Project Title: A Globally Optimal Solution to Low Rank Regression

We provide a closed form solution and global minimizer to the problem of rank-constrained least squares.

1 Problem

We are interested in solving the nonconvex rank-constrained least squares problem

$$\begin{aligned} \underset{W \in \mathbb{R}^{p \times n}}{\operatorname{argmin}} \ & \|Y - XW\|_F^2 \\ \text{s.t. } & \operatorname{rank}(W) \leq k \end{aligned} \tag{1}$$

where $Y \in \mathbb{R}^{m \times n}, X \in \mathbb{R}^{m \times p}$ are given. In this work, we derive an analytical solution to this problem.

2 Notation

Here is some notation. Let $A \in \mathbb{R}^{m \times n}$ be a real-valued matrix.

- $\mathcal{P}_k(A)$ is the k-rank approximation to A which is given by $\mathcal{P}_k(A) = \sum_{i=1}^k \sigma_i u_i v_i^\intercal$. $\mathcal{P}_k(A)$ is the projection of A onto matrices of rank less than or equal to k.
- $[A]_k$ is the matrix of first k rows of A. In Matlab notation, this looks like $A(1:k,\ :)$.
- $[A]^k$ is the matrix of first k columns of A. In Matlab notation, this looks like $A(:,\ 1:k)$.
- Let $\Sigma \in \mathbb{R}^{m \times n}$ be a real-valued matrix. Define Σ_k to be the $k \times k$ leading principal submatrix of Σ obtained by retaining the first k rows and columns and deleting everything else.

3 Intermediate Results

First, we recall $k-$rank matrix approximations.

Lemma 3.1. *Let $A \in \mathbb{R}^{m \times n}$ be a real-valued matrix. The famous Eckart-Young-Mirsky theorem proves that*

$$\begin{aligned} \mathcal{P}_k(A) = \underset{X \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \ & \|A - X\|_F^2 \\ \text{s.t. } & \operatorname{rank}(X) \leq k \end{aligned} \tag{2}$$

In other words, the $k-$rank matrix approximation is optimal in the Frobenius norm sense. It is also optimal in the spectral norm sense.

Now for the main result. Here we prove an analytical solution to the nonconvex rank-constrained least squares:

Theorem 3.2. *Let $X \in \mathbb{R}^{m \times p}$ be the design matrix, and $Y \in \mathbb{R}^{m \times n}$ be the output matrix. Let $U\Sigma V^\intercal$ be the full SVD of X. Note that we do not assume X is full rank. Recall the problem (1) we want to solve:*

$$\begin{aligned} \underset{W \in \mathbb{R}^{p \times n}}{\operatorname{argmin}} \ & \|Y - XW\|_F^2 \\ \text{s.t. } & \operatorname{rank}(W) \leq k \end{aligned}$$

A closed form solution and global minimizer to this problem is

$$W^* = V \begin{bmatrix} \Sigma_{r(X)}^{-1} \mathcal{P}_k \left(\ [U^\intercal Y]_{r(X)} \ \right) \\ \mathbf{0} \end{bmatrix}$$

Note that when $k = r(X)$, we recover the pseudoinverse solution, $X^\dagger Y$. When X is skinny and full rank, this is equivalent to the least squares solution. When X is fat and full rank, this is equivalent to the least norm solution. The least norm problem is

$$W^* = X^\dagger Y = \underset{W}{\text{argmin}} \, \|W\|_F^2 \;\; \text{s.t. } Y = XW$$

Proof. First, note that

$$\|Y - XW\|_F^2 = \|Y - U\Sigma V^\intercal W\|_F^2 = \|U^\intercal Y - \Sigma V^\intercal W\|_F^2 = \|\tilde{Y} - \Sigma \tilde{W}\|_F^2$$

where $\tilde{Y} = U^\intercal Y, \tilde{W} = V^\intercal W$. The above equalities hold because the Frobenius norm is invariant under rotations. It is key to note that $\text{rank}(\tilde{W}) = \text{rank}(V^\intercal W) = \text{rank}(W)$. It is equivalent to consider the problem

$$\begin{aligned} &\underset{\tilde{W}}{\text{argmin}} \; \|\tilde{Y} - \Sigma \tilde{W}\|_F^2 \\ &\quad \text{s.t. } \text{rank}(\tilde{W}) \leq k \end{aligned} \tag{3}$$

