n\mathbf{1}$, we define

$$F(X, Y) = min_{S \in \mathcal{R}^{r \times r}} \mathcal{F}(X, Y, S)$$

$$F(X, Y, S) = \frac{1}{2} \sum_{(i,j) \in E} (M_{ij} - (XSY^T)_{ij})^2$$

- The cleaning step consists in writing $T_r(\tilde{M}^E) = X_0 S_0 Y_0^T$ and minimizing $F(X, Y)$ locally with initial condition $X = X_0$, $Y = Y_0$.
- In geometrix terms, F is a function defined over the cartesian product of two Grassmann manifolds. Newton and conjugate gradient method [2] is applied to solve the problem.

Evaluation

Define the relative root mean square error as

$$RMSE = [\frac{1}{mnr} \|M - T_r(\tilde{M}^E)\|_F^2]^{1/2}$$

where we denote by $\|A\|_F$ the Frobenius norm of matrix A. Notice that the factor $(1/mn)$ corresponds to the usual normalization by the number of entries. The factor $(1/r)$ is instead necessary because the typical size of the entries of M is $\sqrt{r}$.

Main Result

- Assume M to be a rank $r \leq n^{1/2}$ matrix that satisfies the incoherence condition A2. Then with high probablity

$$\frac{1}{mnr}\|M - T_r(\tilde{M}^E)\|_F^2 \leq C(\alpha)\frac{nr}{|E|}$$

- Assume M to be a rank $r \leq n^{1/2}$ matrix that satisfies the incoherence conditions A1 and A2. Further, assume $\Sigma_{min} \leq \Sigma_1, \ldots, \Sigma_r \leq \Sigma_{max}$ with Σ_{min}, Σ_{max} bounded away from 0 and ∞. Then there exists $C^{'}(\alpha)$ such that, if

$$|E| \leq C^{'}(\alpha) nr \max\{logn, r\}$$

then the cleaning procedure in `Spectral Matrix Completion` converges, with high probability, to the matrix M.

Proof of Theorem 1

- There exists $C(\alpha) < \infty$ such that, with high probability

$$\frac{1}{\sqrt{mn}}\|M - T_r(\tilde{M}^E)\|_2 \leq C(\frac{r}{\epsilon})^{1/2}$$

- There exists $C(\alpha) > 0$ such that, with high probability

$$|x^T(\frac{\epsilon}{\sqrt{mn}}M - \tilde{M}^E)y| \leq C\sqrt{r\epsilon}$$

  for any $x \in \mathcal{R}^m$ and $y \in \mathcal{R}^n$ such that $\|x\| = \|y\| = 1$.
- There exists $C(\alpha) > 0$ such that, with high probability

$$|\frac{\sigma_q}{\epsilon} - \Sigma_q| \leq C(\frac{r}{\epsilon})^{1/2}$$

  where it is understood that $\Sigma_q = 0$ for $q > r$.

Proof of Theorem 2

- Assume $\epsilon > A\max\{r\log n, r^2\}$ with A large enough. Then there exists constants $C_1, C_2, \delta > 0$ such that, wiht high probability

$$C_1 n\epsilon(d(\mathbf{x},\mathbf{u})^2 + \|S-\Sigma\|_F^2) \leq F(\mathbf{x}) \leq C_2 n\epsilon d(\mathbf{x},\mathbf{u})^2$$

  for all $\mathbf{x} \in M(m,n) \cap \mathcal{K}(3\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \leq \delta$.
- Assume $\epsilon > A\max\{r\log n, r^2\}$ with A large enough. Then there exists constants $C, \delta > 0$ such that, wiht high probability

$$\|\mathrm{grad}\tilde{F}(\mathbf{x})\|^2 \geq C n\epsilon d(\mathbf{x},\mathbf{u})^2$$

  for all $\mathbf{x} \in M(m,n) \cap \mathcal{K}(3\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \leq \delta$.

Future Directions

- Optimal RMSE with $O(n)$ entries.
  — RMSE can decays much faster with the number of observations per degree of freedom ($|E|/nr$).
- Threshold for exact completion.
  — Prove exact reconstruction for $|E| \leq C'(\alpha) nr \log n$ for all values of r.
- More general models.
  — address the problem when observed entries E is far from uniformly random or the matrices are non-incoherent.

Reference

1 E. J. Candes and B. Recht, Exact matrix completion via convex optimization, arxiv:0805.4471, 2008.

2 A. Edelman, T. A. Arias, and S. T. Smith, The geometry of algorithms with orthogonality constraints, SIAM J. Matr. Anal. Appl. 20 (1999), 303¨C353.