Random projection to dimention reduction of large scale data

Siamak Dehbod

Supervisor Dr. Adel Mohammadpour
Advisor Dr. Hadi Zare

Amirkabir University of Technology

January 20, 2019

Indroduction

Masive Data

Large Scale Data

- more dimensions than records (D $>$ n)
- unable to compute $A^T A$ and using PCA
- online/stream calculation
- unable to store whole data
- covariance is not finite

Big Data

- Volume
- Velocity
- Variety

Heavy tail data

- Common in real data like market data, rare events are more probable than normal distribution
- Random variable X with right side heavy-tail distribution:

$$P(X > x) \sim cx^{-\alpha}, x \to \infty$$

Dimention Reduction

Random Coordinate Sampling

Pros:

- Simplicity $O(nk)$
- Flexability for estimating various summary statistics

Cons:

- Not accurate for losing rare events
- Not suitable for sparse data

Principal components analysis

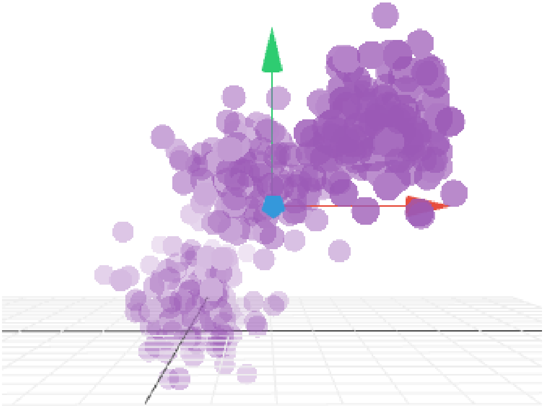


Figure 1:

Principal components analysis

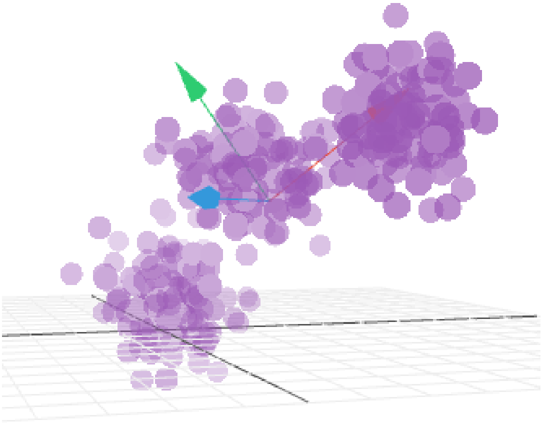


Figure 2:

Clustring

Clustering

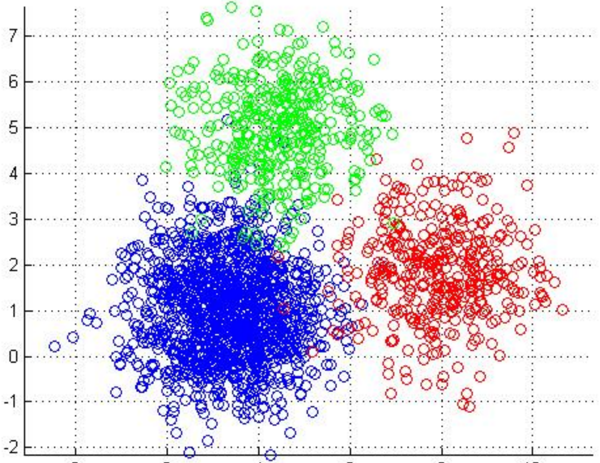


Figure 3: Clustering

k-means

Non-hierarchical clustering method

minimize within-cluster sum of squares

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Adjusted Rand Index

$$\frac{a+b}{a+b+c+d}$$

Adjusted Rand Index

Class \Cluster	v_1	v_2	$\ldots$	v_C	Sums
u_1	n_{11}	n_{12}	$\ldots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\ldots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	$n_{R.}$
Sums	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.C}$	$n_{..} = n$

