Scaled Matrix Completion

Yichen Zhang

Department of Statistics

University of British Columbia

yichen.zhang@stat.ubc.ca

November 3, 2018

Overview

1. Matrix Completion
   - Motivating Example: Netflix Problem
   - Convex Relaxation
   - MC with Noise
2. Scaled Matrix Completion
   - Motivation of Scaled MC
   - Algorithms
3. Simulation Study
   - Comparison between MC and Scaled MC

Matrix Completion and Netflix Problem

Matrix completion: the recovery of an incomplete data matrix based on its observed entries.

Netflix Problem: Given a ratings matrix M in which each entry $M_{i,j}$ represents the rating of movie j by customer i if customer i has watched movie j and is otherwise missing. We would like to predict the remaining entries in order to make good recommendations to customers on what to watch next.

$$\begin{bmatrix} M_{1,1} & ? & M_{1,3} & ? & \cdots & M_{1,n} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ ? & \cdots & \cdots & ? & M_{m,n-1} & M_{m,n} \end{bmatrix}$$

Matrix Completion and Netflix Problem

Netflix Problem:

$$\begin{bmatrix} M_{1,1} & ? & M_{1,3} & ? & \cdots & M_{1,n} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ ? & \cdots & \cdots & ? & M_{m,n-1} & M_{m,n} \end{bmatrix}$$

If we define P_Ω as the projection operator onto the observed entries set Ω, our goal is to find X, such that

$$P_\Omega(X) = P_\Omega(M) \tag{1}$$

Without any restrictions on the number of degrees of freedom in the completed matrix this problem is underdetermined since the hidden entries could be assigned arbitrary values.

Low-rank Matrix Completion

Matrix completion often seeks to find the lowest rank matrix.

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & rank(X) \\ \text{subject to} \quad & P_\Omega(X) = P_\Omega(M). \end{aligned}$$

Why low rank? What does the rank stand for?

Why Low Rank?

In the Netflix problem, if we have m cutomers and n movies to rate. One might imagine that entires of M are determined by a few hidden cutomer features such as movie genre, time of release, etc. In particular, one might imagine a simple linear relationship between average rating, and those features:

$$M_{m\times n} = A_{m\times r} B_{r\times n}$$

$$\begin{array}{c} \\ customer_1 \\ \vdots \\ customer_m \end{array} \begin{array}{c} \begin{array}{cccc} movie_1 & movie_2 & \dots & movie_n \end{array} \\ \begin{pmatrix} M_{11} & M_{12} & \dots & M_{1n} \\ \vdots & \vdots & \dots & \vdots \\ M_{m1} & M_{m2} & \dots & M_{mn} \end{pmatrix} \end{array} =$$

$$\begin{array}{c} \\ customer_1 \\ \vdots \\ customer_m \end{array} \begin{array}{c} \begin{array}{ccc} feature_1 & \dots & feature_r \end{array} \\ \begin{pmatrix} a_{11} & \dots & a_{1r} \\ \vdots & \vdots & \vdots \\ a_{m1} & \dots & a_{mr} \end{pmatrix} \end{array} \begin{array}{c} \\ feature_1 \\ \vdots \\ feature_r \end{array} \begin{array}{c} \begin{array}{ccc} movie_1 & \dots & movie_n \end{array} \\ \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \vdots & \vdots \\ b_{r1} & \dots & b_{rn} \end{pmatrix} \end{array}$$

Convex Relaxation

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & rank(X) \\ \text{subject to} \quad & P_\Omega(X) = P_\Omega(M). \end{aligned} \tag{2}$$

However, the optimization (2) is not trivial to solve for two reasons:

1. Minimizing the rank is NP-hard.
2. The known algorithms which provide the exact solution for (2) require time doubly exponential in the dimension n of the matrix in both theory and practice.

It is standard practice to replace the rank function $rank(X)$ with a convex relaxation, the nuclear norm $||X||_*$.

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & ||X||_* \\ \text{subject to} \quad & P_\Omega(X) = P_\Omega(M). \end{aligned} \tag{3}$$

where $||X||_*$ is called the nuclear norm, which is the sum of all singular values of X.

MC with Noise

In real world application, one often observe M with at least by a small amount of noise. For example, in the Netflix problem, the ratings are uncertain.

If M is observed with some noise Z, $P_\Omega(M) = P_\Omega(X) + P_\Omega(Z)$, assuming that $||P_\Omega(Z)||_F \leq \delta$ for some δ

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & ||X||_* \\ \text{subject to} \quad & ||P_\Omega(M) - P_\Omega(X))||_F \leq \delta. \end{aligned} \tag{4}$$

$||A||_F$ is called the Frobenius norm, where $||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$

Alternatively, we can write (4) as an matrix regularization form with some tuning parameter λ

$$\min_X \frac{1}{2}||P_\Omega(M) - P_\Omega(X)||_F^2 + \lambda||X||_*$$

Motivation of Scaled MC

$$\min_X \frac{1}{2}||P_\Omega(M) - P_\Omega(X)||_F^2 + \lambda||X||_* \tag{5}$$

- How to find the "best" λ ? In practice, by cross-validation. The optimal value of λ has the smallest cross-validation error
- The optimal value of the tuning parameter λ depends on a number of a priori unknown quantities (like the noise Z) and therefore it can be difficult to calibrate in practice

We proposed a new matrix completion problem, which we call Scaled Matrix Completion

1. It gives the identical solution to the matrix completion problem (5)
2. The optimal value of the tuning parameter λ does not depend on the variance of the noise.

