

Robust Matrix Factorization using Density Power Divergence and its Applications

Subhrajyoty Roy

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

Supervisors: Ayanendranath Basu & Abhik Ghosh

Nov 20, 2024

- Video Surveillance Background Modelling
- Robust Singular Value Decomposition
- Robust Principal Component Analysis
- Robust Matrix Rank Estimation
- Asymptotic Breakdown Analysis of Minimum Divergence Estimator

Video Surveillance Background Modelling

Background Modelling Problem



Background Modelling Problem



Applications ranging security, defence, object tracking, motion detection, video filters, etc.

=



+



Background Modelling as Low Rank Decomposition



$$h \times w \times t$$

$h \times w$ -pixels resolution

t -timeframes

Background Modelling as Low Rank Decomposition

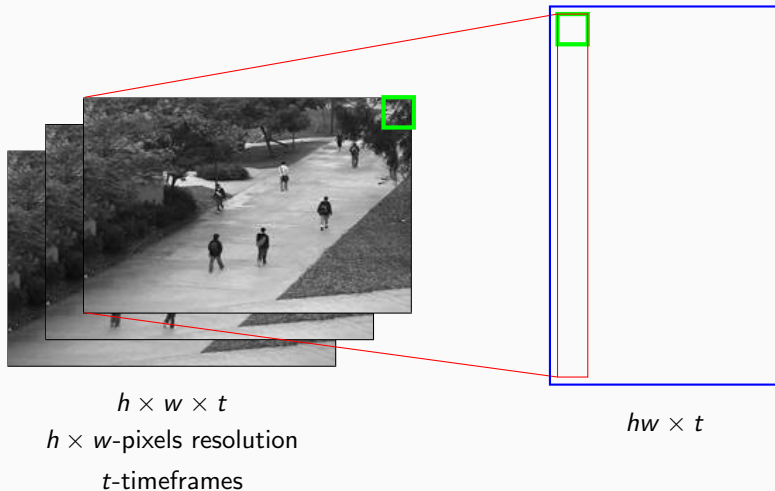


$h \times w \times t$

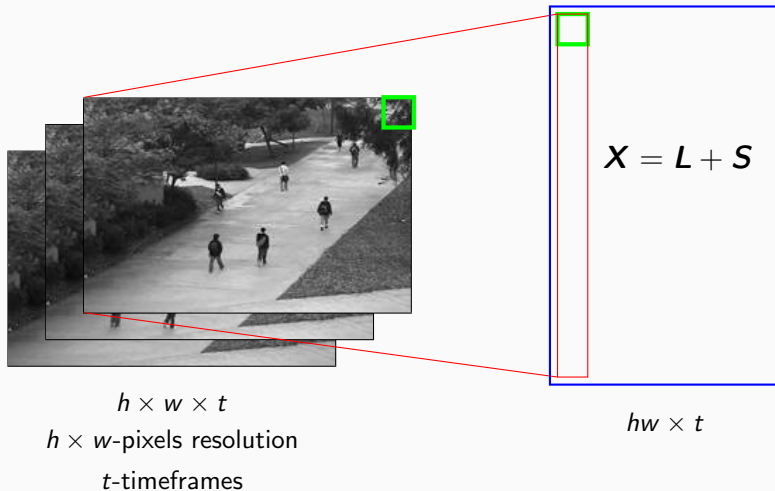
$h \times w$ -pixels resolution

t -timeframes

Background Modelling as Low Rank Decomposition



Background Modelling as Low Rank Decomposition



Singular Value Decomposition (SVD)

For a $n \times p$ matrix \mathbf{M} , its singular value decomposition is defined by the factorization

$$\mathbf{M} = \mathbf{U}_{n \times r} \mathbf{D}_{r \times r} (\mathbf{V}_{p \times r})^T$$

where,

1. \mathbf{U} , \mathbf{V} are unitary matrices (left and right singular vectors).
2. \mathbf{D} is a diagonal matrix comprising singular values, in decreasing order of magnitude.
3. r is the rank of the matrix

- Singular Value Decomposition (SVD) can be used to estimate \mathbf{L} .
- Advantages compared to Deep learning.
 1. Unsupervised. Need no training data.
 2. Low hardware requirements.
 3. Generalizable.
 4. Better theoretical guarantees.