Adjusted Rand Index

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}} \tag{1}$$

C_e

$$C_e = 100(\mathrm{ARI}_d - \mathrm{ARI}_p)$$

$$(d < p)$$

Applications

Distances

$$a = u_1^T u_2 = \sum_{i=1}^{D} u_{1,i} u_{2,i} \tag{2}$$

$$d_{(\alpha)} = \sum_{i=1}^{D} |u_1 - u_2|^{\alpha} \tag{3}$$

Distances

$$A^T A : O(n^2 D)$$

$$O(n^2 \hat{f})$$

Database Query Optimization

joins and execution plan

Sub-linear Nearest Neighbor Searching

$$O(nD) \to O(nk)$$

$$(\alpha > 1) l_\alpha \to O(n^\gamma)(\gamma < 1)$$

Stable Random Projection

Stable Distribution

Stable Distribution

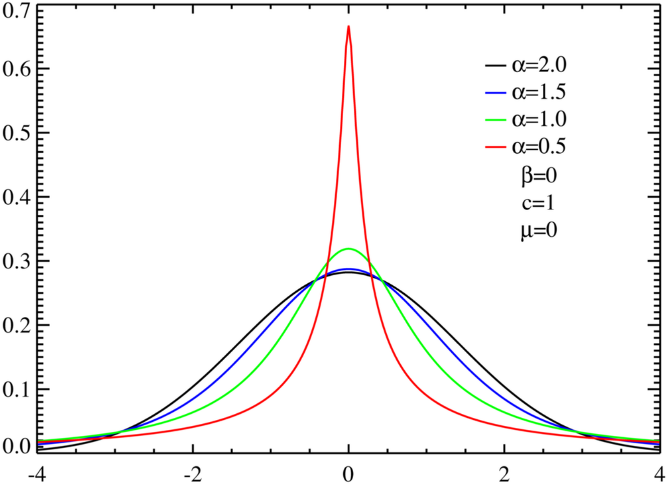


Figure 4: Stable distribution

Stable Distribution

$$X_1 + X_2 + \cdots + X_n =^d c_n X + d_n$$

Gaussian/normal:

$$f(x) = (2\pi)^{1/2} \exp(-x^2/2)$$

Cauchy:

$$f(x) = 1/(\pi(1 + x^2))$$

Stable Normal N(0,1)

