Minimization over Stiefel manifolds: Robust PCA and eigenvalue problem

Arshak Minasyan

Supervisor: Prof. Artur Hovhannisyan

American University of Armenia

arshak_minasyan@edu.aua.am

January 10, 2019

Introduction & Motivation

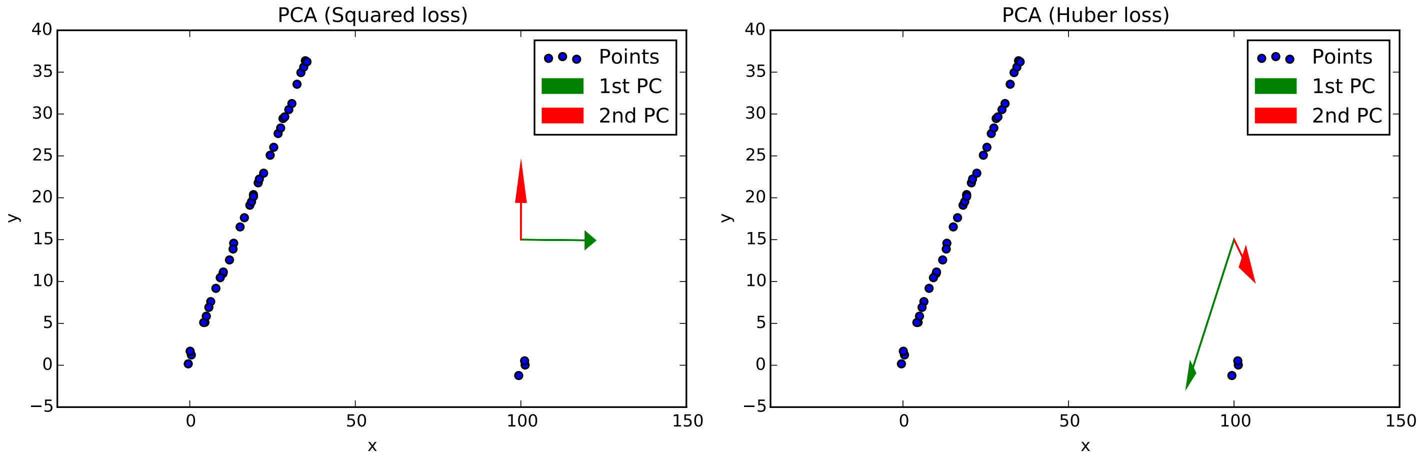


Figure 1: Outlier impact illustration

In this work we discuss the same "robust" setup and apply two completely different optimization methods for solving the proposed optimization problem.

Principal Component Analysis (PCA)

Let $\mathcal{C}_D = \left\{x_1, \ldots, x_N \in \mathbb{R}^D\right\}$, where $D \gg 1$.

$$\overline{x} := \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \Sigma := \frac{1}{N}(x_i - \overline{x})(x_i - \overline{x})^T \in \mathbb{R}^{D\times D}.$$

PCA uses the result of Eckart-Young theorem about SVD to obtain the subspace spaned by eigenvectors of Σ:

$$\Sigma v_i = \lambda_i v_i, \quad \lambda_1 \geq \cdots \geq \lambda_r \geq \cdots \geq \lambda_D \geq 0,$$

where $v_1, \ldots, v_r$ are the eigenvectors of matrix Σ.

Principal Component Analysis (PCA)

Equivalent formulation: aim to find the hyperplane

$$c^T x = \beta, \quad c \in \mathbb{R}^D, \quad \beta \in \mathbb{R} \tag{1}$$

with ℓ_2 best approximation for $\mathcal{C}_D$. The distance between $x_i \in \mathcal{C}_D$ and the hyperplane (1) is $|c^T x_i - \beta|$ given $\|c\|_2 = 1$. Hence,

$$\min_{b, CC^T = I} \sum_{i=1}^{N} \|Cx_i - b\|_2^2, \tag{2}$$

where $C \in \mathbb{M}_{m,D}(\mathbb{R}), b \in \mathbb{R}^m$ yielding the subspace $Cx = b$ with m equations of dimension $D - m$.

Principal Component Analysis (PCA)

Equivalent formulation: aim to find the hyperplane

$$c^T x = \beta, \quad c \in \mathbb{R}^D, \quad \beta \in \mathbb{R} \tag{1}$$

with ℓ_2 best approximation for $\mathcal{C}_D$. The distance between $x_i \in \mathcal{C}_D$ and the hyperplane (1) is $|c^T x_i - \beta|$ given $\|c\|_2 = 1$. Hence,

$$\min_{b, CC^T = I} \sum_{i=1}^{N} \|Cx_i - b\|_2^2, \tag{2}$$

where $C \in \mathbb{M}_{m,D}(\mathbb{R}), b \in \mathbb{R}^m$ yielding the subspace $Cx = b$ with m equations of dimension $D - m$.
Issue: The constraints $CC^T = I$ are not convex, in fact, they are combinatorial.

