

Robust Principal Component Analysis using Density Power Divergence

Subhrajyoty Roy

Research Fellow, Interdisciplinary Statistical Research Unit
Indian Statistical Institute, Kolkata, India

Supervisors: Ayanendranath Basu & Abhik Ghosh

Dec, 2024,
IISA Conference

Principal Component Analysis (PCA)

Given an independent and identically distributed sample X_1, X_2, \dots, X_n , each $X_i \in \mathbb{R}^p$, and a scale measure $S_n(y_1, \dots, y_n)$, first principal component is

$$\hat{\mathbf{v}}_1 = \arg \max_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} S_n(\mathbf{v}^T X_1, \mathbf{v}^T X_2, \dots, \mathbf{v}^T X_n),$$

and the eigenvalue is

$$\hat{\lambda}_1 = S_n(\hat{\mathbf{v}}_1^T X_1, \hat{\mathbf{v}}_1^T X_2, \dots, \hat{\mathbf{v}}_1^T X_n).$$

Subsequent principal components are defined similarly subject to orthogonality conditions.

Effect of outlier on PCA

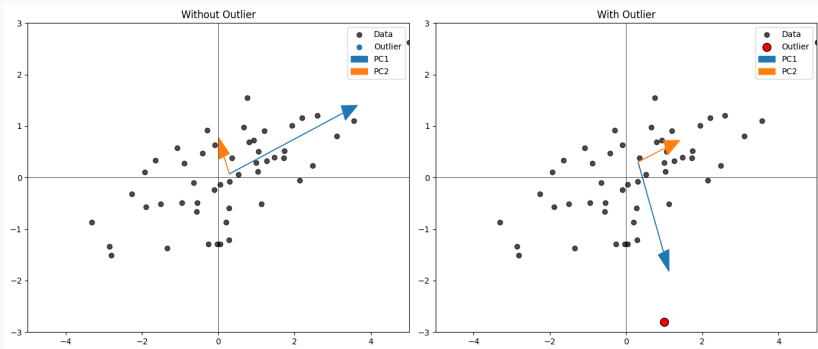


Figure 1: Data generated from multivariate normal. A single outlier can greatly distort the principal component eigenvalues and eigenvectors.

Alternative formulation of PCA

Consider,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \\ \vdots \\ \mathbf{X}_n^\top \end{bmatrix}_{n \times p}$$

and if S_n is the standard deviation, then

$$\frac{1}{n} \mathbf{X}^\top \mathbf{X} = \sum_{k=1}^r \lambda_k \mathbf{v}_k \mathbf{v}_k^\top + \mathbf{E}_1$$

Alternative formulation of PCA

Consider,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \\ \vdots \\ \mathbf{X}_n^\top \end{bmatrix}_{n \times p}$$

and if S_n is the standard deviation, then

$$\frac{1}{n} \mathbf{X}^\top \mathbf{X} = \sum_{k=1}^r \lambda_k \mathbf{v}_k \mathbf{v}_k^\top + \mathbf{E}_1$$

If we consider the same SVD decomposition

$$\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^\top + \mathbf{E} = \mathbf{L} + \mathbf{E},$$

with $\mathbf{D} = \text{Diag}(\sqrt{n\lambda_1}, \dots, \sqrt{n\lambda_p})$, and \mathbf{L} is a low-rank matrix.

$$\mathbf{X} = \mathbf{L} + \mathbf{E}, \text{ where, } \mathbf{E} = ((\epsilon_{ij}))_{i,j}; \epsilon_{ij} \sim (1 - \delta)g + \delta h$$

PCA as alternating regression

$$\mathbf{X} = \mathbf{L} + \mathbf{E}, \text{ where, } \mathbf{E} = ((\epsilon_{ij}))_{i,j}; \epsilon_{ij} \sim (1 - \delta)g + \delta h$$

$$\Rightarrow X_{ij} = \sum_{k=1}^r \sqrt{n\lambda_k} u_{ki} v_{kj} + \epsilon_{ij}, \lambda_k \text{ is eigenvalue, } u_{ki}, v_{kj} \text{ are singular vectors}$$

PCA as alternating regression

$$\mathbf{X} = \mathbf{L} + \mathbf{E}, \text{ where, } \mathbf{E} = ((\epsilon_{ij}))_{i,j}; \epsilon_{ij} \sim (1 - \delta)g + \delta h$$

$$\Rightarrow X_{ij} = \sum_{k=1}^r \sqrt{n\lambda_k} u_{ki} v_{kj} + \epsilon_{ij}, \lambda_k \text{ is eigenvalue, } u_{ki}, v_{kj} \text{ are singular vectors}$$

$$\Rightarrow X_{ij} = \sum_{k=1}^r a_{ki} b_{kj} + \epsilon_{ij}; \text{ assume, } a_{ki} = (n\lambda_k)^{1/4} u_{ki}, b_{kj} = (n\lambda_k)^{1/4} v_{kj}$$