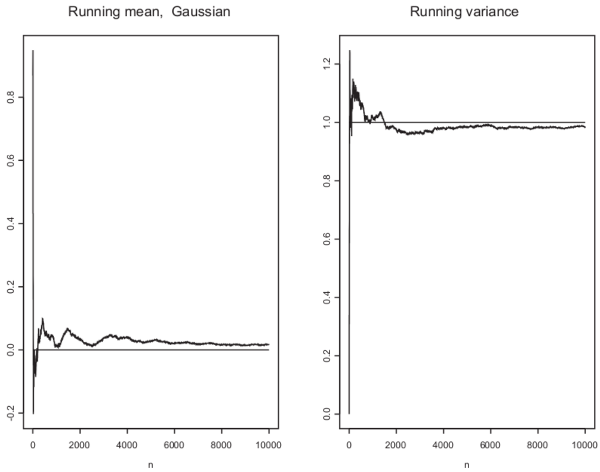


Figure 5: Normal $\alpha = 2$

Stable $\alpha = 1.5$

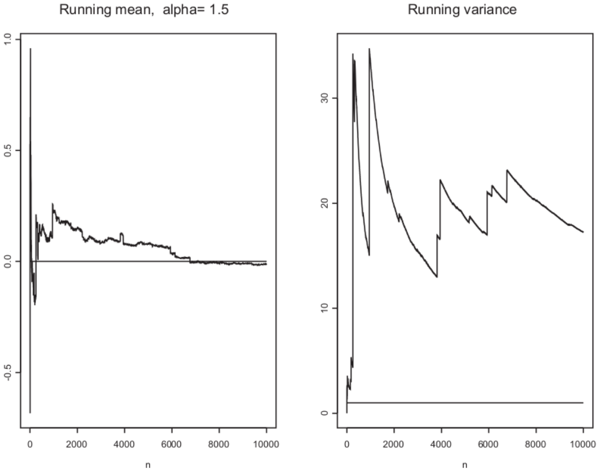


Figure 6: $\alpha = 1.5$

Stable $\alpha = 0.75$

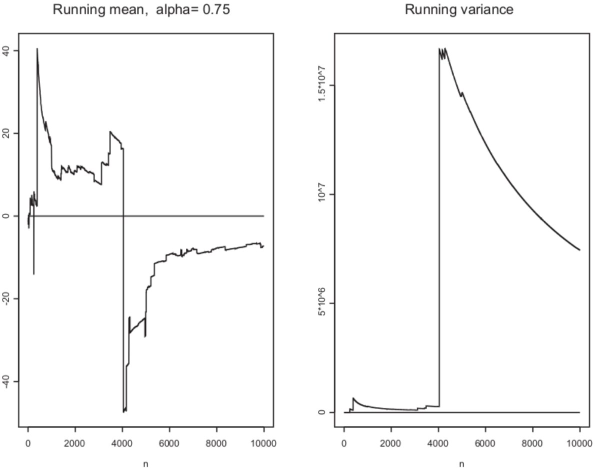


Figure 7: $\alpha = 0.75$

Stable Random Projection

Stable Random Projection

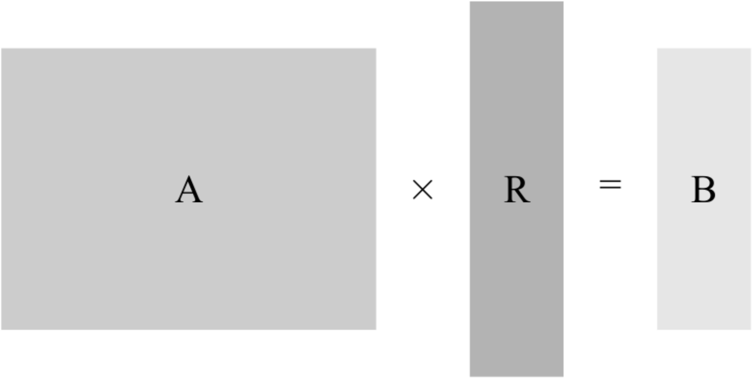


Figure 8:

Stable Random Projection

Johnson-Lindenstrauss Lemma:

$$k = O\left(\frac{\log n}{\epsilon^2}\right)$$

$$l_2 : 1 \pm \epsilon$$

Statistical estimation problem

$$v_{1,j} \sim S\left(\alpha, \sum_{i=1}^{D} |u_{1,i}|^{\alpha}\right), \quad v_{2,j} \sim S\left(\alpha, \sum_{i=1}^{D} |u_{2,i}|^{\alpha}\right), \tag{4}$$

$$x_j = v_{1,j} - v_{2,j} \sim S\left(\alpha, d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^{\alpha}\right). \tag{5}$$

Couchy Random Projection

$$d = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|$$

Very Sparse Random Projection

$$\{-1, 0, 1\}$$

$$\left\{ \frac{1}{2s}, 1 - \frac{1}{s}, \frac{1}{2s} \right\}$$

$$O(Dk) \rightarrow O(Dk/s)$$

l_α Random Projection

$$d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha$$

Data & Implementation

Data summary

Dataset	n	D	N_{class}
Thyroid	215	5	3
Iris	150	4	3
Diabetes	145	3	3
Swiss Banknotes	200	6	2
Seeds	210	7	3
Mice Protein Expression	1080	77	8
Crabs	200	6	2