Robust Background Modelling

- Singular Value Decomposition (SVD) can be used to estimate L .
- Advantages compared to Deep learning.
 1. Unsupervised. Need no training data.
 2. Low hardware requirements.
 3. Generalizable.
 4. Better theoretical guarantees.
- Real-life surveillance has many nuisances.
 1. Weather conditions (Fog / Rain)



Robust Background Modelling

- Singular Value Decomposition (SVD) can be used to estimate L .
- Advantages compared to Deep learning.
 1. Unsupervised. Need no training data.
 2. Low hardware requirements.
 3. Generalizable.
 4. Better theoretical guarantees.
- Real-life surveillance has many nuisances.
 1. Weather conditions (Fog / Rain)
 2. Naturally moving background



Robust Background Modelling

- Singular Value Decomposition (SVD) can be used to estimate L .
- Advantages compared to Deep learning.
 1. Unsupervised. Need no training data.
 2. Low hardware requirements.
 3. Generalizable.
 4. Better theoretical guarantees.
- Real-life surveillance has many nuisances.
 1. Weather conditions (Fog / Rain)
 2. Naturally moving background
 3. Camera tampering



Robust Singular Value Decomposition

SVD as alternating regression

$\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{N}$; \mathbf{L} low-rank, \mathbf{S} sparse and \mathbf{N} dense noise

SVD as alternating regression

$\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{N}$; \mathbf{L} low-rank, \mathbf{S} sparse and \mathbf{N} dense noise

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

SVD as alternating regression

$\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{N}$; \mathbf{L} low-rank, \mathbf{S} sparse and \mathbf{N} dense noise

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

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

SVD as alternating regression

$\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{N}$; \mathbf{L} low-rank, \mathbf{S} sparse and \mathbf{N} dense noise

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

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

$\Rightarrow X_{ij} = \sum_{k=1}^r a_{ki} b_{kj} + \epsilon_{ij}$; assume, $a_{ki} = \sqrt{\lambda_k} u_{ki}$, $b_{kj} = \sqrt{\lambda_k} v_{kj}$

SVD as alternating regression

$\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{N}$; \mathbf{L} low-rank, \mathbf{S} sparse and \mathbf{N} dense noise

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

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

$\Rightarrow X_{ij} = \sum_{k=1}^r a_{ki} b_{kj} + \epsilon_{ij}$; assume, $a_{ki} = \sqrt{\lambda_k} u_{ki}$, $b_{kj} = \sqrt{\lambda_k} v_{kj}$

Fixing j yields,

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

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 a true distribution G with density g and the model family of distributions $\{F_{\theta} : \theta \in \Theta\}$ with densities f_{θ} , the **MDPD functional** is defined as

$$T_{\alpha}(G) = \arg \min_{\theta \in \Theta} d_{\alpha}(g, f_{\theta}).$$

Minimum Density Power Divergence Estimator

Given an **independent and identically distributed** sample X_1, X_2, \dots, X_n modelled by family of densities $\{f_\theta : \theta \in \Theta\}$, the **MDPDE** is defined as

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

Minimum Density Power Divergence Estimator

Given an **independent and identically distributed** sample X_1, X_2, \dots, X_n modelled by family of densities $\{f_\theta : \theta \in \Theta\}$, the **MDPDE** is defined as

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

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

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

Robust SVD using DPD

For background modelling problem, usually $r = 1$, so we have

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

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

Robust SVD using DPD

For background modelling problem, usually $r = 1$, so we have

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

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

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

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

Robust SVD using DPD

For background modelling problem, usually $r = 1$, so we have

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

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

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

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

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

Robust SVD using DPD

For background modelling problem, usually $r = 1$, so we have

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

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

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

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

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

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

Robust SVD using DPD

For background modelling problem, usually $r = 1$, so we have

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

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

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

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

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

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

$$(\sigma^{(t+1)})^2 = \sum_{i,j} (x_{ij} - a_i^{(t+1)} b_j^{(t+1)})^2 \frac{w_{ij}^{(t)}}{np} / \left(\sum_{i,j} \frac{w_{ij}^{(t)}}{np} - \frac{\alpha}{(1+\alpha)} \int f^{1+\alpha} \right).$$

Different choices of model family

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.