MC and Scaled MC

Let q be the number of observed entries, σ be the standard deviation of the noise Z_{ij}.

MC

$$\left(\hat{X}_a\right) := \arg\min_X \frac{1}{2}||P_\Omega(M) - P_\Omega(X)||_F^2 + \lambda||X||_*$$

Scaled MC

$$\left(\hat{X}_b, \hat{\sigma}_b^2\right) := \arg\min_{X,\sigma^2>0} \left\{ \frac{||P_\Omega(M) - P_\Omega(X)||_F^2}{q\sigma} + \sigma + \lambda||X||_* \right\}$$

1. $\hat{X}_a = \hat{X}_b$
2. For Scaled MC, the optimal value of the tuning parameter λ does not depend on σ.

Algorithms

MC

$$\left(\hat{X}_a\right) := \arg\min_X \frac{1}{2}||P_\Omega(M) - P_\Omega(X)||_F^2 + \lambda||X||_*$$

Proximal gradient descent (or generalized gradient descent): taking the generalized gradient at each step

Scaled MC

$$\left(\hat{X}_b, \hat{\sigma}_b^2\right) := \arg\min_{X,\sigma^2>0} \left\{ \frac{||P_\Omega(Y) - P_\Omega(X)||_F^2}{q\sigma} + \sigma + \lambda||X||_* \right\}$$

Coordinate descent: Searching the $\hat{X}_b$ and $\hat{\sigma}_b$ back and forth.

1. Fix $\hat{\sigma}_{b(n)}$, use proximal gradient descent to get the $\hat{X}_{b(n)}$
2. Fix $\hat{X}_{b(n)}$, solve the polynomial to get the $\hat{\sigma}_{b(n+1)}$

Simulation Results (MC and Scaled MC)

