Robust PCA via Regularized REAPER and Matrix-Free Proximal Algorithms

Robert Beinert
Technische Universität Berlin

Classical PCA

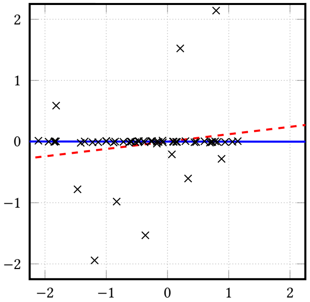
Principle Component Analysis (PCA)

- Given N data points $x_1, \dots, x_N \in \mathbb{R}^n$, find the d-dimensional affine subspace $\{At + b : t \in \mathbb{R}^d\}$ minimizing

$$\sum_{k=1}^{N} \min_{t \in \mathbb{R}^d} \|A\,t + b - x_k\|_2^2 \quad \text{subject to} \quad A^\top A = I_d$$

  over $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n,d}$.
- The optimal offset b is the mean of the data.
- The principle components A may be computed using the covariance matrix of the data.

Example (Classical PCA)



Equivalent Formulation

$$\underset{A \in \mathbb{R}^{n,d}, b \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{k=1}^{N} \|(AA^\top - I_n)(b - x_k)\|_2^2 \quad \text{subject to} \quad A^\top A = I_d.$$

Robust PCA

How to make the classical PCA more robust?

$$\underset{A\in\mathbb{R}^{n,d},b\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{k=1}^{N} \|(AA^\top - I_n)(b - x_k)\|_2^2 \quad \text{subject to} \quad A^\top A = I_d.$$

1. Consider random subsets of points. — Fischler & Bolles 1987
2. Use the 1-norm instead. — Ke & Kanade 2005
3. Introduce weights. — Kriegel et al. 2008
4. Exclude outliers. — Podosinnikova et al. 2014, Rousseeuw & Leroy 2005, Tsakiris & Vidal 2018
5. Decompose data in low-rank and column-sparse part. — Candes et al. 2011, Cherapanamijeri et al. 2017, McCoy & Tropp 2011, Xu et al. 2012
6. Skip the square. — Ding et al. 2006, Lerman & Maunu 2018, Maunu et al. 2019, Neumayer et al. 2020

- We consider the last approach; so skip the square.
- For simplicity, assume that the data are centred, i.e. $b = 0$.

- **Aim:** We want to find a solution of

$$\underset{A \in \mathbb{R}^{n,d}}{\text{minimize}} \quad \sum_{k=1}^{N} \|(AA^\top - I_n)x_k\|_2 \quad \text{subject to} \quad A^\top A = I_d.$$

- Notice that $\boldsymbol{\Pi} := AA^\top$ is the orthogonal projector.
- **Reformulation:** With $X := [x_1|\dots|x_N]$, the problem is equivalent to

$$\underset{\Pi \in \mathcal{S}(n)}{\text{minimize}} \quad \|\boldsymbol{\Pi}X - X\|_{2,1} \quad \text{subject to} \quad \boldsymbol{\lambda}_{\boldsymbol{\Pi}} \in \{0,1\}^n, \ \operatorname{tr}(\boldsymbol{\Pi}) = d.$$

- **Convexification:** Replacing the constraints by its convex hull, we obtain *REAPER*:

$$\underset{P \in \mathcal{S}(n)}{\text{minimize}} \quad \|PX - X\|_{2,1} \quad \text{subject to} \quad \mathbf{0}_{n,n} \preceq P \preceq I_n, \ \operatorname{tr}(P) = d.$$

- To deal with the non-differentiability, Lerman et al. iteratively solve a series of positive semi-definite programs requiring a memory of size $O(nN + n^2)$.

- **Final Step:** Project P onto the orthoprojectors

$$\mathcal{O}_d := \{\boldsymbol{\Pi} \in \mathcal{S}(n) : \boldsymbol{\lambda}_{\boldsymbol{\Pi}} \in \mathcal{E}_d\} \quad \text{with} \quad \mathcal{E}_d := \{\lambda \in \mathbb{R}^n : \boldsymbol{\lambda} \in \{0,1\}^n, \langle \boldsymbol{\lambda}, \mathbf{1}_n \rangle \leq d\}.$$