Theoretical Results (Roy, Basu and Ghosh, 2024)

Assume that,

1. f is twice differentiable.
2. $f'(x) \leq 0$ for all $x > 0$ and $\frac{1}{x} > \alpha \frac{f'(x)}{f(x)} + s'(x) \frac{f(x)}{f'(x)}$, where $s(x)$ is the score function. Equivalently, $\psi'(x) < 0$ for $x > 0$.
3. $x^2\psi(x) = O(1)$ for all $x \geq 0$.

Under these assumptions,

1. The algorithm converges to a global optimum of the restricted DPD objective function.
2. Estimator has equivariance properties.

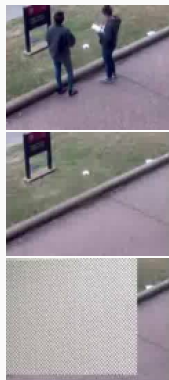
Additionally, suppose,

- The true error density g is symmetric.
- g satisfies $\int z^k e^{-\alpha z^2/2} g(z) dz = O(1)$ for $k = 0, 1, \dots, 4$.
- The error variance $\sigma^2 \sim (np)^{-1/2}$.

Then as both $n \rightarrow \infty, p \rightarrow \infty$, but $n/p \rightarrow c \in (0, \infty)$, the rSVDdpd estimator converges to the true SVD estimates (i.e., the true minimum of DPD between g and $f(\cdot/\sigma)$).

1. Simulation studies show approximately 20 – 30% reduction in RMSE compared to existing robust SVD methods for up to moderate contamination.
2. Massively scalable with no restriction on the dimensions p . A higher dimension results in more contrastingly superior performance. (Explained later).
3. Massively parallelizable and fast. Following are seconds elapsed per frame for a 640×480 video.
 - 3.1 Best existing robust SVD method ([Zhang et al., 2013](#)) - 312.26
 - 3.2 Popular video surveillance method using Robust PCA ([Candes et al., 2011](#)) - 136.41
 - 3.3 Go Decomposition ([Zhou and Tao, 2017](#)) - 12.06
 - 3.4 rSVDdpd (ours) - 2.86

University of Houston Camera Tampering Dataset



Truth



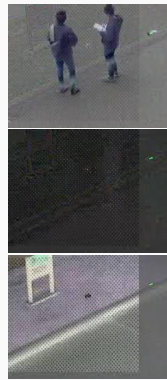
OP



ADMM



GoDec



rSVDdspd

Robust Principal Component Analysis

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.

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

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

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

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

If we consider the same SVD decomposition

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

with $\mathbf{D} = \text{Diag}(\sqrt{n\lambda_1}, \dots, \sqrt{n\lambda_p})$. So, we can apply the same *rSVD* procedure on the stacked matrix \mathbf{X} .

The theoretical analysis is a bit different.

1. Only the rows of \mathbf{E} are independent now.
2. Asymptotic analysis needs to keep p fixed, but $n \rightarrow \infty$.
3. \mathbf{U} is a random orthogonal matrix.

The theoretical analysis is a bit different.

1. Only the rows of \mathbf{E} are independent now.
2. Asymptotic analysis needs to keep p fixed, but $n \rightarrow \infty$.
3. \mathbf{U} is a random orthogonal matrix.