Principal Component Analysis (PCA)

Equivalent formulation: aim to find the hyperplane

$$c^T x = \beta, \quad c \in \mathbb{R}^D, \quad \beta \in \mathbb{R} \tag{1}$$

with ℓ_2 best approximation for $\mathcal{C}_D$. The distance between $x_i \in \mathcal{C}_D$ and the hyperplane (1) is $|c^T x_i - \beta|$ given $\|c\|_2 = 1$. Hence,

$$\min_{b, CC^T = I} \sum_{i=1}^{N} \|Cx_i - b\|_2^2, \tag{2}$$

where $C \in \mathbb{M}_{m,D}(\mathbb{R}), b \in \mathbb{R}^m$ yielding the subspace $Cx = b$ with m equations of dimension $D - m$.
Issue: The constraints $CC^T = I$ are not convex, in fact, they are combinatorial.
Solution: There is a closed form solution for problem (2).

Robust PCA

$$\overline{x} = \arg\min_x \sum_{i=1}^{N} (x_i - x)^2 = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (\ell_2)$$

$$\texttt{med}\ x = \arg\min_x \sum_{i=1}^{N} |x_i - x| \qquad (\ell_1)$$

The trade-off between ℓ_1 and ℓ_2 loss functions is known as Huber function, defined as

$$h(t) = \begin{cases} t^2/2, \text{ if } |t| \leq \delta, \\ \delta|t| - \delta^2/2, \text{ if } |t| > \delta. \end{cases}$$

Robust PCA

$$\overline{x} = \arg\min_x \sum_{i=1}^{N} (x_i - x)^2 = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (\ell_2)$$

$$\texttt{med}\ \ x = \arg\min_x \sum_{i=1}^{N} |x_i - x| \qquad (\ell_1)$$

The trade-off between ℓ_1 and ℓ_2 loss functions is known as Huber function, defined as

$$h(t) = \begin{cases} t^2/2, \text{ if } |t| \leq \delta, \\ \delta|t| - \delta^2/2, \text{ if } |t| > \delta. \end{cases}$$

The "robust" version of PCA reads as follows

$$\min_{b, CC^T = I} \sum_{i=1}^{N} h(\|Cx_i - b\|_2). \tag{3}$$

The trade-off between ℓ_1 and ℓ_2 loss functions is known as Huber function, defined as

$$h(t) = \begin{cases} t^2/2, \ \text{if } |t| \le \delta, \\ \delta|t| - \delta^2/2, \ \text{if } |t| > \delta. \end{cases}$$

The trade-off between ℓ_1 and ℓ_2 loss functions is known as Huber function, defined as

$$h(t) = \begin{cases} t^2/2, \text{ if } |t| \leq \delta, \\ \delta|t| - \delta^2/2, \text{ if } |t| > \delta. \end{cases}$$

The "robust" version of PCA reads as follows

$$\min_{b, CC^T = I} \sum_{i=1}^{N} h(\|Cx_i - b\|_2). \tag{4}$$

The trade-off between ℓ_1 and ℓ_2 loss functions is known as Huber function, defined as

$$h(t) = \begin{cases} t^2/2, & \text{if } |t| \leq \delta, \\ \delta|t| - \delta^2/2, & \text{if } |t| > \delta. \end{cases}$$

The "robust" version of PCA reads as follows

$$\min_{b, CC^T = I} \sum_{i=1}^{N} h(\|Cx_i - b\|_2). \tag{4}$$

Issue: The constraints $CC^T = I$ are not convex, in fact, they are combinatorial.

The trade-off between ℓ_1 and ℓ_2 loss functions is known as Huber function, defined as

$$h(t) = \begin{cases} t^2/2, & \text{if } |t| \leq \delta, \\ \delta|t| - \delta^2/2, & \text{if } |t| > \delta. \end{cases}$$

The "robust" version of PCA reads as follows

$$\min_{b, CC^T = I} \sum_{i=1}^{N} h(\|Cx_i - b\|_2). \tag{4}$$

Issue: The constraints $CC^T = I$ are not convex, in fact, they are combinatorial.
Solution: Manifold optimization.