Regularized REAPER

1. Relax the strict trace constraint.
2. Add the nuclear norm as a regularizer.

- We obtain the reformulation

$$\underset{\boldsymbol{\Pi} \in \mathcal{S}(n)}{\text{minimize}} \quad \|\boldsymbol{\Pi} X - X\|_{2,1} + \alpha \|\boldsymbol{\Pi}\|_{\text{tr}} \quad \text{subject to} \quad \boldsymbol{\lambda}_{\boldsymbol{\Pi}} \in \{0,1\}^n, \ \text{tr}(\boldsymbol{\Pi}) \leq d$$

and its convexification *rREAPER*:

$$\underset{P \in \mathcal{S}(n)}{\text{minimize}} \quad \|PX - X\|_{2,1} + \alpha \|P\|_{\text{tr}} \quad \text{subject to} \quad \mathbf{0}_{n,n} \preceq P \preceq I_n, \ \text{tr}(P) \leq d.$$

Structure of RREAPER

$$\underset{P\in\mathcal{S}(n)}{\text{minimize}} \quad \|PX - X\|_{2,1} + \alpha\|P\|_{\mathrm{tr}} \quad \text{subject to} \quad 0_{n,n} \preceq P \preceq I_n,\ \mathrm{tr}(P) \leq d.$$

PREAPER as Inverse Problem

- The forward operator is given by

$$\mathcal{X} : \mathcal{S}(n) \to \mathbb{R}^{n,N} : P \mapsto PX.$$

- Its adjoint is

$$\mathcal{X}^*(Y) = \tfrac{1}{2}(XY^\top + YX^\top).$$

- With the regularizer $\mathcal{R} : \mathcal{S}(n) \to [0, +\infty]$ defined by

$$\mathcal{R}(P) := \|P\|_{\mathrm{tr}} + \iota_{\mathcal{C}}(P), \qquad \mathcal{C} := \{P \in \mathcal{S}(n) : 0_{n,n} \preceq P \preceq I_n, \mathrm{tr}(P) \leq d\},$$

  rREAPER may be rearrange as

$$\underset{P\in\mathcal{S}(n)}{\text{minimize}} \quad \|\mathcal{X}(P) - X\|_{2,1} + \alpha\mathcal{R}(P).$$

- RREAPER has a global minimizer.

Algorithm (Primal-Dual Method)

INPUT: $X \in \mathbb{R}^{n,N}$ $d \in \mathbb{N}$, *and* $\alpha, \sigma, \tau > 0$ *with* $\sigma\tau < 1/\|X\|_2^2$, *and* $\theta \in (0, 1]$.

INTIALIZATION: $P^{(0)} = \bar{P}^{(0)} = 0_{n,n}$, $Y^{(0)} := 0_{n,N}$.

INTERATION:

$$Y^{(r+1)} := \operatorname{prox}_{\sigma\| \cdot - X\|_{2,1}^*}\Big(Y^{(r)} + \sigma\mathcal{X}\big(\bar{P}^{(r)}\big)\Big),$$

$$P^{(r+1)} := \operatorname{prox}_{\tau\alpha\mathcal{R}}\Big(P^{(r)} - \tau\mathcal{X}^*\big(Y^{(r+1)}\big)\Big),$$

$$\bar{P}^{(r+1)} := (1 + \theta)\, P^{(r+1)} - \theta\, P^{(r)}$$

Proposition (Proximation of the Dual Data Fidelity)

For $X \in \mathbb{R}^{n,N}$ *and* $\sigma > 0$, *we have*

$$\operatorname{prox}_{\sigma\| \cdot - X\|_{2,1}^*} = \operatorname{proj}_{B_{2,\infty}}(\cdot - \sigma X).$$

Proximation of the Regularizer

- Define the halfspace

$$\mathcal{H} := \mathcal{H}(1_n, d) = \{x \in \mathbb{R}^n : \langle x, 1_n \rangle \leq d\}.$$

Proposition (Proximation of the Regularizer)

For $\boldsymbol{P} \in \mathcal{S}(n)$ with spectral decomposition $\boldsymbol{P} = \boldsymbol{U} \operatorname{diag}(\boldsymbol{\lambda}_{\boldsymbol{P}}) \boldsymbol{U}^\top$, it holds

$$\operatorname{prox}_{\tau\alpha\mathcal{R}}(\boldsymbol{P}) = \boldsymbol{U} \operatorname{diag}(\operatorname{proj}_{\mathcal{Q}\cap\mathcal{H}}(\boldsymbol{\lambda}_{\boldsymbol{P}} - \tau\alpha 1_n)) \boldsymbol{U}^\top.$$