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

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

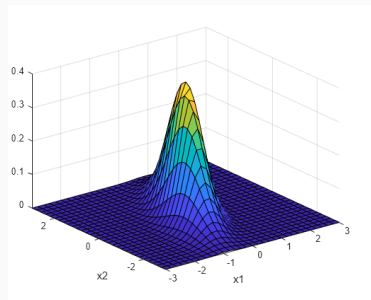
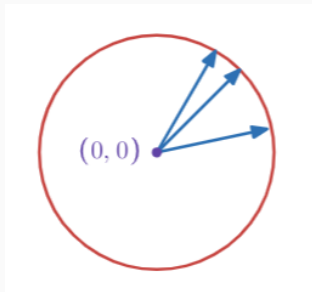
Results on rPCAdpd (Roy, Basu and Ghosh, 2023)

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

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

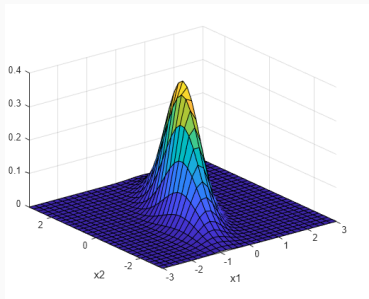
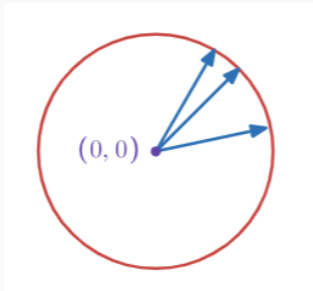
Results on rPCAdpd (Roy, Basu and Ghosh, 2023)

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.

rPCAdpd for Gaussian case (Roy, Basu and Ghosh, 2023)

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 estimates $\hat{\gamma}_j$ and \hat{v}_j are \sqrt{n} -consistent.
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

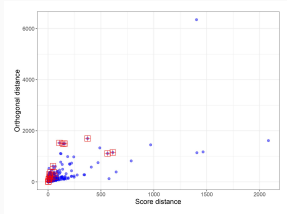
$$\frac{(1 + \alpha)^{p/4}}{(1 + 2\alpha)^{2+p/2}} \sum_{k,l} \left(1 - \frac{\gamma_k}{\gamma_l}\right) \frac{\partial v_k}{\partial \eta} v_k v_l^T \left(\frac{\partial v_l}{\partial \eta}\right)^T$$

4. The estimated eigenvalues $\{\hat{\gamma}_k\}_{k=1}^p$ and the estimated eigenvectors $\{\hat{v}_k\}_{k=1}^p$ are asymptotically independent.

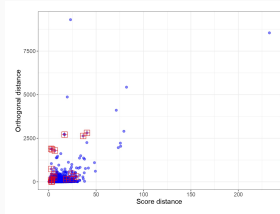
- The estimator has a bounded influence function, if $\hat{\mu}$ has bounded influence.
- Extensive simulation studies have been performed and compared against existing methods.
 - Multiple levels of contaminations from 5% to 20%.
 - Cauchy and t_5 distribution of errors.
 - Dimensions ranging from $p = 10$ to $p = 250$, with $n = 100$.
- rPCAdpd beat existing methods even when $n < p$.
- Several benchmark datasets have been analyzed.

Credit Card Fraud Detection Data Analysis

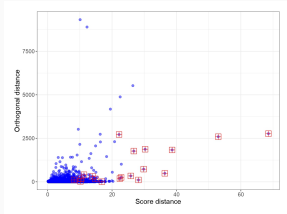
28 anonymized features with $n = 284807$ transactions, with $< 0.1\%$ being frauds. We take first 5 principal components, explaining over 80% of variation.



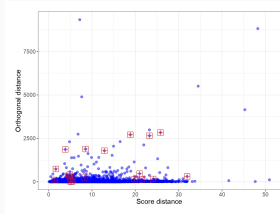
(a) Classical PCA



(b) GMedian



(c) rPCAdpd



(d) ROBPCA

Robust Rank Estimation

Problem Description

Given a data matrix \mathbf{X} with decomposition

$$\mathbf{X} = \mathbf{L} + \frac{\sigma}{\sqrt{\max(n, p)}} \mathbf{E},$$

where \mathbf{L} is low-rank and \mathbf{E} is noise matrix, the problem is to determine the rank of \mathbf{L} .

Problem Description

Given a data matrix \mathbf{X} with decomposition

$$\mathbf{X} = \mathbf{L} + \frac{\sigma}{\sqrt{\max(n, p)}} \mathbf{E},$$

where \mathbf{L} is low-rank and \mathbf{E} is noise matrix, the problem is to determine the rank of \mathbf{L} .