Conjugate Gradient on Riemannian manifolds

- Find the equation of geodesics for the manifold, e.g. Grassmannian, Stiefel manifold. (can be quite involved)
- Perform 1D minimization along geodesics search directions (analog of line-search for flat spaces).
- Parallel transport the tangent vector along geodesics for transferring current information to the next iteration (the tangent spaces differ from point to point on Riemannian manifolds, unlike flat spaces)
- Update the new search direction. (exists a number of choices for step size, e.g. Polak-Ribière, Fletcher-Reeves, etc).

Stiefel manifold

Stiefel manifold: $\texttt{St}_{n,p} = \{X \in \mathbb{M}_{n,p} : X^T X = I_p\}$. Note that $\texttt{St}_{n,1} \equiv \mathcal{S}^{n-1}$ and $\texttt{St}_{n,n} \equiv \mathcal{O}_n$. We embed our manifold with the following (canonical) inner product

$$\langle A, B \rangle_S = \texttt{tr} A^T (I - \frac{1}{2} X X^T) B.$$

The geodesic equation for moving from $X(0) = X$ in the direction of $\dot{X}(0) = H$ on Stiefel manifold has the following form

$$X(t) = XM(t) + QN(t),$$

where $QR = K := (I - XX^T)H$ is the compact QR-decomposition of K, $A = X^T H$ and

$$\begin{pmatrix} M(t) \\ N(t) \end{pmatrix} = \exp\left\{ t \begin{pmatrix} A & -R^T \\ R & 0 \end{pmatrix} \right\} \begin{pmatrix} I_p \\ 0 \end{pmatrix}$$

The gradient of the function $F(X)$ on the Stiefel manifold is defined as

$$\nabla F(X) := F_X - X F_X^T X.$$

Conjugate Gradient on Stiefel manifold

Algorithm 1 Conjugate Gradient on the Stiefel manifold

Given: problem $\min_{b, CC^T = I} F(b, C)$ choose C_0 and some b_0 such that $C_0 C_0^T = I$.

Compute: $G_0 = \nabla F(\cdot, C_0)$ and set $H_0 = -G_0$.

for $k = 0, 1, \ldots$ **do**

  Minimize: $F(b^k, C_k(t))$ over t where

$$C_k(t) = C^k M(t) + QN(t) \qquad \text{[1D minimization along geodesics]},$$

  where $(I - (C^k)^T C^k) H_k = QR$ is a QR-decomposition.

  Update: $C^{k+1} = C_k(t_k)$ with $t_k = \arg\min_t F(b_k, C_k(t))$.

  Put: $w_{ik} = \min\left\{1, \frac{\delta}{\|C_k X_i - b_k\|_2}\right\}$.

  Update: $b^{k+1} = C^{k+1} \overline{X}_{\mathbf{w}}$, where $\overline{X}_{\mathbf{w}} := \frac{\sum_{i=1}^N X_i w_{ik}}{\sum_{i=1}^N w_{ik}}$.

  Compute: $G_{k+1} = \nabla F(b^{k+1}, C^{k+1})$

  Parallel transport:

$$\tau H_k = H_k M(t_k) - C_k R^T N(t_k),$$
$$\tau G_k = G_k \text{ (not parallel)}$$

  Put: $H_{k+1} = -G_{k+1} + \gamma_k \tau H_k$ [Update the new search direction]

$$\gamma_k = \frac{\langle G_{k+1} - \tau G_k, G_{k+1} \rangle_S}{\langle G_k, G_k \rangle_S}$$

end for

Eigenvalue Problem

$$\min_{X^T X = I} F(X) := \frac{1}{2}\texttt{tr} X^T A X \cdot N,$$

where $X \in \mathbb{M}_{k,n}(\mathbb{R})$ for fixed $k < n$ and $N \in \mathcal{S}_p(\mathbb{R})$.

Eigenvalue Problem

$$\min_{X^T X = I} F(X) := \frac{1}{2}\mathtt{tr} X^T A X \cdot N,$$

where $X \in \mathbb{M}_{k,n}(\mathbb{R})$ for fixed $k < n$ and $N \in \mathcal{S}_p(\mathbb{R})$.