PCA as alternating regression

$$\mathbf{X} = \mathbf{L} + \mathbf{E}, \text{ where, } \mathbf{E} = ((\epsilon_{ij}))_{i,j}; \epsilon_{ij} \sim (1 - \delta)g + \delta h$$

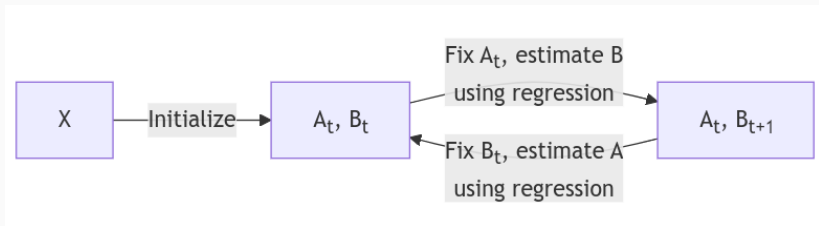
$$\Rightarrow X_{ij} = \sum_{k=1}^r \sqrt{n\lambda_k} u_{ki} v_{kj} + \epsilon_{ij}, \lambda_k \text{ is eigenvalue, } u_{ki}, v_{kj} \text{ are singular vectors}$$

$$\Rightarrow X_{ij} = \sum_{k=1}^r a_{ki} b_{kj} + \epsilon_{ij}; \text{ assume, } a_{ki} = (n\lambda_k)^{1/4} u_{ki}, b_{kj} = (n\lambda_k)^{1/4} v_{kj}$$

Fixing j yields,

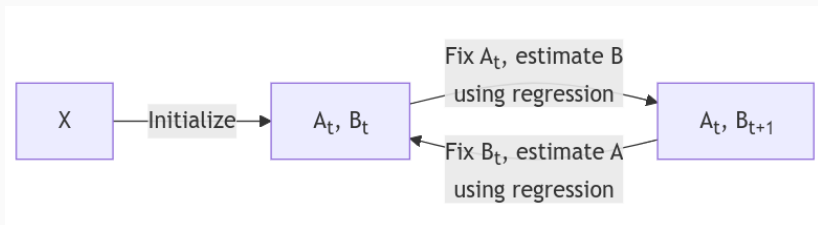
$$\begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{r1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \dots & a_{rn} \end{bmatrix}_{n \times r} \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{rj} \end{bmatrix} + \begin{bmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \vdots \\ \epsilon_{nj} \end{bmatrix}$$

PCA as alternating regression



1. **Scalable:** Avoids inversion of large matrices ($r \ll n, p$).
2. **Fast and Parallelizable:** As each regression problem can be solved independently.

PCA as alternating regression



1. **Scalable:** Avoids inversion of large matrices ($r \ll n, p$).
2. **Fast and Parallelizable:** As each regression problem can be solved independently.

Note: PCA is sensitive to outliers, since solving regression using **ordinary least squares** is sensitive to outliers.

How to perform robust linear regression?

Density Power Divergence (DPD): Given two density functions f and g , Basu et al. (1998) defines the DPD between them as

$$d_{\alpha}(g, f) = \begin{cases} \int f^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \int f^{\alpha} g + \frac{1}{\alpha} \int g^{1+\alpha}, & \alpha > 0 \\ \int f \log(f/g) & \alpha = 0 \end{cases}$$

How to perform robust linear regression?

Density Power Divergence (DPD): Given two density functions f and g , Basu et al. (1998) defines the DPD between them as

$$d_{\alpha}(g, f) = \begin{cases} \int f^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \int f^{\alpha} g + \frac{1}{\alpha} \int g^{1+\alpha}, & \alpha > 0 \\ \int f \log(f/g) & \alpha = 0 \end{cases}$$

Given an **independent but non-identically distributed** sample of observations X_1, \dots, X_n modelled by non-homogeneous families of densities $\{f_{i,\theta} : \theta \in \Theta\}$, the **MDPDE** (Ghosh and Basu) is defined as

$$\hat{\theta}_{\alpha} = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \left[\int f_{i,\theta}^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) f_{i,\theta}^{\alpha}(X_i) \right], \quad \alpha > 0.$$

Robust PCA using DPD

Assume estimation of the first principal component only,

$$X_{ij} = a_i b_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim f(\cdot/\sigma), \sigma \in (0, \infty)$$

Assume form of f is known and it is symmetric around 0.