But, this is not identifiable!

$$\mathbf{X} = \mathbf{L} + \mathbf{1}\mathbf{1}^\top = \mathbf{L}' + \mathbf{0},$$

\mathbf{E} cannot be low-rank itself.

On the other hand, if \mathbf{E} is high-rank, \mathbf{E} may share some eigenvectors of \mathbf{L} itself.

Common restriction: is orthogonal equivariance of the distribution of \mathbf{E} .

Our restriction: is exchangeability of the rows and columns of \mathbf{E} .

Common restriction: is orthogonal equivariance of the distribution of \mathbf{E} .

Our restriction: is exchangeability of the rows and columns of \mathbf{E} .

1. Elbow method and Thresholding. ([Shabalin and Nobel, 2013](#)) - Very adhoc, little theoretical foundation.

Common restriction: is orthogonal equivariance of the distribution of \mathbf{E} .

Our restriction: is exchangeability of the rows and columns of \mathbf{E} .

1. Elbow method and Thresholding. ([Shabalín and Nobel, 2013](#)) - Very adhoc, little theoretical foundation.
2. Penalized methods such as AIC ([Akaike, 1973](#)), BIC([Schwartz, 1978](#)), IC1-3 ([Bai and Ng, 2002](#)), DIC ([Karagrigoriou and Papaioannou, 2008](#)), etc. - does not have theoretical guarantee when both matrix dimensions grow to infinity.

Common restriction: is orthogonal equivariance of the distribution of \mathbf{E} .

Our restriction: is exchangeability of the rows and columns of \mathbf{E} .

1. Elbow method and Thresholding. ([Shabalín and Nobel, 2013](#)) - Very adhoc, little theoretical foundation.
2. Penalized methods such as AIC ([Akaike, 1973](#)), BIC([Schwartz, 1978](#)), IC1-3 ([Bai and Ng, 2002](#)), DIC ([Karagrigoriou and Papaioannou, 2008](#)), etc. - does not have theoretical guarantee when both matrix dimensions grow to infinity.
3. Cross-validation methods ([Wold, 1978](#); [Gabriel, 2002](#); [Owen and Perry, 2009](#)). - extremely computationally expensive.

Divergence Information Criterion (DIC)

Karagrigoriou and Papaioannou, 2008 proposed a robust model selection criterion using density power divergence as

$$\text{DIC}_\alpha(r) = H_\alpha^{(r)}(\hat{\theta}) + r(\alpha + 1)(2\pi)^{-\alpha/2} \left(\frac{1 + \alpha}{1 + 2\alpha} \right)^{3/2},$$

where $H_\alpha^{(r)}(\theta)$ is the form of density power divergence (without the last term which is free of θ).

Divergence Information Criterion (DIC)

[Kurata, 2024](#) extended this to the general case, which for linear regression setup $Y = X\beta + \epsilon$ with $\epsilon \sim f(\cdot/\sigma)$ reduces to

$$\text{DIC}_\alpha(r) = H_\alpha^{(r)}(\hat{\beta}) + \frac{1}{n} \hat{\sigma}^{-\alpha} \text{trace}(\mathbf{X}^\top \mathbf{X}) \frac{C_{2\alpha}}{C_\alpha},$$

where $C_\alpha = \int (f'(x))^2 f^{\alpha-1}(x) dx$ for symmetric density function f .

Divergence Information Criterion (DIC)

[Kurata, 2024](#) extended this to the general case, which for linear regression setup $Y = X\beta + \epsilon$ with $\epsilon \sim f(\cdot/\sigma)$ reduces to

$$\text{DIC}_\alpha(r) = H_\alpha^{(r)}(\hat{\beta}) + \frac{1}{n} \hat{\sigma}^{-\alpha} \text{trace}(\mathbf{X}^\top \mathbf{X}) \frac{C_{2\alpha}}{C_\alpha},$$

where $C_\alpha = \int (f'(x))^2 f^{\alpha-1}(x) dx$ for symmetric density function f .

