DiffRed: Dimensionality Reduction guided by stable rank

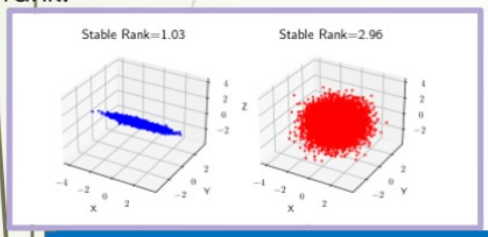
Prarabdh Shukla, Gagan Raj Gupta (CSE, IIT Bhilai), Kunal Dutta (University of Warsaw)

Motivation

- **PCA** takes into account the structure of the data, but (i) is computationally expensive, (ii) doesn't have theoretical guarantees, and (iii) ignores information in low-variance directions.
- **Random Map embeddings** are (i) simple to apply, (ii) data-agnostic, and (iii) have theoretical guarantees on pairwise distortion. But target dimension bounds are much higher.

Question: Is it possible to achieve tighter bounds on distortion metrics such as *Stress* and *M1* by incorporating the structure of the data?

Key Insight: We quantify the structure of the data as a property of the data matrix A in terms of its stable rank.



$$\rho(A) = \frac{\sum_{i=1}^{rank(A)} \sigma_i^2}{\sigma_1^2}$$

STABLE RANK

DiffRed Algorithm

Algorithm 1: *DiffRed* Algorithm

Input: A, k_1, k_2, η
compute SVD $A = U\Sigma V^\top$
compute $A_{k_1} \leftarrow \sum_{i=1}^{k_1} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ and $A^* \leftarrow A - A_{k_1}$
Let V_{k_1} be the matrix with the k_1 leftmost columns of V
$Z \leftarrow AV_{k_1}$ // Project A along V_{k_1}
Initialize $min = \infty$
Initialize $T, T_{min} \in \mathbb{R}^{n \times k_2}$
// η Monte Carlo iterations
for $i = 0, \cdots, \eta$ **do**
  Sample $G \in \mathbb{R}^{D \times k_2}$ where $G_{ij} \sim \mathcal{N}(0,1)$ i.i.d.
  $G \leftarrow \frac{1}{\sqrt{k_2}} G$
  $T \leftarrow A^* G$
  if $\Lambda_{M_1}(A^*, T) < min$ **then**
    $T_{min} \leftarrow T$
$R \leftarrow T_{min}$
// T_{min} is the projection with least Λ_{M_1}
$\tilde{A} \leftarrow [Z|R]$
return $\tilde{A}$

A: Data Matrix
$k_1 + k_2 = d$ (Target Dimension)
η: Monte Carlo iterations

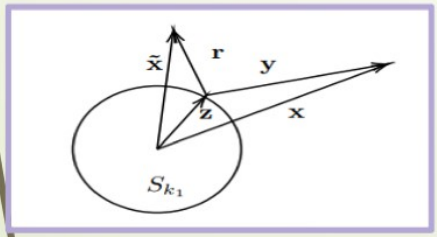
Calculating residual matrix A^*

Monte Carlo search to find best random projection for A^*

Running Time

$O(Dn \cdot \min(D, n) + \eta k_2 D)$

Analysis



DiffRed maps each vector $\boldsymbol{x} \in \mathbb{R}^D$ to $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$. Each vector's component $\boldsymbol{z}$ in the subspace of the first k_1 principal vectors

Theoretical Results

- **Theorem 2 [Bound on *M1*]:** Given a data matrix $A \in \mathbb{R}^{n \times D}$ and non negative integers k_1 and k_2, let the application of *DiffRed* algorithm on A with target dimensions k_1 and k_2 return the embedding matrix $\tilde{A} \in \mathbb{R}^{n \times d}$ where $d = k_1 + k_2$. Then,

$$\mathbb{P}[\Lambda_{M1}(A) \geq \varepsilon] \leq 2e^{\left(-\frac{c_1 \varepsilon^2 k_2 \rho(A^*)}{(1-p)^2}\right)}$$

where $c_1 > 0$ is a constant.

- *Proof Sketch:* We know $\tilde{A} = [Z|R]$, therefore, $\|\tilde{A}\|_F^2 = \|Z\|_F^2 + \|R\|_F^2$. Now, $\|Z\|_F^2 = p\|A\|_F^2$. We then bound the residual term $\|R\|_F^2$ using Hanson Wright inequality[1] to get the stated inequality.

