Lecture: Matrix Completion

http://bicmr.pku.edu.cn/~wenzw/bigdata2020.html

Acknowledgement: this slides is based on Prof. Jure Leskovec and Prof. Emmanuel Candes's lecture notes

Recommendation systems

References:

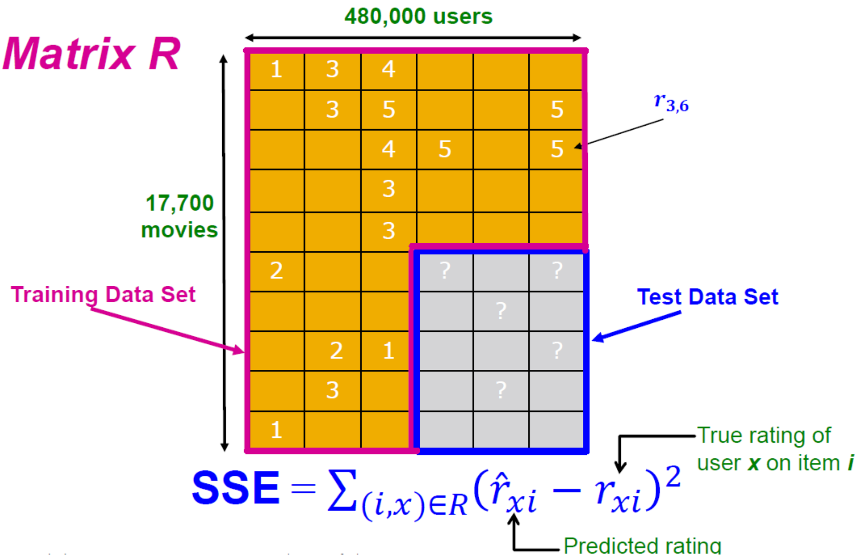
http://bicmr.pku.edu.cn/~wenzw/bigdata/07-recsys1.pdf

http://bicmr.pku.edu.cn/~wenzw/bigdata/08-recsys2.pdf

The Netflix Prize

- Training data
  - 100 million ratings, 480,000 users, 17,770 movies
  - 6 years of data: 2000-2005
- Test data
  - Last few ratings of each user (2.8 million)
  - Evaluation criterion: root mean squared error (RMSE): $\sqrt{\sum_{xi}(r_{xi} - r^*_{xi})^2}$: r_{xi} and r^*_{xi} are the predicted and true rating of x on i
  - Netflix Cinematch RMSE: 0.9514
- Competition
  - 2700+ teams
  - $1 million prize for 10% improvement on Cinematch

Netflix: evaluation



$$\text{SSE} = \sum_{(i,x)\in R} (\hat{r}_{xi} - r_{xi})^2$$

Collaborative Filtering: weighted sum model

$$\hat{r}_{xi} = b_{xi} + \sum_{j \in N(i;x)} w_{ij}(r_{xj} - b_{xj})$$

- baseline estimate for r_{xi}: $b_{xi} = \mu + b_x + b_i$
  μ: overall mean rating
  b_x: rating deviation of user x = (avg. rating of user x) - μ
  b_i: (avg. rating of movie i) - μ
- We sum over all movies j that are similar to i and were rated by x
- w_{ij} is the interpolation weight (some real number). We allow: $\sum_{j \in N(i,x)} w_{ij} \neq 1$
- w_{ij} models interaction between pairs of movies (it does not depend on user x)
- $N(i;x)$: set of movies rated by user x that are similar to movie i

Finding weights w_{ij}?

Find w_{ij} such that they work well on known (user, item) ratings:

$$\min_{w_{ij}} \quad F(w) := \sum_x \left(\left[b_{xi} + \sum_{j \in N(i;x)} w_{ij}(r_{xj} - b_{xj}) \right] - r_{xi} \right)^2$$

- Unconstrained optimization: quadratic function

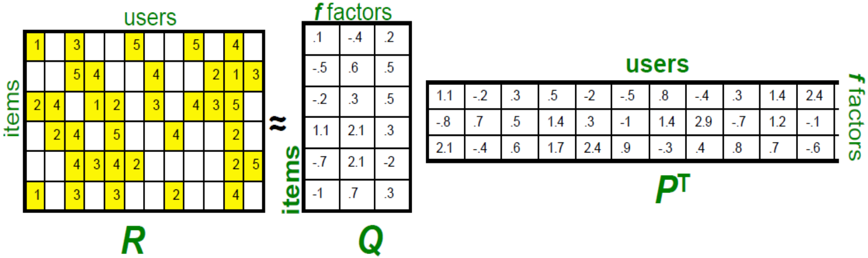
$$\nabla_{w_{ij}} F(w) = 2 \sum_x \left(\left[b_{xi} + \sum_{k \in N(i;x)} w_{ik}(r_{xk} - b_{xk}) \right] - r_{xi} \right) (r_{xj} - b_{xj}) = 0$$

  for $j \in \{N(i,x), \forall i, x\}$

- Equivalent to solving a system of linear equations?
- Steepest gradient descent method: $w^{k+1} = w^k - \tau \nabla F(w)$
- Conjugate gradient method

Latent factor models

- “SVD” on Netflix data: $R \approx Q \cdot P^T$



- For now let’s assume we can approximate the rating matrix R as a product of “thin” $Q \cdot P^T$
  R has missing entries but let’s ignore that for now! Basically, we will want the reconstruction error to be small on known ratings and we don’t care about the values on the missing ones

Ratings as products of factors

- How to estimate the missing rating of user x for item i?