Proposition (Projection onto the Truncated Hypercube, Beck 2017)

For any $\boldsymbol{\lambda} \in \mathbb{R}^n$ and any $d \in (0, n]$, the projection to the truncated hypercube is given by

$$\operatorname{proj}_{\mathcal{Q}\cap\mathcal{H}}(\boldsymbol{\lambda}) = \begin{cases} \operatorname{proj}_{\mathcal{Q}}(\boldsymbol{\lambda}) & \text{if } \langle \operatorname{proj}_{\mathcal{Q}}(\boldsymbol{\lambda}), 1_n \rangle \leq d, \\ \operatorname{proj}_{\mathcal{Q}}(\boldsymbol{\lambda} - \hat{t} 1_n) & \text{otherwise,} \end{cases}$$

where $\hat{t}$ is the smallest root of the function

$$\phi(t) := \langle \operatorname{proj}_{\mathcal{Q}}(\boldsymbol{\lambda} - t 1_n), 1_n \rangle - d, \qquad t \geq 0.$$

Lemma (Properties of ϕ)

For $\boldsymbol{\lambda} \in \mathbb{R}^n$ with $\langle \mathrm{proj}_{\mathcal{Q}}(\boldsymbol{\lambda}), \mathbf{1}_n \rangle > d$, the function $\phi : [0, \infty) \to \mathbb{R}$ satisfies:

(i) *ϕ is Lipschitz continuous, monotone decreasing, and piecewise linear.*

(ii) *We can construct a sequence $0 = s_0 < s_1 < s_2 < \ldots < s_M$ with $M \leq 2n$ such that*

$$\phi(t) = \phi(s_l) - k_l(t - s_l),$$

for $t \in [s_l, s_{l+1})$, $l = 0, \ldots, M-1$, where

$$k_l := |\{j \in \{1, \ldots, n\} : (\boldsymbol{\lambda} - s_l \mathbf{1}_n)_j \in (0, 1]\}|.$$

Moreover, $\phi(t) = -d$ for $t \geq s_M$.

(iii) *The smallest root $\hat{t}$ of ϕ is given by*

$$\hat{t} = s_m + \frac{1}{k_m}\phi(s_m),$$

where m is the unique index such that $\phi(s_m) > 0$ and $\phi(s_{m+1}) \leq 0$.

Algorithm (Primal-Dual RREAPER)

(i) *DUAL UPDATE:* $Y^{(r+1)} := \operatorname{proj}_{B_{2,\infty}}(Y^{(r)} + \sigma(\mathcal{X}(\bar{P}^{(r)}) - X))$.

(ii) *PRIMAL UPDATE:*

 (a) $U \operatorname{diag}(\boldsymbol{\lambda})\, U^{\mathsf{T}} := P^{(r)} - \tau \mathcal{X}^*(Y^{(r+1)})$,

 (b) $\hat{\boldsymbol{\lambda}} := \operatorname{proj}_{\mathcal{Q}\cap\mathcal{H}}(\boldsymbol{\lambda} - \tau\alpha \mathbf{1}_n)$,

 (c) $P^{(r+1)} := U \operatorname{diag}(\hat{\boldsymbol{\lambda}})\, U^*$.

(iii) *EXTRAPOLATION:* $\bar{P}^{(r+1)} := (1+\theta)\, P^{(r+1)} - \theta\, P^{(r)}$.

- **Matrix-free representation:** $P^{(r)} := \sum_{k=1}^{d_r} \lambda_k^{(r)} u_k^{(r)} (u_k^{(r)})^{\mathsf{T}}$.
- **Thick-restarted LANCZOS process, Wu & Simon 2000:** Find an orthogonal matrix E such that

$$E^* P E = T = \begin{bmatrix} \breve{\lambda}_1 & & & \rho_1 & & & \\ & \ddots & & \vdots & & & \\ & & \breve{\lambda}_\ell & \rho_\ell & & & \\ \rho_1 & \cdots & \rho_\ell & \beta_{\ell+1} & \gamma_{\ell+1} & & \\ & & & \gamma_{\ell+1} & \beta_{\ell+1} & \ddots & \\ & & & & \ddots & \ddots & \gamma_{k_{\max}-1} \\ & & & & & \gamma_{k_{\max}-1} & \beta_{k_{\max}} \end{bmatrix}.$$

Compute $T = Y \Lambda Y^{\mathsf{T}}$, set $U = EY$, and repeat.