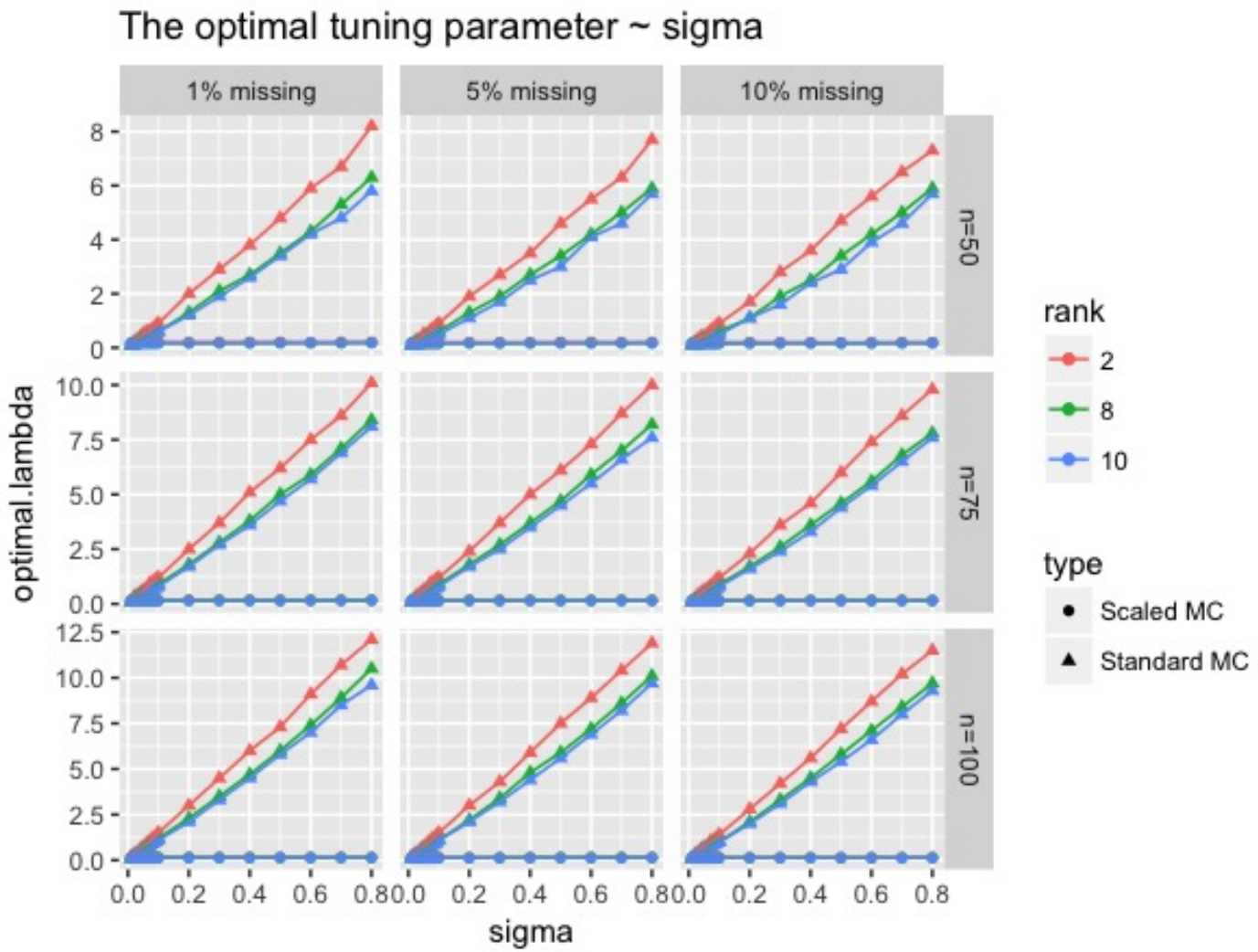

Figure: Optimal tuning parameter with sigma for scaled MC and standard MC

Simulation Results (Scaled MC)

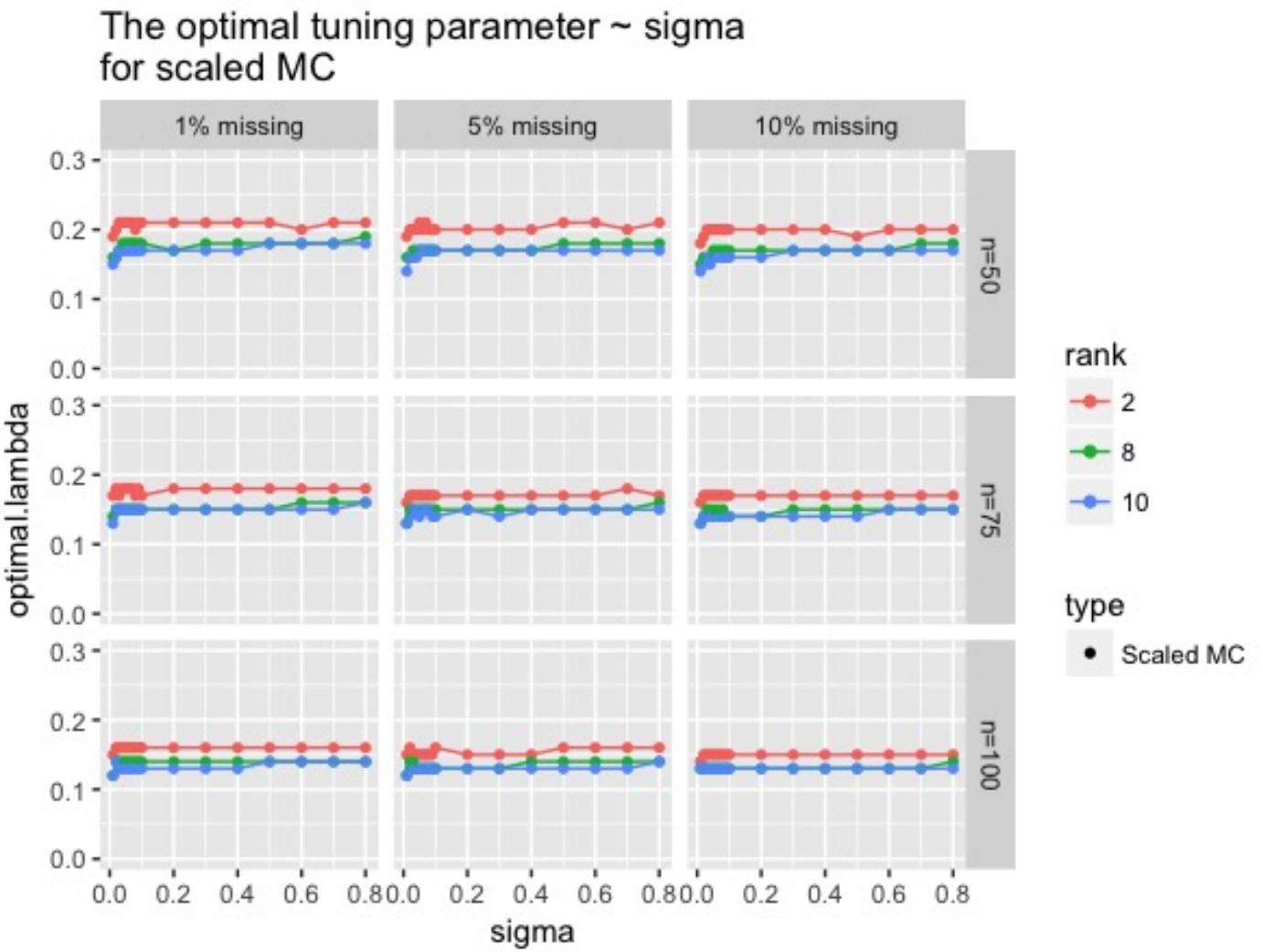

Figure: Optimal tuning parameter with sigma for scaled MC