Sparse Matrix Decomposition

Until now, we've been focusing on "simple" vectors:

$$\min_x \|y - x\|^2 \text{ s.t. } x \in \mathcal{C} : \{\text{"set of simple vectors"}\}$$

$$\mathcal{C} : \{x : \|x\|_0 \leq k.\}.$$

What if now we formulate this for matrices?

$$\min_M \|Z - M\|_F^2 \text{ s.t. } M \in \mathcal{C}$$

- Option 1: sparse matrices $\Rightarrow \mathcal{C} : \{M : \|M\|_0 \leq K\}$.

$$\Rightarrow \min_M \|Z - M\|_F^2 \text{ st. } \|M\|_0 \leq k$$

or

$$(P_1^\lambda): \quad \min_M \|Z - M\|^2 + \lambda \|M\|_1$$

$$\hat{M} = S_\lambda(Z) \Rightarrow \operatorname{sign}(z_{ij})(|z_{ij}| - \lambda)_+ = M_{ij}.$$

Option 2: Just like vectors, they aren't sparse themselves, but rather in some other domain. We will use some "natural decomposition".

Consider $X \in \mathbb{R}^{D \times N}$, and we want

$$X \simeq \sum_{i=1} A_i \cdot c_i \; ; \; A_i \in \mathbb{R}^{D \times N} ; \; c_i \in \mathbb{R}.$$

↳ "basic" matrices.

"Basic" because $A_i = (u_i v_i^T)$, $u_i \in \mathbb{R}^D$, $v_i \in \mathbb{R}^N$

$\Rightarrow A_i$: rank 1 matrices.

Thus, let's say our "simplicity" pursuit problem becomes:

$$\min_{A_i, c_i} \| X - \sum_{i=1}^{d} c_i A_i \|_2^2 \quad \text{s.t.} \quad \text{rank}(A_i) = 1$$

$$d < \min(N, D).$$

This is the same as

$$\min_{A} \| X - A \|_F^2 \quad \text{s.t.} \quad \text{rank}(A) \leq d.$$

Recall that $\forall \cancel{A} : \cancel{A} = U \Sigma V^T$; Σ: diagonal $D \times D$.

and $\text{rank}(X) = \| \Sigma \|_0$. Thus,

$$\boxed{\min_{A} \| X - A \|_F^2 \quad \text{s.t.} \quad \| \Sigma \|_0 \leq d.}$$ where $A = U \Sigma V^T$

Assume hereafter that X is "centered" — column mean $= 0$.

How to solve?

Note:

$$\min_{A} \| X - A \|_F^2 = \min_{U, \Sigma_A, V_A} \| \Sigma_X - U \Sigma_A V_A^T \|_F^2 \qquad \begin{cases} U = U_X^T U_A \\ V = V_X^T V_A \end{cases}$$

$$= \min_{U, \Sigma_A, V_A} \| \Sigma_X \|_F^2 - 2 \langle \Sigma_X, U \Sigma_A V^T \rangle + \| \Sigma_A \|_F^2$$

Assume Σ_A : fixed.

$$\Rightarrow \max_{U, V} \langle \Sigma_X, U \Sigma_A V^T \rangle.$$

From Van Neumann's Inequality, $\langle F, G \rangle \leq \sum_{i=1}^{n} \sigma_i(F) \sigma(G)$ $\quad tr(F^T G) =$

and equality holds iff $\exists\, V, U$: $F = U \Sigma_F V^T$, $G = U \Sigma_G V^T$.

$$\Rightarrow \max_{U, V} \langle \Sigma_X, U \Sigma_A V^T \rangle \leq \sum_{i=1}^{d} \sigma(\Sigma_X)_i \, \sigma(A)_i$$

To maximize this, sig. vectors should be the same

$U = U_x^T U_A = I \quad \wedge \quad V = V_x^T V_A = I.$ (2)

$\Rightarrow U_A = U_x, \quad V_A = V_x.$

$\Rightarrow$ from here also: $\min \| X - A\|_F^2 \Rightarrow \min \| \Sigma_x - \Sigma_A\|_F^2$ s.t. $\|\Sigma_A\|_0 \leq d$.

Thus, $\min_A \|X - A\|_F^2 \Rightarrow \min_{\Sigma_A} \|\Sigma_A\|_F^2 - 2\langle \Sigma_x, \Sigma_A \rangle.$

$\min_{\sigma_i(A)} \sum_i^d \sigma_i(A)^2 - \sigma_i(X)\sigma(A)$

$\Rightarrow \sigma_i(A) = \sigma_i(X)$. for $1 \leq i \leq d$

Thus, $A_d = \arg\min_A \|X - A\|_F^2$ s.t. $\|\Sigma_x\|_0 \leq d$

$A_d = U_d \Sigma_d V_d^T \qquad U_d, V_d$: top d left and right s-v.

Note that we can write $A = U_x H_d(\Sigma_x) V_x^T$. //

(Also, look @ Mantas Mazeika, Sing. V. Decomp. and low rank approx.).