$$\hat{r}_{xi} = q_i \cdot p_x^T = \sum_f q_{if} p_{xf},$$

where q_i is row i of Q and p_x is column x of P^T

items

.1	-.4	.2
-.5	**.6**	**.5**
-.2	.3	.5
1.1	2.1	.3
-.7	2.1	-2
-1	.7	.3

f factors

Q $\bullet$

users

f factors

1.1	-.2	.3	.5	**-2**	-.5	.8	-.4	.3	1.4	2.4	-.9
-.8	.7	.5	1.4	**.3**	-1	1.4	2.9	-.7	1.2	-.1	1.3
2.1	-.4	.6	1.7	**2.4**	.9	-.3	.4	.8	.7	-.6	.1

P^T

SVD - Properties

Theorem: SVD

If A is a real m-by-n matrix, then there exits

$$U = [u_1, \ldots, u_m] \in \mathbb{R}^{m \times m} \text{ and } V = [v_1, \ldots, v_n] \in \mathbb{R}^{n \times n}$$

such that $U^T U = I$, $V^T V = I$ and

$$U^T A V = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}, \quad p = \min(m, n),$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.

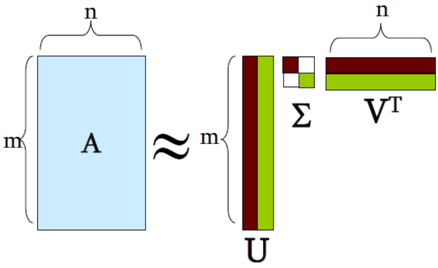
Eckart & Young, 1936

Let the SVD of $A \in \mathbb{R}^{m \times n}$ be given in Theorem: SVD. If $k < r = rank(A)$ and $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, then

$$\min_{rank(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

What is SVD?

- A: Input data matrix
- U: Left singular vecs
- V: Right singular vecs
- Σ: Singular values



- SVD gives minimum reconstruction error (SSE!)

$$\min_{U,V,\Sigma} \sum_{ij} (A_{ij} - [U\Sigma V^T]_{ij})^2$$

- In our case, "SVD" on Netflix data: $R \approx Q \cdot P^T$, i.e., $A = R, Q = U, P^T = V^T$
- But, we are not done yet! R has missing entries!

Latent factor models

- Minimize SSE on training data!
- Use specialized methods to find P, Q such that $\hat{r}_{xi} = q_i \cdot p_x^T$

$$\min_{P,Q} \sum_{(i,x)\in \text{training}} (r_{xi} - q_i \cdot p_x^T)^2$$

We don't require cols of P, Q to be orthogonal/unit length

- P, Q map users/movies to a latent space
- Add regularization:

$$\min_{P,Q} \sum_{(i,x)\in \text{training}} (r_{xi} - q_i \cdot p_x^T)^2 + \lambda \left[\sum_x \|p_x\|_2^2 + \sum_i \|q_i\|_2^2 \right]$$

λ is called regularization parameters

Gradient descent method

$$\min_{P,Q} \quad F(P,Q) := \sum_{(i,x)\in \text{training}} (r_{xi} - q_i \cdot p_x^T)^2 + \lambda \left[\sum_x \|p_x\|_2^2 + \sum_i \|q_i\|_2^2 \right]$$

Gradient decent:

- Initialize P and Q (using SVD, pretend missing ratings are 0)
- Do gradient descent:
  $P^{k+1} \leftarrow P^k - \tau \nabla_P F(P^k, Q^k)$,
  $Q^{k+1} \leftarrow Q^k - \tau \nabla_Q F(P^k, Q^k)$,
  where $(\nabla_Q F)_{if} = -2\sum_{x,i}(r_{xi} - q_i p_x^T) p_{xf} + 2\lambda q_{if}$. Here q_{if} is entry f of row q_i of matrix Q
- Computing gradients is slow when the dimension is huge

Stochastic gradient descent method

Observation: Letq_{if} be entry f of row q_i of matrix Q

$$
\begin{aligned}
(\nabla_Q F)_{if} &= \sum_{x,i} \left(-2(r_{xi} - q_i p_x^T) p_{xf} + 2\lambda q_{if}\right) = \sum_{x,i} \nabla_Q F(r_{xi}) \\
(\nabla_P F)_{xf} &= \sum_{x,i} \left(-2(r_{xi} - q_i p_x^T) q_{xf} + 2\lambda p_{if}\right) = \sum_{x,i} \nabla_P F(r_{xi})
\end{aligned}
$$

Stochastic gradient decent:

- Instead of evaluating gradient over all ratings, evaluate it for each individual rating and make a step
- $P \leftarrow P - \tau \nabla_P F(r_{xi})$
  $Q \leftarrow Q - \tau \nabla_Q F(r_{xi})$
- Need more steps but each step is computed much faster

Latent factor models with biases

predicted models:

$$\hat{r}_{xi} = \mu + b_x + b_i + q_i \cdot p_x^T$$

μ: overall mean rating, b_x: Bias for user x, b_i: Bias for movie i

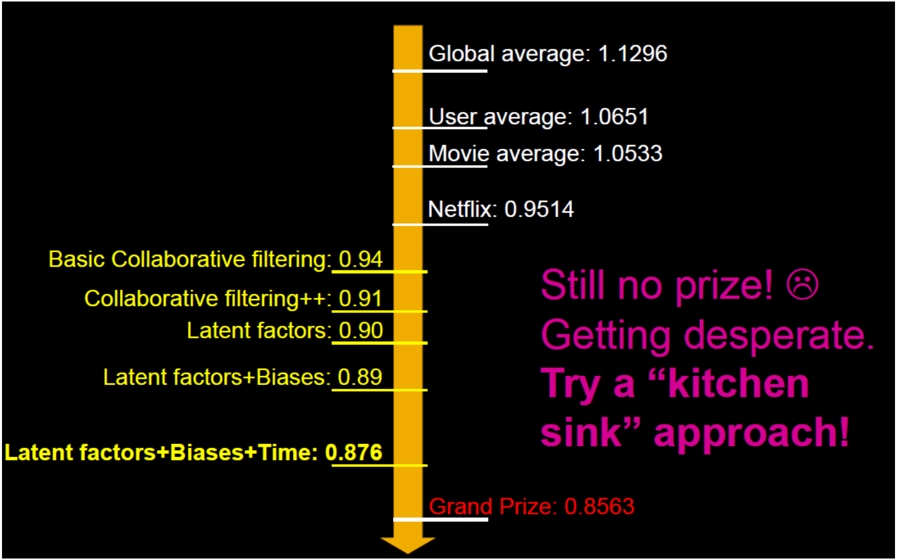
New model:

$$\min_{P,Q,b_x,b_i} \sum_{(i,x)\in \text{training}} (r_{xi} - (\mu + b_x + b_i + q_i \cdot p_x^T))^2 + \lambda \left[\sum_x \|p_x\|_2^2 + \sum_i \|q_i\|_2^2 + \|b_x\|_2^2 + \|b_i\|_2^2 \right]$$

- Both biases b_x, b_i as well as interactions q_i, p_x are treated as parameters (we estimate them)
- Add time dependence to biases:

$$\hat{r}_{xi} = \mu + b_x(t) + b_i(t) + q_i \cdot p_x^T$$

Netflix: performance



Netflix: performance

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update | Download

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	[illegible]	[illegible]	[illegible]
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22

General matrix completion

Matrix completion

- Matrix $M \in \mathbb{R}^{n_1 \times n_2}$
- Observe subset of entries
- Can we guess the missing entries?

$$\begin{bmatrix} \times & ? & ? & ? & \times & ? \\ ? & ? & \times & \times & ? & ? \\ \times & ? & ? & \times & ? & ? \\ ? & ? & \times & ? & ? & \times \\ \times & ? & ? & ? & ? & ? \\ ? & ? & \times & \times & ? & ? \end{bmatrix}$$

Which algorithm ?

Hope: only one low-rank matrix consistent with the sampled entries

Recovery by minimum complexity

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & X_{ij} = M_{ij}, \quad (i,j) \in \Omega \end{array}$$

Problem

- This is NP-hard
- Doubly exponential in n (?)

Nuclear-norm minimization

Singular value decomposition

$$X = \sum_{k=1}^{r} \sigma_k u_k v_k^*$$

- $\{\sigma_k\}$: singular values, $\{u_k\}, \{v_k\}$: singular vectors

Nuclear norm ($\sigma_i(X)$ is ith largest singular value of X)

$$\|X\|_* = \sum_{i=1}^{n} \sigma_i(X)$$

Heuristic

$$\begin{array}{rl} \text{minimize} & \|X\|_* \\ \text{subject to} & X_{ij} = M_{ij}, \quad (i,j) \in \Omega \end{array}$$

- Convex relaxation of the rank minimization program

Connections with compressed sensing

General setup

Rank minimization

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & \mathcal{A}(X) = b \end{array}$$

Convex relaxation

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \mathcal{A}(X) = b \end{array}$$

Suppose $X = \operatorname{diag}(x), x \in \mathbb{R}^n$

- $\operatorname{rank}(X) = \sum_i 1_{(x_i \neq 0)} = \|x\|_{\ell_0}$
- $\|X\|_* = \sum_i |x_i| = \|x\|_{\ell_1}$

Rank minimization

$$\begin{array}{ll} \text{minimize} & \|x\|_{\ell_0} \\ \text{subject to} & Ax = b \end{array}$$

Convex relaxation

$$\begin{array}{ll} \text{minimize} & \|x\|_{\ell_1} \\ \text{subject to} & Ax = b \end{array}$$

This is compressed sensing!

Correspondence

parsimony concept	cardinality	rank
Hilbert Space norm	Euclidean	Frobenius
sparsity inducing norm	ℓ_1	nuclear
dual norm	ℓ_∞	operator
norm additivity	disjoint support	orthogonal row and column spaces
convex optimization	linear programming	semidefinite programming

Table: From Recht Parrilo Fazel (08)

Semidefinite programming (SDP)

- Special class of convex optimization problems
- Relatively natural extension of linear programming (LP)
- “Efficient” numerical solvers (interior point methods)

LP: $x \in \mathbb{R}^n$

$$\begin{array}{ll} \text{minimize} & \langle c, x \rangle \\ \text{subject to} & Ax = b \\ & x \geq 0 \end{array}$$

SDP: $X \in \mathbb{R}^{n \times n}$

$$\begin{array}{ll} \text{minimize} & \langle C, X \rangle \\ \text{subject to} & \langle A_k, X \rangle = b_k \\ & X \succcurlyeq 0 \end{array}$$

Standard inner product: $\langle C, X \rangle = \text{trace}(C^* X)$

SOCP/SDP Duality

(P) $\min \quad c^\top x$
s.t. $\quad Ax = b, x_Q \succeq 0$

(D) $\max \quad b^\top y$
s.t. $\quad A^\top y + s = c, s_Q \succeq 0$

$$\textbf{(P)} \quad \begin{aligned} \min \quad & \langle C, X\rangle \\ \text{s.t.} \quad & \langle A_1, X\rangle = b_1 \\ & \dots \\ & \langle A_m, X\rangle = b_m \\ & X \succeq 0 \end{aligned}$$

$$\textbf{(D)} \quad \begin{aligned} \max \quad & b^\top y \\ \text{s.t.} \quad & \sum_i y_i A_i + S = C \\ & S \succeq 0 \end{aligned}$$

Strong duality

- If $p^* > -\infty$, (P) is **strictly** feasible, then (D) is feasible and $p^* = d^*$
- If $d^* < +\infty$, (D) is **strictly** feasible, then (P) is feasible and $p^* = d^*$
- If (P) and (D) has **strictly** feasible solutions, then both have optimal solutions.