Robust PCA using DPD

Assume estimation of the first principal component only,

$$X_{ij} = a_i b_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim f(\cdot/\sigma), \sigma \in (0, \infty)$$

Assume form of f is known and it is symmetric around 0.

Define weight function $\psi(\cdot)$

$$w_{ij}^{(t)} = \psi \left(|X_{ij} - a_i^{(t)} b_j^{(t)}| / \sigma^{(t)} \right), \quad \psi(x) = -f^{\alpha-1}(|x|) f'(|x|) / |x|$$

Robust PCA using DPD

Assume estimation of the first principal component only,

$$X_{ij} = a_i b_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim f(\cdot/\sigma), \sigma \in (0, \infty)$$

Assume form of f is known and it is symmetric around 0.

Define weight function $\psi(\cdot)$

$$w_{ij}^{(t)} = \psi\left(|X_{ij} - a_i^{(t)} b_j^{(t)}|/\sigma^{(t)}\right), \quad \psi(x) = -f^{\alpha-1}(|x|)f'(|x|)/|x|$$

Then, estimates are simply weighted averages as $x_{ij}/b_j \approx a_i$

$$a_i^{(t+1/2)} = \left[\sum_j (b_j^{(t)})^2 w_{ij}^{(t)} \right]^{-1} \left[\sum_j (x_{ij}/b_j^{(t)}) (b_j^{(t)})^2 w_{ij}^{(t)} \right], \quad i = 1, \dots, n;$$

$$a_i^{(t+1)} = a_i^{(t+1/2)} / \|\mathbf{a}^{(t+1/2)}\|, \quad \text{as we want unitary matrices}$$

Different choices of model family

All the weight functions are decreasing!

So more errors \implies less weights.

Density family	$\propto f(x)$	$\psi(x)$
Normal	$e^{-x^2/2}$	$e^{-\alpha x^2/2}$
Laplace	$e^{- x }$	$e^{-\alpha x }/ x $
t_ν	$(1 + x^2/\nu)^{-(1+\nu)/2}$	$(1 + 1/\nu)(1 + x^2/\nu)^{-\alpha(\nu+1)/2-1}$
Logistic	$e^{-x}/(1 + e^{-x})^2$	$e^{-\alpha x}(1 - e^{-x})/x(1 + e^{-x})^{2\alpha+1}$

Table 1: The choices of $\psi(\cdot)$ functions for different elliptically symmetric family of densities.

Results on rPCAdpd (Roy, Basu and Ghosh, 2024)

We assume X_1, \dots, X_n are from an elliptically symmetric density family,

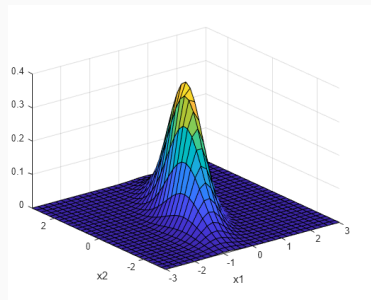
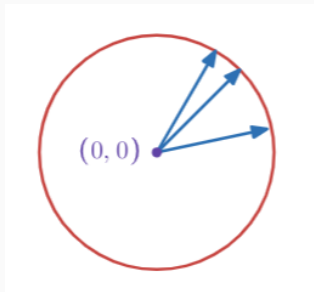
$$f_{\theta}(\mathbf{x}) \propto \det(\mathbf{\Sigma})^{-1/2} \exp \left[g \left((\mathbf{x} - \boldsymbol{\mu})^{\top} \sum_{k=1}^p \gamma_k^{-1} \mathbf{v}_k \mathbf{v}_k^{\top} (\mathbf{x} - \boldsymbol{\mu}) \right) \right]$$

If g is decreasing and twice differentiable, the minimizer of DPD exists and the rPCAdpd algorithm converges.

If the mean $\hat{\boldsymbol{\mu}}$ is orthogonally equivariant, then the rPCAdpd estimates are orthogonally equivariant.

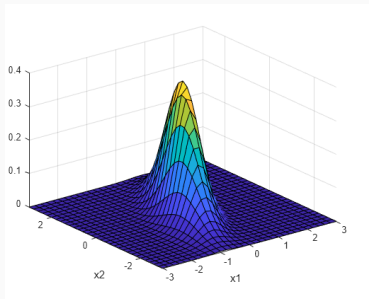
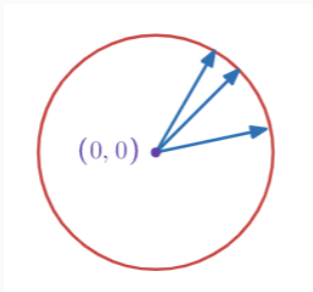
Results on rPCAdpd (Roy, Basu and Ghosh, 2023)

Typical consistency and asymptotic normality results with one caveat.