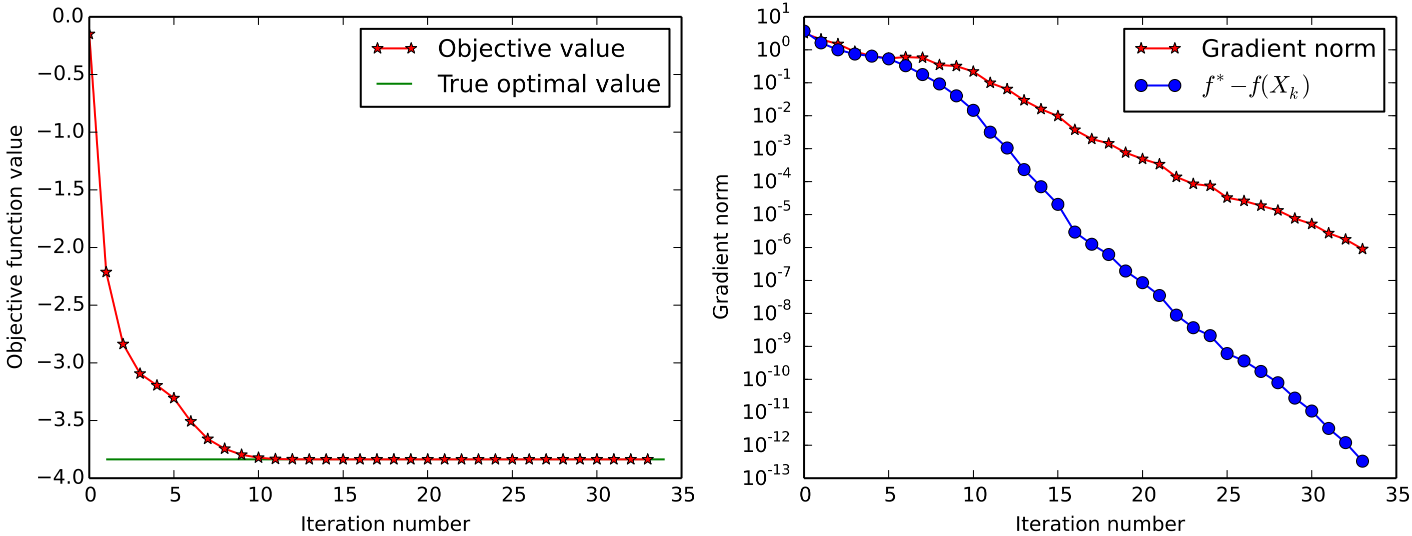


Figure 2: Matrix $A \in \mathcal{S}_n(\mathbb{R})$ with $n = 20$ and $p = 5$ and $N = I_p$.

Eigenvalue Problem

Take $A = \mathrm{diag}(1,2,3,4)$, $\quad N = \mathrm{diag}(1,2)$ on $\mathtt{St}_{4,2}(n=4, p=2)$.

$$\min_{\substack{(x_1,x_2)=0,\\ \|x_1\|=\|x_2\|=1}} \quad 0.5\left[x_1^T A x_1 + 2x_2^T A x_2\right].$$

Eigenvalue Problem

Take $A = \operatorname{diag}(1,2,3,4)$, $\quad N = \operatorname{diag}(1,2)$ on $\mathtt{St}_{4,2}(n=4, p=2)$.

$$\min_{\substack{(x_1,x_2)=0,\\ \|x_1\|=\|x_2\|=1}} \quad 0.5\left[x_1^T A x_1 + 2x_2^T A x_2\right].$$

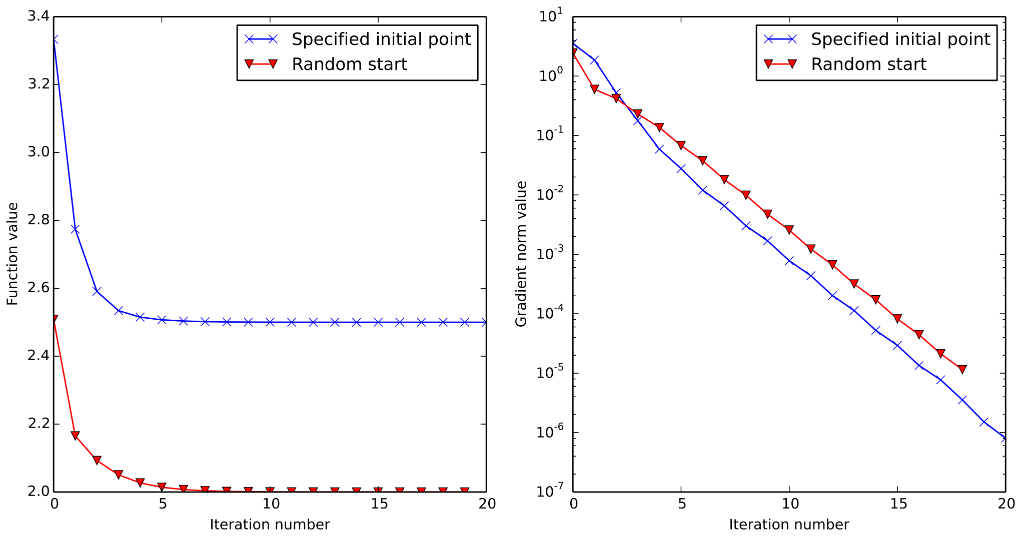


Figure 3: Example of stucking in local minimum point.