Semidefinite program

$$(D)\quad \begin{array}{ll} \min & -b^\top y \\ \text{s.t.} & y_1 A_1 + \ldots + y_m A_m \preceq C \end{array}$$

- $A_i, C \in \mathcal{S}^k$, multiplier is matrix $X \in \mathcal{S}^k$
- Lagrangian $\mathcal{L}(y, X) = -b^\top y + \langle X, y_1 A_1 + \ldots + y_m A_m - C \rangle$
- dual function

$$g(X) = \inf_y \quad \mathcal{L}(y, X) = \begin{cases} -\langle C, X \rangle, & \langle A_i, X \rangle = b_i \\ -\infty & \text{otherwise} \end{cases}$$

The dual of (D) is

$$\begin{array}{ll} \min & \langle C, X \rangle \\ \text{s.t.} & \langle A_i, X \rangle = b_i, X \succeq 0 \end{array}$$

$p^* = d^*$ if primal SDP is strictly feasible.

SDP Relaxtion of Maxcut

$$
\begin{array}{ll} \min & x^\top W x \\ \text{s.t.} & x_i^2 = 1 \end{array}
\quad \Longrightarrow \quad
\begin{array}{ll} \max & -1^\top \nu \\ \text{s.t.} & W + \operatorname{diag}(\nu) \succeq 0 \end{array}
\quad \Longleftrightarrow \quad
\begin{array}{ll} \min & \operatorname{Tr}(WX) \\ \text{s.t.} & X_{ii} = 1 \\ & X \succeq 0 \end{array}
$$

- a nonconvex problem; feasible set contains 2n discrete points
- interpretation: partition $\{1, \ldots, n\}$ in two sets; W_{ij} is cost of assigning i, j to the same set;; $-W_{ij}$ is cost of assigning to different sets

dual function

$$
g(\nu) = \inf_x \left(x^\top W x + \sum_i \nu_i (x_i^2 - 1) \right) = \inf_x \; x^\top (W + \operatorname{diag}(\nu)) x - 1^\top \nu
$$

$$
= \begin{cases} -1^\top \nu & W + \operatorname{diag}(\nu) \succeq 0 \\ -\infty & \text{otherwise} \end{cases}
$$

Positive semidefinite unknown: SDP formulation

Suppose unknown matrix X is positive semidefinite

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \sigma_i(X) \\ \text{s.t.} \quad & X_{ij} = M_{ij} \quad (i,j) \in \Omega \\ & X \succcurlyeq 0 \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \min \quad & \text{trace}(X) \\ \text{s.t.} \quad & X_{ij} = M_{ij} \quad (i,j) \in \Omega \\ & X \succcurlyeq 0 \end{aligned}$$

Trace heuristic: Mesbahi & Papavassilopoulos (1997), Beck & D'Andrea (1998)

General SDP formulation

Let $X \in \mathbb{R}^{m\times n}$. For a given norm $\|\cdot\|$, the dual norm $\|\cdot\|_d$ is defined as

$$\|X\|_d := \sup\{\langle X, Y\rangle : Y \in \mathbb{R}^{m\times n}, \|Y\| \le 1\}$$

Nuclear norm and spectral norms are dual:

$$\|X\| := \sigma_1(X), \quad \|X\|_* = \sum_i \sigma_i(X).$$

(P) $\begin{array}{ll} \max\limits_Y & \langle X, Y\rangle \\ \text{s.t.} & \|Y\|_2 \le 1 \end{array}$ $\Leftrightarrow$ $\begin{array}{l} \max\limits_Y 2\langle X, Y\rangle \\ \text{s.t.} \begin{bmatrix} I_m & Y \\ Y^\top & I_n \end{bmatrix} \succcurlyeq 0 \end{array}$ $\Leftrightarrow$ $\begin{array}{l} \min\limits_Z -\left\langle Z, \begin{bmatrix} 0 & X \\ X^\top & 0 \end{bmatrix}\right\rangle \\ \text{s.t. } Z_1 = I_m \\ \quad Z_2 = I_n \\ \quad Z = \begin{bmatrix} Z_1 & Z_3 \\ Z_3^\top & Z_2 \end{bmatrix} \succeq 0 \end{array}$

General SDP formulation

The Lagrangian dual problem is:

$$\max_{W_1, W_2} \min_{Z \succeq 0} \quad - \left\langle Z, \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix} \right\rangle + \langle Z_1 - I_m, W_1 \rangle + \langle Z_2 - I_n, W_2 \rangle$$

strong duality after a scaling of $1/2$ and change of variables X to $-X$

$$\text{(D)} \quad \begin{array}{ll} \text{minimize} & \frac{1}{2}(\text{trace}(W_1) + \text{trace}(W_2)) \\ \text{subject to} & \begin{bmatrix} W_1 & X \\ X^\top & W_2 \end{bmatrix} \succcurlyeq 0 \end{array}$$

Optimization variables: $W_1 \in \mathbb{R}^{n_1 \times n_1}, W_2 \in \mathbb{R}^{n_2 \times n_2}$.