Results

C_e

$$C_e = 100(\mathrm{ARI}_d - \mathrm{ARI}_p)$$

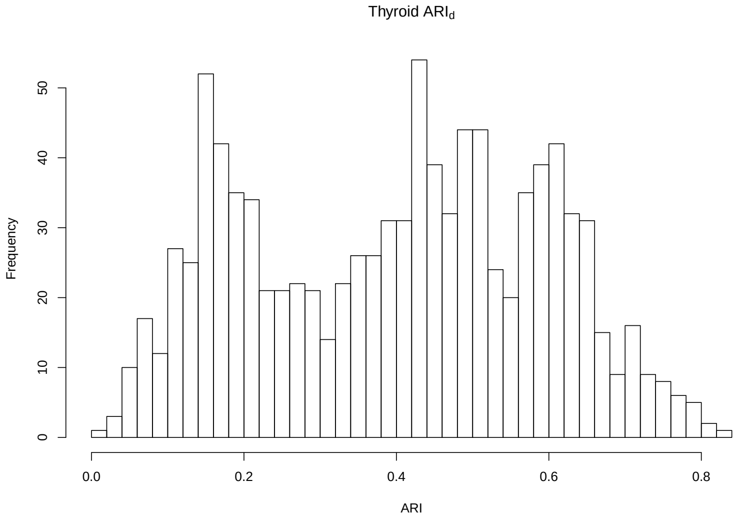
$$(d < p)$$

Normal $\alpha = 2, d = 2$

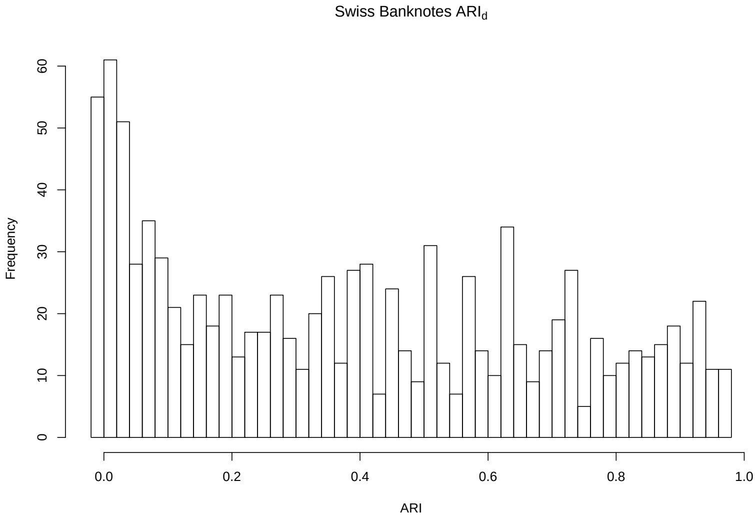
Tabel $\alpha = 2, d = 2$

Dataset	ARI_p	ARI_d	C_e
Thyroid	0.58	0.40	-18
Iris	0.62	0.47	-15
Diabetes	0.38	0.36	-2
Swiss Banknotes	0.85	0.39	-46
Seeds	0.77	0.45	-33
Mice Protein Expression	0.13	0.07	-7
Crabs	0.05	0.04	0

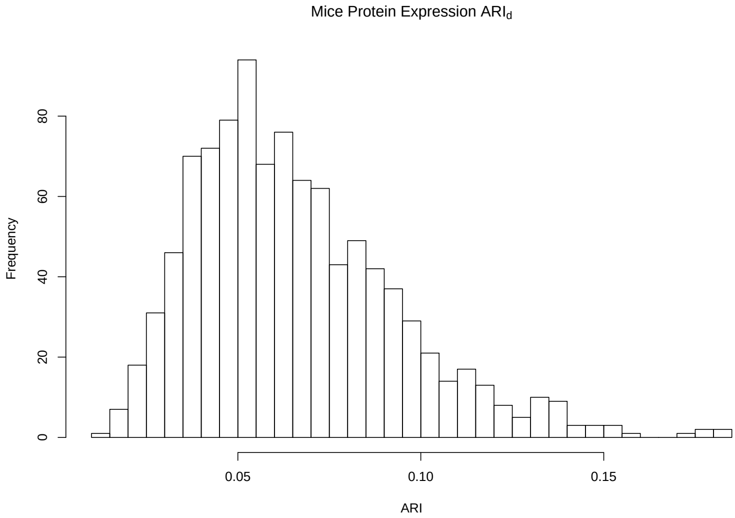
Histogram 2 peak



Hisogram undefined



Hisogram efficient



Normal $\alpha = 2, d = 3$

Tabel $\alpha = 2, d = 3$

Dataset	ARI_p	ARI_d	C_e
Thyroid	0.58	0.43	-15
Iris	0.62	0.54	-8
Diabetes	0.38	0.38	0
Swiss Banknotes	0.85	0.47	-37
Seeds	0.77	0.53	-24
Mice Protein Expression	0.13	0.08	-5
Crabs	0.05	0.05	0

Cauchy $\alpha = 1, d = 2$

Tabel $\alpha = 1, d = 2$

Dataset	ARI_p	ARI_d	C_e
Thyroid	0.58	0.36	-23
Iris	0.62	0.51	-11
Diabetes	0.38	0.33	-5
Swiss Banknotes	0.85	0.40	-44
Seeds	0.77	0.45	-32
Mice Protein Expression	0.13	0.06	-7
Crabs	0.05	0.05	0