Results on rPCAdpd (Roy, Basu and Ghosh, 2023)

Typical consistency and asymptotic normality results with one caveat.



If the true eigenvalues are distinct, then for any n^c -consistent mean estimator $\hat{\mu}$ ($c \geq 1/2$), the rPCAdpd estimates of eigenvalues and **a natural parameter η for the eigenvectors** are \sqrt{n} -consistent and are jointly asymptotically normal.

If X_1, \dots, X_n are modelled using p -variate Gaussian family. Then for any n^c -consistent mean estimator $\hat{\mu}$ ($c \geq 1/2$),

1. The estimated eigenvalues $\{\hat{\gamma}_k\}_{k=1}^p$ and the estimated eigenvectors $\{\hat{v}_k\}_{k=1}^p$ are asymptotically independent.
2. $\sqrt{n}(\hat{\gamma} - \gamma)$ is asymptotically p -variate normally distributed with mean $\mathbf{0}$ and variance

$$\frac{(1 + \alpha)^{p+4}}{(1 + 2\alpha)^{p/2}} \mathbf{M}^{-1} \left(A_\alpha \mathbf{J}_\gamma + \frac{1}{2(1 + 2\alpha)^2} \text{Diag}(\gamma)^{-2} \right) \mathbf{M}^{-1}$$

where

$$\mathbf{J}_\gamma = (\text{Diag}(\gamma)^{-1})(\text{Diag}(\gamma)^{-1})^\top$$

$$\mathbf{M} = \left(\frac{\alpha^2}{4} \mathbf{J}_\gamma + \frac{1}{2} \text{Diag}(\gamma)^{-2} \right)$$

$$A_\alpha = \alpha^2 \left[\frac{1}{(1 + 2\alpha)^2} - \frac{(1 + 2\alpha)^{p/2}}{4(1 + \alpha)^{p+2}} \right]$$

If X_1, \dots, X_n are modelled using p -variate Gaussian family. Then for any n^c -consistent mean estimator $\hat{\mu}$ ($c \geq 1/2$),

3. For the natural parameter $\sqrt{n}(\hat{\eta} - \eta)$ is asymptotically p -variate normally distributed with 0 and a variance

$$\frac{(1 + \alpha)^{p/4}}{(1 + 2\alpha)^{2+p/2}} \sum_{k,l} \left(1 - \frac{\gamma_k}{\gamma_l}\right) \frac{\partial v_k}{\partial \eta} v_k v_l^T \left(\frac{\partial v_l}{\partial \eta}\right)^T$$

- The estimator has a bounded influence function, if $\hat{\mu}$ has bounded influence.
- The rPCAdpd estimator has asymptotic breakdown point at least $\alpha/(1 + \alpha)$ which is free of p .
- Extensive simulation studies have been performed and compared against existing methods (PCP, spherical PCA, ROBPCA, projection pursuit- based methods, Geometric median-based PCA, etc.).
 - Multiple levels of contaminations from 5% to 20%.
 - Cauchy and t_5 distribution of errors.
 - Dimensions ranging from $p = 10$ to $p = 250$, with $n = 50$.
- rPCAdpd beat existing methods even when $n < p$.
- Several benchmark datasets have been analyzed.

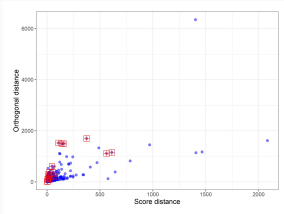
Simulation Results

Table 2: Estimated Bias and Mean Absolute Error of eigenvalues and Subspace Recovery Error of eigenvectors for different PCA algorithms (with $n = 50$).