For our case, we have $\mathbf{X} = \mathbf{A}\mathbf{B}^\top + \mathbf{E}$. Therefore,

$$\text{DIC}_\alpha(r) \mid \mathbf{A} \approx H_\alpha(\mathbf{A}, \hat{\mathbf{B}}, \hat{\sigma}^2) + \frac{p}{n} \hat{\sigma}^{-\alpha} \text{trace}(\mathbf{A}^\top \mathbf{A}) \frac{C_{2\alpha}}{C_\alpha}$$

Divergence Information Criterion (DIC)

[Kurata, 2024](#) extended this to the general case, which for linear regression setup $Y = X\beta + \epsilon$ with $\epsilon \sim f(\cdot/\sigma)$ reduces to

$$\text{DIC}_\alpha(r) = H_\alpha^{(r)}(\hat{\beta}) + \frac{1}{n} \hat{\sigma}^{-\alpha} \text{trace}(\mathbf{X}^\top \mathbf{X}) \frac{C_{2\alpha}}{C_\alpha},$$

where $C_\alpha = \int (f'(x))^2 f^{\alpha-1}(x) dx$ for symmetric density function f .

For our case, we have $\mathbf{X} = \mathbf{A}\mathbf{B}^\top + \mathbf{E}$. Therefore,

$$\text{DIC}_\alpha(r) \mid \mathbf{A} \approx H_\alpha(\mathbf{A}, \hat{\mathbf{B}}, \hat{\sigma}^2) + \frac{p}{n} \hat{\sigma}^{-\alpha} \text{trace}(\mathbf{A}^\top \mathbf{A}) \frac{C_{2\alpha}}{C_\alpha}$$

but $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_r$. So,

$$\text{DIC}_\alpha(r) \mid \mathbf{A} \approx H_\alpha(\mathbf{A}, \hat{\mathbf{B}}, \hat{\sigma}^2) + r \frac{p}{n} \hat{\sigma}^{-\alpha} \frac{C_{2\alpha}}{C_\alpha},$$

which is free of \mathbf{A} .

Divergence Information Criteria for Matrix Rank Estimation (DICMR) is proposed as

$$Q_{\alpha}^{(r)}(\theta) = H_{\alpha}(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\sigma}^2) + r \frac{(1/n + 1/p)}{2} \hat{\sigma}^{-\alpha} \frac{C_{2\alpha}}{C_{\alpha}}$$

- Resulting criterion can be shown to achieve selection consistency in the asymptotic regime where both n and p tend to ∞ subject to the restriction $(n/p) \rightarrow c \in (0, \infty)$.
- It is first-order B-robust (i.e., the influence curve is bounded).
- Simple and easy to implement. As fast as the *rSVDdpd* algorithm.

- Resulting criterion can be shown to achieve selection consistency in the asymptotic regime where both n and p tend to ∞ subject to the restriction $(n/p) \rightarrow c \in (0, \infty)$.
- It is first-order B-robust (i.e., the influence curve is bounded).
- Simple and easy to implement. As fast as the *rSVDdpd* algorithm.
- Numerical simulations show that

Existing all method with classical SVD \ll

Existing penalized methods with *rSVDdpd* \ll

Existing CV methods with *rSVDdpd* \approx

DICMR with *rSVDdpd*

Asymptotic Breakdown Analysis

What is breakdown point?

For a functional $T(G)$, its asymptotic breakdown point is

$$\epsilon^*(T) := \sup \left\{ \epsilon : \epsilon \in [0, 1/2] \text{ and, } \inf_{\theta_\infty \in \partial\Theta} \liminf_{m \rightarrow \infty} \|T(G_{\epsilon,m}) - \theta_\infty\| \text{ for all } \{K_m\}_{m=1}^\infty \right\},$$

where $G_{\epsilon,m} = (1 - \epsilon)G + \epsilon K_m$.

It is the maximum proportion of contamination that an estimator can tolerate before producing an egregiously bad estimate, i.e., in the boundary of parameter space.

Popular classical robust estimators such as M-estimators have an asymptotic breakdown point $\leq 1/(p + 1)$.

But, empirical verification shows that even for large dimensional matrices, rSVDdpd and rPCAdpd perform well.

Popular classical robust estimators such as M-estimators have an asymptotic breakdown point $\leq 1/(p + 1)$.