It is key to note that

$$\Sigma \tilde{W} = \begin{bmatrix} \Sigma_{r(X)} [\tilde{W}]_{r(X)} \\ \mathbf{0} \end{bmatrix}$$

If X is skinny ($m \geq p$), then we have

$$\Sigma \tilde{W} = \begin{bmatrix} \Sigma_p \\ \mathbf{0} \end{bmatrix} \tilde{W} = \begin{bmatrix} \Sigma_p \tilde{W} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \Sigma_{r(X)} [\tilde{W}]_{r(X)} \\ \mathbf{0} \end{bmatrix}$$

If X is fat ($m < p$), then we have

$$\Sigma \tilde{W} = \begin{bmatrix} \Sigma_m & \mathbf{0} \end{bmatrix} \tilde{W} = \begin{bmatrix} \Sigma_{r(X)} [\tilde{W}]_{r(X)} \\ \mathbf{0} \end{bmatrix}$$

The bottom $m - r(X)$ rows of the argument to the Frobenius norm, $\tilde{Y} - 0$, does not depend on $\tilde{W}$, so eq (3) can now be rewritten as

$$\begin{aligned} &\underset{\tilde{W}}{\text{argmin}} \; \left\| [\tilde{Y}]_{r(X)} - \Sigma_{r(X)} [\tilde{W}]_{r(X)} \right\|_F^2 \\ &\text{s.t. } \text{rank}(\tilde{W}) \leq k \end{aligned} \tag{4}$$

Another key thing to note here is that $\text{rank}(\tilde{W}) \geq \text{rank}([\tilde{W}]_{r(X)})$, but if we set

$$\tilde{W} = \begin{bmatrix} [\tilde{W}]_{r(X)} \\ \mathbf{0} \end{bmatrix}$$

then we have that $\text{rank}(\tilde{W}) = \text{rank}([\tilde{W}]_{r(X)})$. We rewrite Eq (4) to solve an even simpler problem

$$\begin{aligned} \underset{[\tilde{W}]_{r(X)}}{\text{argmin}} \quad & \left\| [\tilde{Y}]_{r(X)} - \Sigma_{r(X)} [\tilde{W}]_{r(X)} \right\|_F^2 \\ \text{s.t. } & \text{rank}([\tilde{W}]_{r(X)}) \leq k \end{aligned} \tag{5}$$

$\Sigma_{r(X)}$ is a diagonal positive definite matrix, so we can make the substitution $\hat{W} = \Sigma_{r(X)}[\tilde{W}]_{r(X)}$ (note that $\text{rank}(\hat{W}) = \text{rank}(\Sigma_{r(X)}[\tilde{W}]_{r(X)})$) and solve

$$\begin{aligned} \underset{\hat{W}}{\text{argmin}} \quad & \left\| [\tilde{Y}]_{r(X)} - \hat{W} \right\|_F^2 \\ \text{s.t. } & \text{rank}(\hat{W}) \leq k \end{aligned} \tag{6}$$

We invoke the Lemma 3.1 and chain the substitutions together to get

$$\begin{aligned} \hat{W}^* &= \mathcal{P}_k([\tilde{Y}]_{r(X)}) \\ [\tilde{W}]^*_{r(X)} &= \Sigma^{-1}_{r(X)} \mathcal{P}_k([\tilde{Y}]_{r(X)}) \\ \tilde{W}^* &= \begin{bmatrix} \Sigma^{-1}_{r(X)} \mathcal{P}_k([\tilde{Y}]_{r(X)}) \\ \mathbf{0} \end{bmatrix} \\ W^* &= V \begin{bmatrix} \Sigma^{-1}_{r(X)} \mathcal{P}_k([\tilde{Y}]_{r(X)}) \\ \mathbf{0} \end{bmatrix} \\ &= V \begin{bmatrix} \Sigma^{-1}_{r(X)} \mathcal{P}_k(\ [U^\intercal Y]_{r(X)} \ ) \\ \mathbf{0} \end{bmatrix} \end{aligned}$$

This concludes the proof. □

A small generalization of this result arises when we consider the case when W is sandwiched by two matrices:

Theorem 3.3. *Let $B \in \mathbb{R}^{m \times r}, \Phi \in \mathbb{R}^{s \times n}$, and $Y \in \mathbb{R}^{m \times n}$ be the output matrix. Let $B = U \Sigma V^\intercal$ and $\Phi = GDH^\intercal$ be the full SVD. Note that we do not assume B nor Φ are full rank. The slightly more general problem we'd like to solve is*

$$\begin{aligned} \underset{W \in \mathbb{R}^{r \times s}}{\text{argmin}} \quad & \| Y - BW\Phi \|_F^2 \\ \text{s.t. } & \text{rank}(W) \leq k \end{aligned} \tag{7}$$

A closed form solution and global minimizer is given by

$$W^* = V \begin{bmatrix} \Sigma^{-1}_{r(B)} \mathcal{P}_k \left(\ [U^\intercal Y H]^{r(\Phi)}_{r(B)} \ \right) D^{-1}_{r(\Phi)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G^\intercal$$

Note that when $\Phi = I_s$, our solution defaults to the solution of Theorem 3.2.

Proof. The proof of this is very similar to the proof of Theorem 3.2. We start by rewriting the loss function.

$$\begin{aligned} \|Y - BW\Phi\|_F^2 &= \|Y - U\Sigma V^\intercal WGDH^\intercal\|_F^2 = \|U^\intercal YH - \Sigma V^\intercal WGD\|_F^2 \\ &= \|\tilde{Y} - \Sigma\tilde{W}D\|_F^2 \end{aligned}$$

where $\tilde{Y} = U^\intercal YH$ and $\tilde{W} = V^\intercal WG$. By similar arguments as in the previous proof,

$$\Sigma\tilde{W}D = \begin{bmatrix} \Sigma_{r(B)}[\tilde{W}]_{r(B)}^{r(\Phi)}D_{r(\Phi)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Eq (7) can now be rewritten as

$$\begin{aligned} &\underset{\tilde{W}\in\mathbb{R}^{r\times s}}{\operatorname{argmin}} \left\| [\tilde{Y}]_{r(B)}^{r(\Phi)} - \Sigma_{r(B)}[\tilde{W}]_{r(B)}^{r(\Phi)}D_{r(\Phi)} \right\|_F^2 \\ &\quad \text{s.t. } \operatorname{rank}(\tilde{W}) \le k \end{aligned} \tag{8}$$

Set $\hat{W} = \Sigma_{r(B)}[\tilde{W}]_{r(B)}^{r(\Phi)}D_{r(\Phi)}$, then solve

$$\begin{aligned} &\underset{\hat{W}\in\mathbb{R}^{r(B)\times r(\Phi)}}{\operatorname{argmin}} \left\| [\tilde{Y}]_{r(B)}^{r(\Phi)} - \hat{W} \right\|_F^2 \\ &\quad \text{s.t. } \operatorname{rank}(\hat{W}) \le k \end{aligned}$$

Invoking lemma 3.1 and chaining the substitutions together gives

$$\begin{aligned} \hat{W}^* &= \mathcal{P}_k\left([\tilde{Y}]_{r(B)}^{r(\Phi)} \right) \\ \left([\tilde{W}]_{r(B)}^{r(\Phi)}\right)^* &= \Sigma_{r(B)}^{-1}\mathcal{P}_k\left([\tilde{Y}]_{r(B)}^{r(\Phi)} \right) D_{r(\Phi)}^{-1} \\ \tilde{W}^* &= \begin{bmatrix} \Sigma_{r(B)}^{-1}\mathcal{P}_k\left([\tilde{Y}]_{r(B)}^{r(\Phi)} \right) D_{r(\Phi)}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ W^* &= V\begin{bmatrix} \Sigma_{r(B)}^{-1}\mathcal{P}_k\left([\tilde{Y}]_{r(B)}^{r(\Phi)} \right) D_{r(\Phi)}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G^\intercal \\ &= V\begin{bmatrix} \Sigma_{r(B)}^{-1}\mathcal{P}_k\left([U^\intercal YH]_{r(B)}^{r(\Phi)} \right) D_{r(\Phi)}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G^\intercal \end{aligned}$$

□

Ideas for Next Steps:

- Comparing analytical solution vs. nuclear norm regularization when the data is slightly perturbed.
- Explore Randomized SVD guarantees plugged into inexact proximal gradient guarantees to get convergence rates, perhaps with high probability.