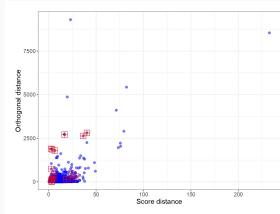
Metric	p	Classical	LOC	ROBPCA	Proj	RobCov	Grid	Gmed	PCP	DPD (0.25)	DPD (0.5)	DPD (0.75)	DPD (1)
Bias	10	0.321	0.757	0.381	0.429	0.589	0.757	0.14	1.065	0.329	0.281	0.138	0.067
	25	0.553	2.198	0.368	0.635	1.004	1.344	0.235	2.451	0.568	0.364	0.073	0.036
	50	1.467	4.602	0.829	1.796	NA	3.221	0.583	4.617	1.48	0.97	0.323	0.182
	100	2.66	9.414	1.028	2.692	NA	6.019	1.235	9.159	2.766	2.005	0.533	0.2
	250	7.033	23.805	3.245	8.006	NA	15.969	2.746	22.799	7.089	4.447	1.446	0.299
MAE	10	41.99	75.693	43.08	52.712	60.646	82.448	30.185	106.511	45.196	42.803	29.261	22.409
	25	85.197	219.781	63.246	93.376	112.498	165.495	65.545	245.114	90.83	74.713	45.453	41.413
	50	194.589	460.223	130.581	236.929	NA	406.956	144.635	462.199	209.841	172.386	110.373	96.321
	100	364.678	941.397	221.517	400.614	NA	665.195	267.786	916.897	394.475	317.498	173.981	142.885
	250	957.207	2380.505	618.404	1066.532	NA	1696.499	658.65	2283.607	1060.277	838.361	545.85	432.927
SRE	10	1.812	2.049	1.109	2.346	1.424	2.886	1.889	1.197	1.811	1.774	1.405	1.111
	25	2.14	2.422	1.021	2.645	2.212	4.19	2.26	1.276	2.152	1.832	1.111	1.03
	50	2.219	2.472	1.02	2.828	NA	4.985	2.314	2.265	2.24	1.819	1.201	1.049
	100	2.227	2.453	1.043	2.868	NA	3.86	2.326	2.272	2.242	1.868	1.153	1.007
	250	2.249	2.549	1.066	2.976	NA	3.901	2.362	2.302	2.262	1.767	1.16	1.007

Credit Card Fraud Detection Data Analysis

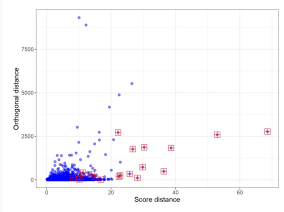
28 anonymized features with $n = 284807$ transactions, with $< 0.1\%$ being frauds. We take first 5 principal components, explaining over 80% of variation.



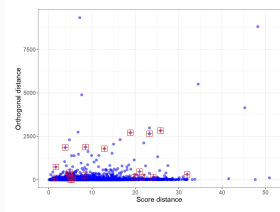
(a) Classical PCA



(b) GMedian



(c) rPCAdpd



(d) ROBPCA

Video Surveillance Background Modelling

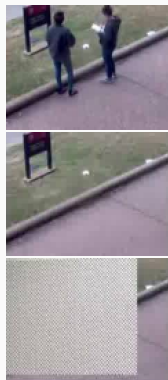


$$\mathbf{X}(\text{Video}) = \mathbf{L}(\text{Background}) + \mathbf{E}(\text{Foreground})$$

Following are seconds elapsed per frame for a 640×480 video.

1. Best existing robust SVD method ([Zhang et al., 2013](#)) - 312.26
2. Popular video surveillance method using Robust PCA ([Candes et al., 2011](#)) - 136.41
3. Go Decomposition ([Zhou and Tao, 2017](#)) - 12.06
4. rPCAdpd (ours) - 2.86

University of Houston Camera Tampering Dataset



Truth



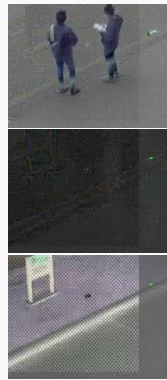
OP



ADMM








GoDec



rSVDdpd

References

-  Roy, Subhrajyoty, Ayanendranath Basu, and Abhik Ghosh. Robust Principal Component Analysis Using Density Power Divergence. *Journal of Machine Learning Research* 25, no. 324 (2024): 1–40. <http://jmlr.org/papers/v25/23-1096.html>.
-  Roy, Subhrajyoty, Abhik Ghosh. & Ayanendranath Basu. Robust singular value decomposition with application to video surveillance background modelling. *Stat Comput* 34, 178 (2024). <https://doi.org/10.1007/s11222-024-10493-7>.
-  Roy, Subhrajyoty, Abir Sarkar, Abhik Ghosh & Ayanendranath Basu. "Breakdown Point Analysis of the Minimum S-Divergence Estimator." arXiv preprint arXiv:2304.07466 (2023). - *in review*.
-  Basu, Ayanendranath, et al. "Robust and efficient estimation by minimising a density power divergence." *Biometrika* 85.3 (1998): 549-559.
-  Ghosh, Abhik, and Ayanendranath Basu. "Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression." (2013): 2420-2456.

Thank you!
Questions?