Convergence and Projection

Theorem (Convergence of Primal-Dual RREAPER)

Let

$$\mathbf{Q}^{(r)} := \mathbf{P}^{(r-1)} - \tau\mathcal{X}^*(\mathbf{Y}^{(r)}) = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\lambda})\,\mathbf{U}^\top$$

be the argument in the rth primal-dual iteration, and $\tilde{\mathbf{Q}}^{(r)}$ *the computed low-rank approximation of the leading positive eigenvalues greater than* $\tau\alpha$*. Define*

$$E_r := \|(\tilde{\mathbf{Q}}^{(r)} - \tau\alpha\mathbf{I}_n)_{\geq 0} - (\mathbf{Q}^{(r)} - \tau\alpha\mathbf{I}_n)_{\geq 0}\|_F.$$

Then, for $\sigma\tau < 1/\|\mathbf{X}\|_2^2$ *and* $\theta = 1$*, primal-dual RREAPER converges to a global minimizer whenever* $\sum_{r=0}^{\infty} E_r^{1/2}(\|\mathbf{Q}^{(r)}\|_F + \sqrt{d})^{1/2} < \infty$.

Proposition (Projection onto the Orthoprojectors)

For $\mathbf{P} \in S(n)$ *with eigenvalue decomposition* $\mathbf{P} = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\lambda}_{\mathbf{P}})\,\mathbf{U}^\top$*, and for every* $1 \leq p \leq \infty$*, the projection onto* $\mathbb{O}_d$ *with respect to the Schatten p-norm is given by*

$$\mathrm{proj}_{\mathbb{O}_d}(\mathbf{P}) = \mathbf{U}\,\mathrm{diag}(\mathrm{proj}_{\mathcal{E}_d}(\boldsymbol{\lambda}))\,\mathbf{U}^\top.$$

- Define the *distance of the data from the subspace* and the *permeance statistic* by

$$\mathcal{R}_L = \mathcal{R}_L(X) := \|(I_n - \boldsymbol{\Pi}_L)X\|_{2,1}, \qquad \text{and} \qquad \mathcal{P}_L = \mathcal{P}_L(X) := \min_{\substack{\boldsymbol{u}\in L\\ \|\boldsymbol{u}\|=1}} \sum_{k=1}^{N} |\langle u, x_k\rangle|.$$

Theorem (Recovery Error)

Let $\boldsymbol{\Pi}_L$ be the orthogonal projector onto a subspace L of $\mathbb{R}^n$ of dimension d_L and $x_k \in \mathbb{R}^n$, $k = 1, \ldots, N$, $N \geq d_L$ such that their projections onto L form a frame of L. Set $\gamma_L := 1/2\sqrt{2d_L}\,\mathcal{P}_L$. Let $\hat{P}$ be the solution of RREAPER *and $\hat{\boldsymbol{\Pi}}$ the projection of $\hat{P}$ onto $\mathcal{O}_d$. Then, for $\alpha \leq 2\gamma_L$ and $d \geq d_L$, it holds*

$$\|\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi}_L\|_{\mathrm{tr}} \leq \min\left\{\frac{6\mathcal{R}_L}{\gamma_L - |\gamma_L - \alpha|}, \frac{6\mathcal{R}_L + 2d|\alpha - \gamma_L|}{\gamma_L}\right\}.$$

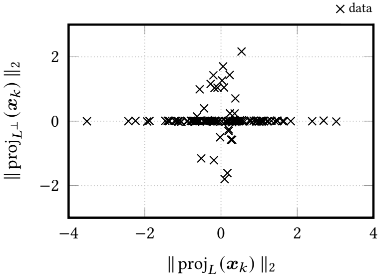
- Dividing the data in in- and outlier, we obtain

Lerman et al. 2015

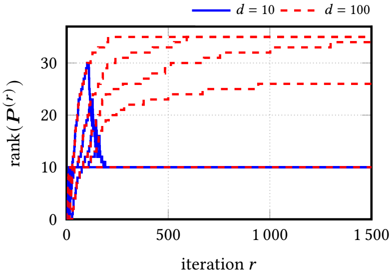
$$\|\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi}_L\|_{\mathrm{tr}} \leq \left(16\sqrt{d_L}\mathcal{R}_L(X_{\mathrm{in}}) + 8\alpha d_L\sqrt{d_L}\right) \Big/ \left(\mathcal{P}_L(X_{\mathrm{in}}) - 4\sqrt{d_L}\mathcal{R}_L(X_{\mathrm{in}}) - 4\sqrt{d_L}\|X_{\mathrm{out}}\|\,\|((I_n - \boldsymbol{\Pi}_L)X_{\mathrm{out}})^{\sim}\|\right)_+.$$