Proposition 2.1 in "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization", Benjamin Recht, Maryam Fazel, Pablo A. Parrilo

General SDP formulation

Nuclear norm minimization

$$\begin{aligned} &\min \ \|X\|_* \\ &\text{s.t. } \mathcal{A}(X) = b \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} &\max \ b^\top y \\ &\text{s.t. } \|\mathcal{A}^*(y)\| \leq 1 \end{aligned}$$

SDP Reformulation

$$\begin{aligned} &\min \ \frac{1}{2}(\text{trace}(W_1) + \text{trace}(W_2)) \\ &\text{s.t. } \mathcal{A}(X) = b \\ &\qquad \begin{bmatrix} W_1 & X \\ X^\top & W_2 \end{bmatrix} \succcurlyeq 0 \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} &\max \ b^\top y \\ &\text{s.t. } \begin{bmatrix} I & \mathcal{A}^*(y) \\ (\mathcal{A}^*(y))^\top & I \end{bmatrix} \succcurlyeq 0 \end{aligned}$$

Matrix recovery

$$M = \sum_{k=1}^{2} \sigma_k u_k u_k^*, \quad \begin{aligned} u_1 &= (e_1 + e_2)/\sqrt{2}, \\ u_2 &= (e_1 - e_2)/\sqrt{2} \end{aligned}$$

$$M = \begin{bmatrix} * & * & 0 & \dots & 0 & 0 \\ * & * & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

Cannot be recovered from a small set of entries

Rank-1 matrix $M = xy^*$

$$M_{ij} = x_i y_j$$

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ & & & & & \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \end{bmatrix}$$

If single row (or column) is not sampled $\rightarrow$ recovery is not possible

What happens for almost all sampling sets?

Ω subset of m entries selected uniformly at random

References

- Jianfeng Cai, Emmanuel Candes, Zuowei Shen, *Singular value thresholding algorithm for matrix completion*
- Shiqian Ma, Donald Goldfarb, Lifeng Chen, *Fixed point and Bregman iterative methods for matrix rank minimization*
- Zaiwen Wen, Wotao Yin, Yin Zhang, *Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm*
- Onureena Banerjee, Laurent El Ghaoui, Alexandre d'Aspremont, *Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data*
- Zhaosong Lu, *Smooth optimization approach for sparse covariance selection*

Matrix Rank Minimization

Given $X \in \mathbb{R}^{m\times n}$, $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, $b \in \mathbb{R}^p$, we consider

- the matrix rank minimization problem:

$$\min \ \mathrm{rank}(X), \ \text{s.t. } \mathcal{A}(X) = b$$

- matrix completion problem:

$$\min \ \mathrm{rank}(X), \ \text{s.t. } X_{ij} = M_{ij}, (i,j) \in \Omega$$

- nuclear norm minimization:

$$\min \ \|X\|_* \ \text{ s.t. } \mathcal{A}(X) = b$$

where $\|X\|_* = \sum_i \sigma_i$ and $\sigma_i = i$th singular value of matrix X.

Quadratic penalty framework

- Unconstrained Nuclear Norm Minimization:

$$\min\ F(X) := \mu\|X\|_* + \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2.$$

- Optimality condition:

$$\mathbf{0} \in \mu\partial\|X^*\|_* + \mathcal{A}^*(\mathcal{A}(X^*) - b),$$

  where $\partial\|X\|_* = \{UV^\top + W : U^\top W = 0, WV = 0, \|W\|_2 \leq 1\}$.
- Linearization approach (g is the gradient of $\frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$):

$$\begin{aligned} X^{k+1} &:= \arg\min_X\ \mu\|X\|_* + \langle g^k, X - X^k\rangle + \frac{1}{2\tau}\|X - X^k\|_F^2 \\ &= \arg\min_X\ \mu\|X\|_* + \frac{1}{2\tau}\|X - (X^k - \tau g^k)\|_F^2 \end{aligned}$$

Matrix Shrinkage Operator

For a matrix $Y \in \mathbb{R}^{m \times n}$, consider:

$$\min_{X \in \mathbb{R}^{m \times n}} \ \nu \|X\|_* + \frac{1}{2}\|X - Y\|_F^2.$$

The optimal solution is:

$$X := S_\nu(Y) = U\mathrm{Diag}(s_\nu(\sigma))V^\top,$$

- SVD: $Y = U\mathrm{Diag}(\sigma)V^\top$
- Thresholding operator:

$$s_\nu(x) := \bar{x}, \text{ with } \bar{x}_i = \begin{cases} x_i - \nu, & \text{if } x_i - \nu > 0 \\ 0, & \text{o.w.} \end{cases}$$

Fixed Point Method (Proximal gradient method)

Fixed Point Iterative Scheme

$$\begin{cases} Y^k = X^k - \tau \mathcal{A}^*(\mathcal{A}(X^k) - b) \\ X^{k+1} = S_{\tau\mu}(Y^k). \end{cases}$$

Lemma: Matrix shrinkage operator is non-expansive. i.e.,

$$\|S_\nu(Y_1) - S_\nu(Y_2)\|_F \leq \|Y_1 - Y_2\|_F.$$

Theorem: The sequence $\{X^k\}$ generated by the fixed point iterations converges to some $X^* \in \mathcal{X}^*$, where $\mathcal{X}^*$ is the optimal solution set.