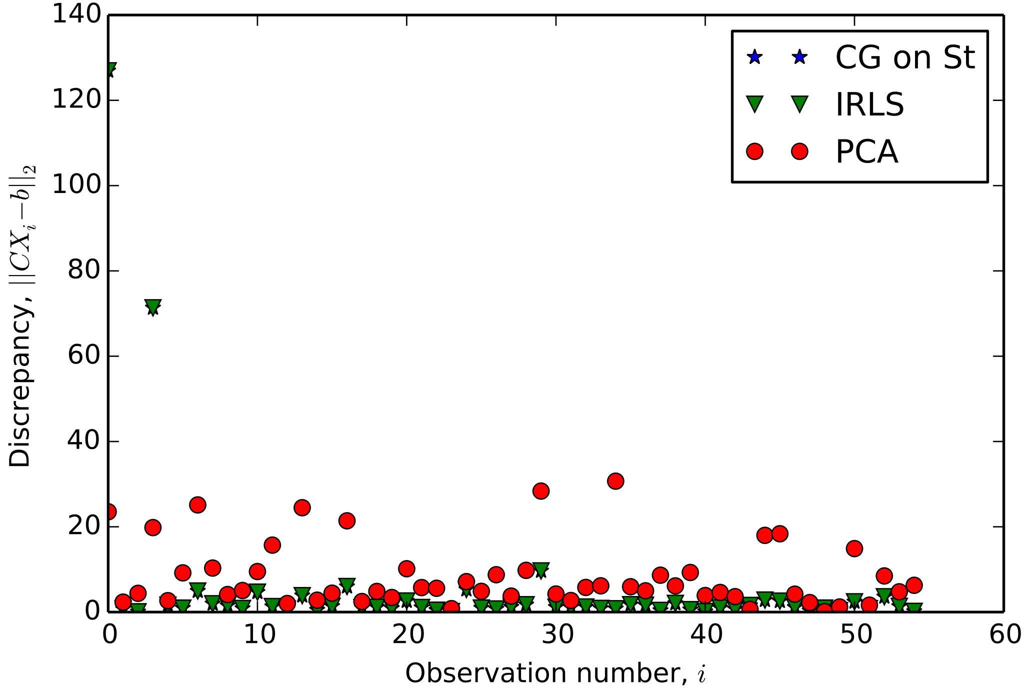
Robust PCA

R-PCA: Sleep in Mammals:[1] $\min_{b, CC^T=I} \sum_{i=1}^{N} h(\|Cx_i - b\|_2)$.

[1] http://www.statsci.org/data/general/sleep.html

Robust PCA

R-PCA: Sleep in Mammals:[1] $\min_{b, CC^T=I} \sum_{i=1}^{N} h(\|Cx_i - b\|_2)$.



[1] http://www.statsci.org/data/general/sleep.html

R-PCA: Wine Quality[2]

	White Wine		Red Wine	
Method	Time [sec]	# Iterations	Time [sec]	# Iterations
IRLS	16.25	16	12.12	14
CG on St	9.12	156	7.64	119

	White Wine		Red Wine	
Method	Value converged	Relative error	Value converged	Relative error
IRLS	3704.9952	$5.2 \cdot 10^{-9}$	1338.0631	$1.1 \cdot 10^{-9}$
CG on St	3704.9981	$2.5 \cdot 10^{-7}$	1338.5662	$7.5 \cdot 10^{-7}$

[2] http://www3.dsi.uminho.pt/pcortez/wine/

Bibliography

- B. T. Polyak and M.V. Khlebnikov. Principal Component Analysis: Robust versions, *Automation and Remote Control* **78** (3): 490–506, 2017.
- E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56**(5), 2053-2080.
- A. Edelman, T. A. Arias, S.T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM J. on M. Analysis and App.*. 1998, 20(2), 303–353.
- Peter J. Huber, Elvezio M. Ronchetti, Robust Statistics, 2nd Edition, pages 1-380 pages, March 2009.
- G.E.P. Box. Non-Normality and Tests on Variances. *Biometrika*. 1953. V. 40. P. 318–335.