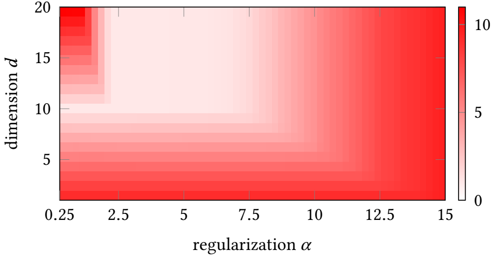
Parameter Choise

Example (Rank Evolution & Recovery Error)

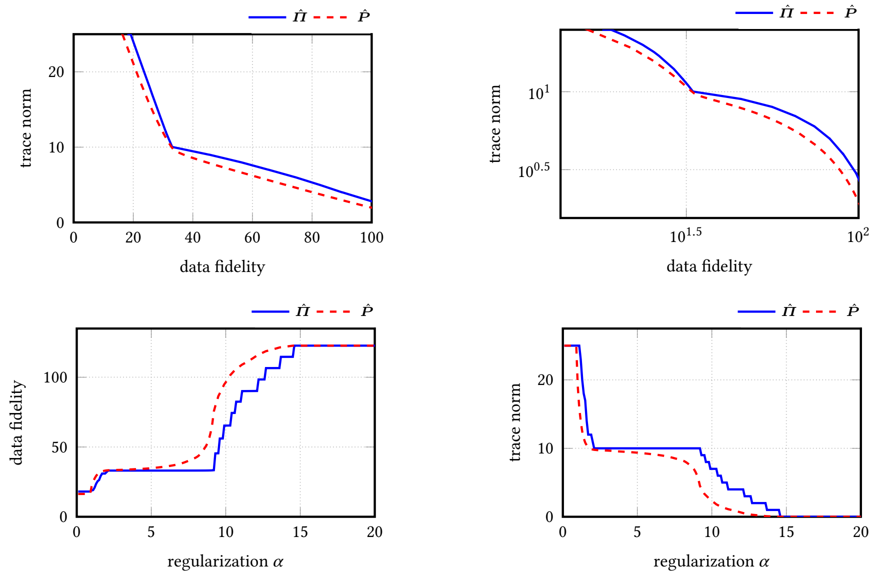


- 10-dimensional subspace in $\mathbb{R}^{100}$
- 100 inliner, 25 outlier
- GAUßian noise
- Mean recovery error $\|\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi}\|_{\mathrm{tr}}$





Example (L-Curve Method, d=25)



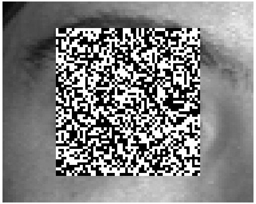
Eigenfaces

- Extended Yale Face Data Set B. Georghiades et al. 2001, Lee et al. 2005
- Series of 64 face images of size 640×480 under various illuminations.
- Four images have artefacts to serve as outliers.
- $P^{(r)} \in \mathbb{R}^{307\,200 \times 307\,200}$ would require 703.125 GiB.

Example (Artefacts and noise)

Example (Comparison Between Robust PCA Methods)

Original sample



Artefact

RREAPER

PCA-L1[1]

Nested Weiszfeld[2]

R1-PCA[3]

FMS[4]

GGD[5]

[1] Kwak 2008, [2] Neumayer et al., [3] Ding et al. 2006, Neumayer et al. 2020, [4] Lerman & Maunu 2018, [5] Maunu et al. 2019.

Summary/Outlook

- Convex variational method for robust PCA that can handle high-dimensional data.
- Reduce the required memory from $O(nN + n^2)$ to $O(nN + nr)$.
- Guaranteed convergence to the global minimizer even for inexact eigenvalue decompositions.
- Determine the subspace dimension via the L-curve method.
- The regularizer smooth the principle components.

- How to choose the offset $\boldsymbol{b}$.
- Employ other sparsity promoting norms to incorporate different noise types.

Thank you for the attention.

- Robert BEINERT, Gabriele STEIDL. Robust PCA via Regularized REAPER with a Matrix-Free Proximal Algorithm. *Journal of Mathematical Imaging and Vision*, 21, 1181–1232, 2021.