SVT

Linearized Bregman method:

$$\begin{aligned} V^{k+1} &:= V^k - \tau\mathcal{A}^*(\mathcal{A}(X^k) - b) \\ X^{k+1} &:= S_{\tau\mu}(V^{k+1}) \end{aligned}$$

Convergence to

$$\min \ \tau\|X\|_* + \frac{1}{2}\|X\|_F^2, \text{ s.t. } \mathcal{A}(X) = b$$

Accelerated proximal gradient (APG) method

Complexity of the fixed point method:

$$F(X^k) - F(X^*) \le \frac{L_f \|X^0 - X^*\|^2}{2k}$$

APG algorithm ($t^{-1} = t^0 = 1$):

$$\begin{aligned}
Y^k &= X^k + \frac{t^{k-1} - 1}{t^k}(X^k - X^{k-1}) \\
G^k &= Y^k - (\tau^k)^{-1}\mathcal{A}^*(\mathcal{A}(Y^k) - b) \\
X^{k+1} &= S_{\tau^k}(G^k), \quad t^{k+1} = \frac{1 + \sqrt{1 + 4(t^k)^2}}{2}
\end{aligned}$$

Complexity:

$$F(X^k) - F(X^*) \le \frac{2L_f \|X^0 - X^*\|^2}{(k+1)^2}$$

Low-rank factorization model

- Finding a low-rank matrix W so that $\|\mathcal{P}_\Omega(W-M)\|_F^2$ or the distance between W and $\{Z \in \mathbb{R}^{m\times n}, Z_{ij} = M_{ij}, \forall (i,j) \in \Omega\}$ is minimized.
- Any matrix $W \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(W) \le K$ can be expressed as $W = XY$ where $X \in \mathbb{R}^{m\times K}$ and $Y \in \mathbb{R}^{K\times n}$.

New model

$$\min_{X,Y,Z} \frac{1}{2}\|XY - Z\|_F^2 \text{ s.t. } Z_{ij} = M_{ij}, \forall (i,j) \in \Omega$$

- Advantage: SVD is no longer needed!
- Related work: the solver `OptSpace` based on optimization on manifold

Nonlinear Gauss-Seidel scheme

First variant of alternating minimization:

$$\begin{aligned} X_+ &\leftarrow ZY^\dagger \equiv ZY^\top (YY^\top)^\dagger, \\ Y_+ &\leftarrow (X_+)^\dagger Z \equiv (X_+^\top X_+)^\dagger (X_+^\top Z), \\ Z_+ &\leftarrow X_+Y_+ + \mathcal{P}_\Omega(M - X_+Y_+). \end{aligned}$$

Let $\mathcal{P}_A$ be the orthogonal projection onto the range space $\mathcal{R}(A)$

- $X_+Y_+ = \left(X_+(X_+^\top X_+)^\dagger X_+^\top\right) Z = \mathcal{P}_{X_+} Z$
- One can verify that $\mathcal{R}(X_+) = \mathcal{R}(ZY^\top)$.
- $X_+Y_+ = \mathcal{P}_{ZY^\top} Z = ZY^\top (YZ^\top ZY^\top)^\dagger (YZ^\top) Z.$
- idea: modify X_+ or Y_+ to obtain the same product X_+Y_+

Nonlinear Gauss-Seidel scheme

Second variant of alternating minimization:

$$\begin{aligned} X_+ &\leftarrow ZY^\top, \\ Y_+ &\leftarrow (X_+)^\dagger Z \equiv (X_+^\top X_+)^\dagger (X_+^\top Z), \\ Z_+ &\leftarrow X_+ Y_+ + \mathcal{P}_\Omega(M - X_+ Y_+). \end{aligned}$$

Third variant of alternating minimization: $V = \mathtt{orth}(ZY^\top)$

$$\begin{aligned} X_+ &\leftarrow V, \\ Y_+ &\leftarrow V^\top Z, \\ Z_+ &\leftarrow X_+ Y_+ + \mathcal{P}_\Omega(M - X_+ Y_+). \end{aligned}$$

Sparse and low-rank matrix separation

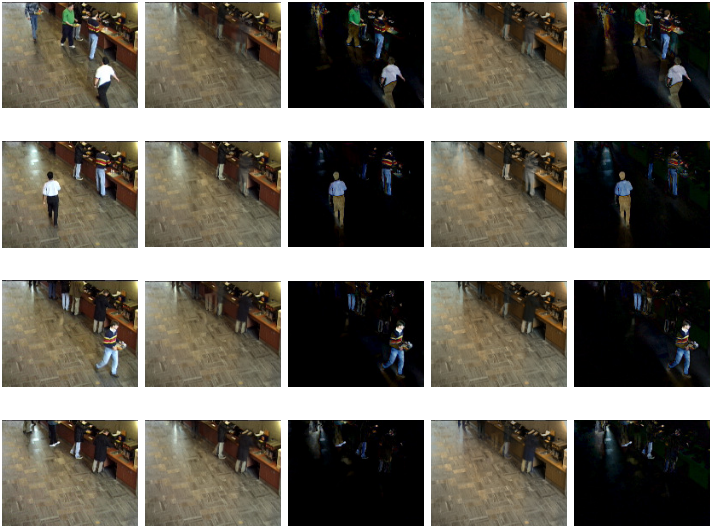
- Given a matrix M, we want to find a low rank matrix W and a sparse matrix E, so that $W + E = M$.
- Convex approximation:

$$\min_{W,E} \; \|W\|_* + \mu\|E\|_1, \text{ s.t. } W + E = M$$

- Robust PCA

Video separation

- Partition the video into moving and static parts



ADMM

Convex approximation:

$$\min_{W,E} \ \|W\|_* + \mu\|E\|_1, \text{ s.t. } W + E = M$$

Augmented Lagrangian function:

$$L(W, E, \Lambda) := \|W\|_* + \mu\|E\|_1 + \langle \Lambda, W + E - M \rangle + \frac{1}{2\beta}\|W + E - M\|_F^2$$

Alternating direction Augmented Lagrangian method

$$\begin{aligned}
W^{j+1} &:= \arg\min_W \ L(W,\ E^j,\ \Lambda^j), \\
E^{j+1} &:= \arg\min_E \ L(W^{j+1},\ E,\ \Lambda^j), \\
\Lambda^{j+1} &:= \Lambda^j + \frac{\gamma}{\beta}(W^{j+1} + E^{j+1} - M).
\end{aligned}$$