Matrix Completion

Say we now observe only a subset of entries in X, Ω: $P_\Omega(X)$.

$\min_A \|P_\Omega(X) - P_\Omega(A)\|_F^2 \Rightarrow$ infinite solutions!

$P_\Omega(X)_{ij} = \begin{cases} X_{i,j} & \forall (i,j) \in \Omega. \\ 0 & \text{otherwise}. \end{cases}$

$\Rightarrow \min_A \|P_\Omega(X) - P_\Omega(A)\|_F^2$ st. $rk(A) \leq d$.

This is now NP-hard, however.

Some Heuristic: iterative projections:

init: $A^{(0)} = X$

iter: - $Z \leftarrow \arg\min_Z \|A^{(k)} - Z\|_F^2$ st. $rk(Z) \leq d$. (low-rank app)

- $A^{k+1} \leftarrow P_\Omega(X) + Z_{\Omega^c}$ (better estimate closer to measurements)

depend greatly on initialization and not very robust.

Just Relax (again).

Instead of doing $\min_A \| P_\Omega(X) - P_\Omega(A) \|_F^2$ s.t. $\| \Sigma_A \|_0 \leq d$

do $\min_A \| P_\Omega(X) - P_\Omega(A) \|_F^2 + \lambda \underbrace{\| \Sigma_A \|_1}$

$\sum_i \sigma_i(A) = \|A\|_*$.

"Nuclear Norm" or "trace norm".

Let's start by looking at the simpler problem:

$$\min_X \| Y - X \|_F^2 + \lambda \|X\|_* \quad ; \ (P_\lambda^*).$$

$\to$ convex – why?

Recall that this is equivalent to

$$\min_X \| \Sigma_Y - U \Sigma_X V^T \|_F^2 + \lambda \|X\|_*$$

$$\min_X \| \Sigma_Y - \Sigma_X \|_F^2 + \lambda \sum_i \sigma_i(X) \quad \text{from Von Neumann's.}$$

define $S_\cdot = \text{diag}(\Sigma_\cdot)$, then

$$\min \| S_Y - S_X \|_2^2 + \lambda \| S_Y \|_1$$

$$\Rightarrow \quad \hat{S}_X = S_\lambda(S_Y) \quad \text{- soft thresholding the singular values.}$$

Thus, $\hat{X} = U S_\lambda(\Sigma_Y) V^T = \arg\min_X \| Y - X \|_F^2 + \lambda \| X \|_*$.

$\hat{X} = D_\lambda(Y) = \text{prox}_{\lambda \|\cdot\|_*}(Y)$

An alternative (interesting) proof) is in Cai 2005.

Given $h(Y) = \min_X \| Y - X \|_F^2 + \lambda \|X\|_*$, recall that

$$Z \in \partial f(X_0) \text{ if } f(X) \geq f(X_0) + \langle Z, X - X_0 \rangle \quad \forall X.$$

and $\hat{X} = S_\lambda(Y)$ minimizes $h(Y)$ iif $0 \in \partial h(\hat{X})$:

$$0 \in \hat{X} - Y + \lambda \, \partial \| \hat{X} \|_* \qquad (1)$$

Let $X = U \Sigma V^T$. It is known that $(X \in \mathbb{R}^{n \times m})$

$$\partial \|X\|_* = \{ UV^T + W : W \in \mathbb{R}^{n \times m},\ U^T W = 0,\ WV = 0,\ \|W\|_2 \leq 1 \}.$$

Excersise: Verify that indeed (1) holds.
for $\hat{X} = U S_\lambda(\Sigma) V^T$.

So, going back to $(P^*_{\lambda,\Omega})$: $\min_{X} \| P_\Omega(X) - P_\Omega(Y) \|_2^2 + \lambda \|A\|_*$

This is convex and can be solved with standard (Semi Def. Program) but its expensive. Better: proximal gradient Descent:

Given Y, init $\hat{X} = 0$, then:

$\cancel{X^{k+1} = S_\lambda(P_\Omega(Y) + P_\Omega^\perp(X^k))}$

$$\Rightarrow X^{k+1} = \operatorname{prox}_{\lambda\|\cdot\|_*}\left(X^k - \nabla f(X^k)\right).$$

$$X^{k+1} = S_\lambda^s\left(X^k - \left(P_\Omega(X^k) - P_\Omega(Y)\right)\right)$$

Noting that $P_\Omega^\perp(X^k) = X^k - P_\Omega(X^k)$

$$\Rightarrow \boxed{X^{k+1} = S_\lambda^s\left(P_\Omega(Y) + P_\Omega^\perp(X^k)\right)};$$

This converges to global optimum with rate $O(1/k)$.

Even better, note that $P_\Omega(Y) + P_\Omega^\perp(X^k) = \underbrace{P_\Omega(Y) - P_\Omega(X^k)}_{\text{sparse}} + \underbrace{X^k}_{\text{low rank}}$.

So this is efficient.