But, empirical verification shows that even for large dimensional matrices, rSVDdpd and rPCAdpd perform well.

Turns out, MDPDE has a dimension-free lower bound to its asymptotic breakdown point, in many scenarios.

Popular classical robust estimators such as M-estimators have an asymptotic breakdown point $\leq 1/(p + 1)$.

But, empirical verification shows that even for large dimensional matrices, rSVDdpd and rPCAdpd perform well.

Turns out, MDPDE has a dimension-free lower bound to its asymptotic breakdown point, in many scenarios.

Moreover, this is true for a larger class of minimum divergence estimators.

Generalized S-divergence

Ghosh et al, 2017 proposed a family of super-divergences (SD) as

$$d_{(\alpha,\lambda)}(g, f_\theta) = \frac{1}{A} \int f_\theta^{1+\alpha} - \frac{1+\alpha}{AB} \int f_\theta^B g^A + \frac{1}{B} \int g^{1+\alpha}$$

where $A = 1 + \lambda(1 - \alpha)$, $B = \alpha - \lambda(1 - \alpha)$.

Generalized S-divergence

Ghosh et al, 2017 proposed a family of super-divergences (SD) as

$$d_{(\alpha,\lambda)}(g, f_\theta) = \frac{1}{A} \int f_\theta^{1+\alpha} - \frac{1+\alpha}{AB} \int f_\theta^B g^A + \frac{1}{B} \int g^{1+\alpha}$$

where $A = 1 + \lambda(1 - \alpha)$, $B = \alpha - \lambda(1 - \alpha)$.

Maji et al, 2016 proposed a family of logarithmic super-divergence (LSD)

$$d_{(\alpha,\lambda)}^*(g, f_\theta) = \frac{1}{A} \ln \left(\int f_\theta^{1+\alpha} \right) - \frac{1+\alpha}{AB} \ln \left(\int f_\theta^B g^A \right) + \frac{1}{B} \ln \left(\int g^{1+\alpha} \right)$$

Generalized S-divergence

Ghosh et al, 2017 proposed a family of super-divergences (SD) as

$$d_{(\alpha,\lambda)}(g, f_\theta) = \frac{1}{A} \int f_\theta^{1+\alpha} - \frac{1+\alpha}{AB} \int f_\theta^B g^A + \frac{1}{B} \int g^{1+\alpha}$$

where $A = 1 + \lambda(1 - \alpha)$, $B = \alpha - \lambda(1 - \alpha)$.

Maji et al, 2016 proposed a family of logarithmic super-divergence (LSD)

$$d_{(\alpha,\lambda)}^*(g, f_\theta) = \frac{1}{A} \ln \left(\int f_\theta^{1+\alpha} \right) - \frac{1+\alpha}{AB} \ln \left(\int f_\theta^B g^A \right) + \frac{1}{B} \ln \left(\int g^{1+\alpha} \right)$$

In general, consider any ψ function, to have

$$d_{(\alpha,\lambda)}^\psi(g, f_\theta) = \frac{1}{A} \psi \left(\int f_\theta^{1+\alpha} \right) - \frac{1+\alpha}{AB} \psi \left(\int f_\theta^B g^A \right) + \frac{1}{B} \psi \left(\int g^{1+\alpha} \right).$$

It is a valid divergence if and only if $\Psi(x) := \psi(e^x)$ is increasing and convex.

Results on Asymptotic Breakdown Point

Under the (informal) assumptions,

1. $A \geq 0$ and $B > 0$.
2. The densities k_m and g are mutually singular as $m \rightarrow \infty$.
3. For any $\theta_m \rightarrow \theta_\infty$, f_{θ_m} and g are mutually singular.
4. $\sup_{\theta \in \Theta \setminus \partial\Theta} \int f_\theta^{1+\alpha} < \infty$ and $\sup_m \int k_m^{1+\alpha} < \infty$.