- **Theorem 7 [Bound on *Stress*]:** Given a data matrix consisting of data points $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n} \in \mathbb{R}^D$, k_1 and k_2, let the application of *DiffRed* algorithm return the points $\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n \in \mathbb{R}^d$. Then with probability at least 1/2,

$$\Lambda_S = O\left(\sqrt{\frac{1-p}{k_2}}\right)$$

- *Proof Sketch:* Again, we split the difference vectors into their components along Z and R. Then we use the triangle inequality to upper bound this in terms of the stress of the residual, which we then bound using Bartal et. al's result[2].

Experimental Results

Results on ***Stress*** (Λ_S)

$$\Lambda_S = \left(\frac{\sum_{i,j}(\|\boldsymbol{d}_{ij}\| - \|\widetilde{\boldsymbol{d}_{ij}}\|)^2}{\sum_{i,j}\|\boldsymbol{d}_{ij}\|^2}\right)^{\frac{1}{2}}$$

Dataset	D	d	Λ_{M_1}						
			DiffRed	PCA	RMap	S-PCA	K-PCA	UMap	T-SNE ($d=2$)
Bank	17	5	**2.82e-05**	0.54	0.38	0.58	0.95	94.89	2659.70
Hatespeech	100	10	**1.91e-04**	0.66	0.06	0.68	0.99	240.50	2298.09
FMnist	784	10	**1.92e-04**	0.60	0.11	0.64	1.00	241.35	829.54
Cifar10	3072	10	**1.31e-04**	0.49	0.09	0.54	1.00	166.84	604.71
geneRNASeq	20.5k	10	**7.96e-05**	0.94	0.31	0.95	1.00	328.72	8,761.41
Reuters30k	30.9k	10	**1.27e-04**	0.88	0.03	0.88	1.00	196.97	2393.31
APTOS 2019	509k	10	**4.09e-05**	0.81	0.24	-	-	-	-
DIV2k	6.6M	10	**7.07e-05**	0.66	0.05	-	-	-	-

Stress is the normalized value of the root-mean-squared (RMS) value of the distortion of pairwise distances. Minimizing stress is important for preserving important structures such as clusters and nearest neighbors. *Stress* finds use in MDS, Psychology, 3D face recognition, medical imaging and also as an important metric to measure projection quality.

Theorem 7

$$\Lambda_S = O\left(\sqrt{\frac{1-p}{k_2}}\right)$$

Results on ***M1*** (Λ_{M_1})

$$\Lambda_{M_1}(A, \tilde{A}) = \left|1 - \frac{\sum_{i=1}^n \|\tilde{\boldsymbol{x}}\|_2^2}{\sum_{i=1}^n \|\boldsymbol{x}\|_2^2}\right|$$

Dataset	D	d	Λ_S								
			DiffRed	PCA	RMap	S-PCA	K-PCA	UMap	UMap2	T-SNE ($d=2$)	T-SNE2 ($d=2$)
Bank	17	6	**0.02**	0.03	0.17	0.04	0.47	7.07	0.35	52.44	0.72
Hatespeech	100	10	**0.15**	0.36	0.16	0.36	0.65	5.29	0.46	32.86	0.38
FMnist	784	10	**0.12**	0.19	0.15	0.21	0.68	4.02	0.42	24.49	0.38
Cifar10	3072	10	**0.13**	0.21	0.16	0.24	0.69	1.26	0.60	16.88	0.31
geneRNASeq	20.5k	10	**0.13**	0.21	0.16	0.25	0.70	18.72	0.47	164.89	1.21
Reuters30k	30.9k	10	**0.155**	0.49	0.157	0.49	0.71	3.35	0.44	18.02	0.31
APTOS 2019	509k	10	**0.10**	0.12	0.16	-	-	-	-	-	-
DIV2k	6.6M	10	**0.14**	0.31	0.16	-	-	-	-	-	-

M1 distortion is the distortion of the mean-squared pair-wise distance. Minimizing *M1* ensures that the low dimensional representations $\tilde{A}$ have similar *energy* or *total variance* as the original data matrix A.

Theorem 2

$$\Lambda_{M_1} = O\left(\frac{1-p}{\sqrt{k_2 \rho(A)}}\right)$$