Guarantees for matrix Completion

- No noise.

- How many samples from a low rank matrix do we need to recover it exactly?

- First observation: if ~~do~~ no observations of a row ~~and~~ or columns, then it's impossible to recover the matrix exactly even if it is of rank 1.

- So how many samples do we need to observe at least 1 entry from every row & column? (with high probability)

$$N > p \log p \qquad (p \times p \text{ matrix}).$$

actually: $\Pr[N > \beta p \log p] \leq p^{-\beta}$

$\Rightarrow$ "coupon collector problem".

Second Observation:

We need a "good" matrix, => incoherent with the canonical basis.

Consider $Z = e_1 e_1^T := \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 0 & & \\ 0 & & \ddots & \\ \vdots & & & \end{pmatrix}$ => if we observe $N \ll p^2$ we won't see anything.

Intuition: we need the singular vectors to be "spread out".

Def: Matrix incoherence:

$X \in \mathbb{R}^{p \times m}$ is ν-incoherent w.r.t. sparse matrices if
G of rank $= r$.

$$\max_i \|u_i^T\|_2 \leq \nu \sqrt{\frac{r}{p}}, \quad \max_j \|v_j^T\|_2 \leq \nu \sqrt{\frac{r}{m}},$$

$$\|UV^T\|_\infty \leq \frac{\nu \sqrt{r}}{\sqrt{pm}}.$$

u_i^T, v_j^T are the rows of U, V.

=> This measures "spikyness"; it's low if entries in U, V are not concentrated.

Note that $\|u_i\|_2 = 1$ (columns). If these are concentrated $(U = I)$ then $\max_i \|u_i^T\|_2 = 1 \Rightarrow \nu \geq \sqrt{p/r} > 1.$ $(\gg 1)$

If u_i is "spread out" maximally $\Rightarrow u_i = \pm\sqrt{1/p} \Rightarrow \nu \gtrsim 1.$
$\nu = 1$

In fact, this property is satisfied for Gaussian matrices as long as p is large enough.

Low Rank Matrix Recovery through Convex Optim.

Let $X \in \mathbb{R}^{p \times p}$, ν-incoherent, and M is the expected number of observed entries sampled at random (e.g. $M = p^2 \cdot t$)

prob. that a sample is observed.

Then, $\exists$ c: constant such that if
numerical

$$M \geq c\, \nu^4 d\, p (\log p)^2 = \Omega(dp \operatorname{polylog}(p)).$$

[Candes & Recht, '09 Candes & Tao '10]

then $X = \underset{A}{\operatorname{argmin}} \|A\|_*$ s.t. $P_\Omega(X) = P_\Omega(A)$. with $\Pr \geq 1 - p^{-3}$

Note that this is pretty tight:

How many degrees of freedom in a low rank matrix?
r $p \times p$

$$\text{DoF} = 2rp - r^2$$

Say we choose the first r columns of dim p (so $p \cdot r$). Say these are l.i.
Then, for the reminder colums, they will be linear combinations of the previos r colums. (r coeff)

Thus $\text{DoF} = rp + (p - r) \cdot r = 2rp - r^2$.

So up to a poly log factor, this is the lower bound.

Robust PCA:

So convex relaxation of the rank can recover a low rank matrix if there's no noise. If some entries in X are corrupted, then the l_2 will be very sensitive.

Upgrading this model: $X = L_0 + E_0$, where $\text{rank}(L_0) \le r$.

E_0: "outliers"

Thus: and sparse $\|E\|_0 \le k$.

$$\min_{L,E} \text{rank}(L) + \lambda \|E\|_0 \quad \text{s.t. } X = L + E.$$

This looks impossible!:

- p^2 equations (linear) for $2p^2$ unknowns.
- both losses are non-convex
- Solution might not be unique:

imagine $Z = e_1 e_1^T \Rightarrow (L,E) = (Z,0)$ or $(L,E) = (0,Z)$?

The wellposed-ness of this problem will naturally depend on similar "incoherence" properties for the L_0, and the outliers cannot be "conspicuosly located", say block and entire column/row.

$\Rightarrow$ Double Relax!

Principal Component Pursuit (PCP):

$$\boxed{\min_{L,E} \|L\|_* + \lambda \|E\|_1 \quad \text{s.t. } X = L + E}$$

Guarantees: Say $X: p \times p$. ; $L_0 = U \Sigma V^T$, $U, V: p \times r$,

and $supp(E_0)$ is uniformly distributed, if

$$\operatorname{rank}(L_0) \leq \frac{a \cdot p}{\nu^* \log^2 p} \quad \text{and} \quad \|E\|_0 \leq b \cdot p^2$$

for some constants $\{a, b\} > 0$, numerical constants. $\Rightarrow \exists\, c > 0$: $\hat{L} = L_0$, $\hat{E} = E_0$ with

$P_r \geq 1 - c p^{-10}$ (as long as $\lambda = \frac{1}{\sqrt{p}}$). [Candes '11].