W-subproblem

Convex approximation:

$$
\begin{aligned}
W^{j+1} \quad &:= \quad \arg\min_W \; L(W,\; E^j,\; \Lambda^j) \\
&= \quad \arg\min_W \|W\|_* + \frac{1}{2\beta} \left\| W - \left(M - E^j - \beta\Lambda^j\right) \right\|_F^2 \\
&= \quad S_\beta(M - E^j - \beta\Lambda^j) := U\mathrm{Diag}(s_\beta(\sigma))V^\top
\end{aligned}
$$

- SVD: $M - E^j - \beta\Lambda^j = U\mathrm{Diag}(\sigma)V^\top$
- Thresholding operator:

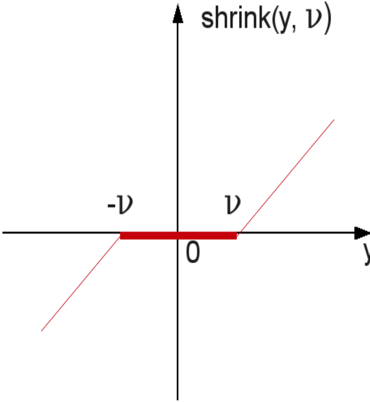
$$
s_\nu(x) := \bar{x}, \text{ with } \bar{x}_i = \begin{cases} x_i - \nu, & \text{if } x_i - \nu > 0 \\ 0, & \text{o.w.} \end{cases}
$$

E-subproblem

Convex approximation:

$$
\begin{aligned}
W^{j+1} &:= \arg\min_{E} \; L(W^{j+1},\; E,\; \Lambda^{j}) \\
&= \arg\min_{E} \|E\|_1 + \frac{1}{2\beta\mu} \left\| E - \left(M - W^{j+1} - \beta\Lambda^{j}\right) \right\|_F^2 \\
&= s_{\beta\mu}(M - W^{j+1} - \beta\Lambda^{j})
\end{aligned}
$$

$$
\begin{aligned}
s_{\nu}(y) : &= \arg\min_{x \in \mathbb{R}} \; \nu\|x\|_1 + \frac{1}{2}\|x - y\|_2^2 \\
&= \begin{cases} y - \nu \mathrm{sgn}(y), & \text{if } |y| > \nu \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$



Low-rank factorization model for matrix separation

- Consider the model

$$\min_{Z,S} \ \|S\|_1 \ \text{ s.t. } \ Z + S = D, \ \operatorname{rank}(Z) \leq K$$

- Low-rank factorization: $Z = UV$

$$\min_{U,V,Z} \ \|Z - D\|_1 \ \text{ s.t. } \ UV - Z = 0$$

- Only the entries D_{ij}, $(i,j) \in \Omega$, are given. $\mathcal{P}_\Omega(D)$ is the projection of D onto Ω.

New model

$$\min_{U,V,Z} \quad \|\mathcal{P}_\Omega(Z - D)\|_1 \quad \text{s.t.} \quad UV - Z = 0$$

- Advantage: SVD is no longer needed!

ADMM

Consider:

$$\min_{U,V,Z} \quad \|\mathcal{P}_\Omega(Z-D)\|_1 \quad \text{s.t.} \quad UV - Z = 0$$

Introduce the augmented Lagrangian function

$$\mathcal{L}_\beta(U,V,Z,\Lambda) = \|\mathcal{P}_\Omega(Z-D)\|_1 + \langle \Lambda, UV - Z\rangle + \frac{\beta}{2}\|UV - Z\|_F^2,$$

Alternating direction augmented Lagrangian framework (Bregman):

$$\begin{aligned}
U^{j+1} &:= \arg\min_{U\in\mathbb{R}^{m\times k}} \mathcal{L}_\beta(U,\ V^j,\ Z^j,\ \Lambda^j),\\
V^{j+1} &:= \arg\min_{V\in\mathbb{R}^{k\times n}} \mathcal{L}_\beta(U^{j+1},\ V,\ Z^j,\ \Lambda^j),\\
Z^{j+1} &:= \arg\min_{Z\in\mathbb{R}^{m\times n}} \mathcal{L}_\beta(U^{j+1},\ V^{j+1},\ Z,\ \Lambda^j),\\
\Lambda^{j+1} &:= \Lambda^j + \gamma\beta(U^{j+1}V^{j+1} - Z^{j+1}).
\end{aligned}$$

ADMM subproblems

- Let $B = Z - \Lambda/\beta$, then

$$U_+ = BV^\top (VV^\top)^\dagger \text{ and } V_+ = (U_+^\top U_+)^\dagger U_+^\top B$$

  Since $U_+V_+ = U_+(U_+^\top U_+)^\dagger U_+^\top B = \mathcal{P}_{U_+}B$, then:

$$Q := \texttt{orth}(BV^\top), \quad U_+ = Q \text{ and } V_+ = Q^\top B$$

- Variable Z:

$$\begin{aligned}
\mathcal{P}_\Omega(Z_+) &= \mathcal{P}_\Omega\left(\mathcal{S}\left(U_+V_+ - D + \frac{\Lambda}{\beta}, \frac{1}{\beta}\right) + D\right) \\
\mathcal{P}_{\Omega^c}(Z_+) &= \mathcal{P}_{\Omega^c}\left(U_+V_+ + \frac{\Lambda}{\beta}\right)
\end{aligned}$$

Nonnegative matrix factorization completion

Model problem:

$$\begin{array}{ll} \min & \|\mathcal{P}_\Omega(XY - M)\|_F \\ \text{s.t.} & X \geq 0, Y \geq 0, \end{array} \iff \begin{array}{ll} \min & \frac{1}{2}\|XY - Z\|_F^2 \\ \text{s.t.} & X = U, Y = V, \\ & U \geq 0,\ V \geq 0, \\ & \mathcal{P}_\Omega(Z - M) = 0 \end{array}$$

Augmented Lagrangian function:

$$\mathcal{L}_A(X, Y, Z, U, V, \Lambda, \Pi) = \frac{1}{2}\|XY - Z\|_F^2 + \Lambda \bullet (X - U) \\ + \Pi \bullet (Y - V) + \frac{\alpha}{2}\|X - U\|_F^2 + \frac{\beta}{2}\|Y - V\|_F^2,$$

where $A \bullet B := \sum_{i,j} a_{ij} b_{ij}$.