Let $C = \limsup_m \int k_m^{1+\alpha}$. Then, the minimum GSD functional has an asymptotic breakdown point satisfying $\epsilon^* \geq \epsilon_0$, where

$$\psi \left((1 - \epsilon_0)^A \int f_{\theta_g}^B g^A \right) = \frac{B}{1 + \alpha} \psi \left(\int f_{\theta_g}^{1+\alpha} \right) + \frac{A}{1 + \alpha} \psi((C\epsilon_0)^{1+\alpha}).$$

Breakdown point for specific setups

For location estimation, the asymptotic breakdown point of MGSD functional is $1/2$.

Breakdown point for specific setups

For location estimation, the asymptotic breakdown point of MGSD functional is $1/2$.

In many other cases, e.g. scale estimation for normal family, exponential distribution, and shape estimation for gamma family, the breakdown is at least

$$\begin{cases} (B/(1+\alpha))^{1/A} & A > 0, \\ e^{-1/(1+\alpha)} & A = 0, \end{cases}$$

for $\psi(x) = x, \alpha > 0$ case, which is data-dimension free.

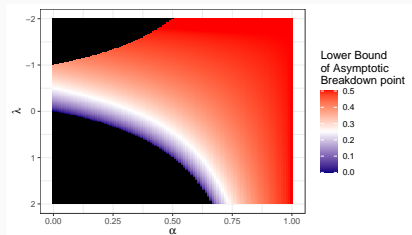
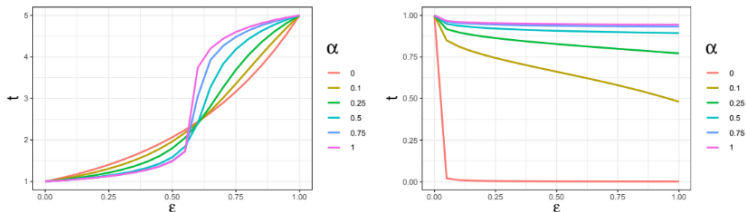







Figure 5.5: Behaviour of MDPD estimates under gamma density model as a function of the contamination proportion ϵ , for two values of shape parameter $t_0 = 10$ (Left) and $t_0 = 0.001$ (Right).










More examples are present in [Roy et al. \(2023\)](#).

- SVD and PCA are popular matrix factorization techniques with many applications, including background modelling in video data.
- Due to outlier sensitivity, real-life applications demand robust matrix factorization techniques.
- Using MDPDE, we can obtain a simple and scalable alternating regression-based algorithm to compute PCA and SVD robustly.
- For a wider class of minimum divergence estimators (including MDPDE), the asymptotic breakdown point remains bounded away from 0 even when data-dimension p is arbitrarily large.

References

-  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, 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, Abir Sarkar, Abhik Ghosh & Ayanendranath Basu. "Breakdown Point Analysis of the Minimum S-Divergence Estimator." arXiv preprint arXiv:2304.07466 (2023). - *in review*.
-  Roy, Subhrajyoty, Supratik Basu, Abhik Ghosh & Ayanendranath Basu. "Characterization of Generalized Alpha-Beta Divergence and its Associated Properties." - *in preparation*.
-  Roy, Subhrajyoty, Abhik Ghosh & Ayanendranath Basu. "Divergence Information Criterion for Robust Rank Selection" - *in preparation*.

Additional References

-  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.
-  Karagrigoriou, Alex, and Takis Papaioannou. "On measures of information and divergence and model selection criteria." *Statistical Models and Methods for Biomedical and Technical Systems* (2008): 503-518.
-  Shabalin, Andrey A., and Andrew B. Nobel. "Reconstruction of a low-rank matrix in the presence of Gaussian noise." *Journal of Multivariate Analysis* 118 (2013): 67-76.
-  Bai, Jushan, and Serena Ng. "Determining the number of factors in approximate factor models." *Econometrica* 70.1 (2002): 191-221.
-  Owen, Art B., and Patrick O. Perry. "Bi-cross-validation of the SVD and the nonnegative matrix factorization." (2009): 564-594.
-  Kurata, Sumito. "On robustness of model selection criteria based on divergence measures: Generalizations of BHHJ divergence-based method and comparison." *Communications in Statistics-Theory and Methods* 53.10 (2024): 3499-3516.

Thank you!
Questions?