ADMM

$$
\begin{aligned}
X_{k+1} &= (Z_k Y_k^T + \alpha U_k - \Lambda_k)(Y_k Y_k^T + \alpha I)^{-1}, && \text{(3a)}\\
Y_{k+1} &= (X_{k+1}^T X_{k+1} + \beta I)^{-1}(X_{k+1}^T Z_k + \beta V_k - \Pi_k), && \text{(3b)}\\
Z_{k+1} &= X_{k+1} Y_{k+1} + \mathcal{P}_\Omega(M - X_{k+1} Y_{k+1}), && \text{(3c)}\\
U_{k+1} &= \mathcal{P}_+(X_{k+1} + \Lambda_k/\alpha), && \text{(3d)}\\
V_{k+1} &= \mathcal{P}_+(Y_{k+1} + \Pi_k/\beta), && \text{(3e)}\\
\Lambda_{k+1} &= \Lambda_k + \gamma\alpha(X_{k+1} - U_{k+1}), && \text{(3f)}\\
\Pi_{k+1} &= \Pi_k + \gamma\beta(Y_{k+1} - V_{k+1}), && \text{(3g)}
\end{aligned}
$$

where $\gamma \in (0, 1.618)$ and $(\mathcal{P}_+(A))_{ij} = \max\{a_{ij}, 0\}$.

Sparse covariance selection (A. d'Aspremont)

We estimate a covariance matrix Σ from empirical data

- Infer independence relationships between variables
- Given $m+1$ observations $x_i \in \mathbb{R}^n$ on n random variables, compute $S := \frac{1}{m}\sum_{i=1}^{m+1}(x_i - \bar{x})(x_i - \bar{x})$
- Choose a symmetric subset I of matrix coefficients and denote by J the complement
- Choose a covariance matrix $\hat{\Sigma}$ such that
  - $\hat{\Sigma}_{ij} = S_{ij}$ for all $(i,j) \in I$
  - $\hat{\Sigma}_{ij}^{-1} = 0$ for all $(i,j) \in J$
- Benefits: maximum entropy, maximum likelihood, existence and uniqueness
- Applications: Gene expression data, speech recoginition and finance

Maximum likelihood estimation

Consider estimation:

$$\max_{X \in S^n} \log\det X - \mathrm{Tr}(SX) - \rho\|X\|_0$$

Convex relaxations:

$$\max_{X \in S^n} \log\det X - \mathrm{Tr}(SX) - \rho\|X\|_1,$$

whose dual problem is:

$$\max \log\det W \ \text{ s.t. } \|W - S\|_\infty \le \lambda$$

APG

Zhaosong Lu (*smooth optimization approach for sparse covariance selection*) consider

$$\begin{aligned} \max \ & \log\det X - \mathrm{Tr}(SX) - \rho\|X\|_1 \\ \text{s.t.} \ & \mathcal{X} := \{X \in S^n : \beta I \succeq X \succeq \alpha I\}, \end{aligned}$$

which is equivalent to ($\mathcal{U} := \{U \in S^n : |U_{ij}| \leq 1, \forall ij\}$)

$$\max_{X\in\mathcal{X}} \min_{U\in\mathcal{U}} \log\det X - \langle S + \rho U, X\rangle$$

Let $f(U) := \max_{X\in\mathcal{X}} \log\det X - \langle S + \rho U, X\rangle$

- $\log\det X$ is strongly concave on $\mathcal{X}$
- $f(U)$ is continuous differentiable
- $\nabla f(U)$ is Lipschitz cont. with $L = \rho\beta^2$

Therefore, APG can be applied to the dual problem

$$\min_{U\in\mathcal{U}} f(U)$$

Extension

Consider

$$\max_{x \in \mathcal{X}} \; g(x) := \min_{u \in \mathcal{U}} \phi(x,u)$$

Assume:

- $\phi(x,u)$ is a cont. fun. which is strictly concave in $x \in \mathcal{X}$ for every fixed $u \in \mathcal{U}$, and convex diff. in $u \in \mathcal{U}$ for every fixed $x \in \mathcal{X}$. Then $f(u) := \max_{x \in \mathcal{X}} \phi(x,u)$ is diff.
- $\nabla f(u)$ is Lipschitz cont.

Then

- the primal and the dual $\min_{u \in U} f(u)$ are both solvable and have the same optimal value;
- Nesterov's smooth minimization approach can be applied to the dual

Nesterov's smoothing technique

Consider

$$\max_{x \in \mathcal{X}} \min_{u \in \mathcal{U}} \phi(x, u)$$

Question: What if the assumptions do not hold?

- Add a strictly convex function $\mu d(u)$ to the obj. fun.

$$g(u) := \arg\min_{u \in \mathcal{U}} \phi(x, u) + \mu d(u)$$

- $g(u)$ is differentiable
- Apply Nesterov's smooth minimization
- Complexity of finding a ϵ-suboptimal point: $O(\frac{1}{\epsilon})$ iterations
- Other smooth technique?