Cauchy $\alpha = 1, d = 3$

Tabel $\alpha = 1, d = 3$

Dataset	ARI_p	ARI_d	C_e
Thyroid	0.58	0.37	-22
Iris	0.62	0.54	-8
Diabetes	0.38	0.35	-3
Swiss Banknotes	0.85	0.43	-41
Seeds	0.77	0.47	-30
Mice Protein Expression	0.13	0.07	-7
Crabs	0.05	0.05	0

Sparse $s = 2, d = 2$

Tabel $s = 2, d = 2$

Dataset	ARI_p	ARI_d	C_e
Thyroid	0.58	0.40	-18
Iris	0.62	0.49	-13
Diabetes	0.38	0.35	-3
Swiss Banknotes	0.85	0.40	-44
Seeds	0.77	0.45	-33
Mice Protein Expression	0.13	0.06	-7
Crabs	0.05	0.05	0

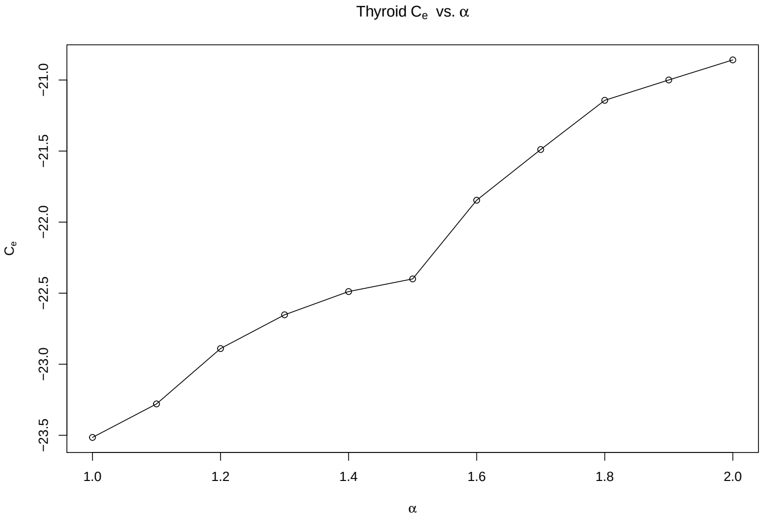
Sparse $s = 2, d = 3$

Tabel $s = 2, d = 3$

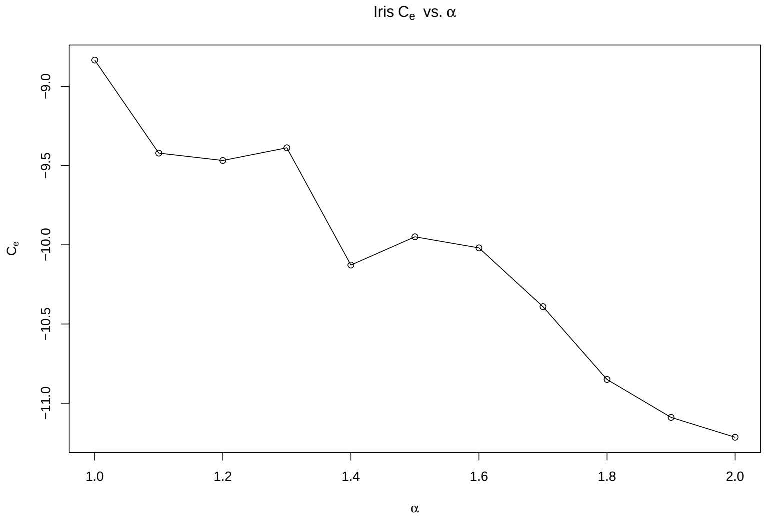
Dataset	ARI_p	ARI_d	C_e
Thyroid	0.58	0.44	-14
Iris	0.62	0.54	-8
Diabetes	0.38	0.37	-1
Swiss Banknotes	0.85	0.49	-36
Seeds	0.77	0.53	-24
Mice Protein Expression	0.13	0.08	-5
Crabs	0.05	0.05	0

C_e versus α for $d = 2$

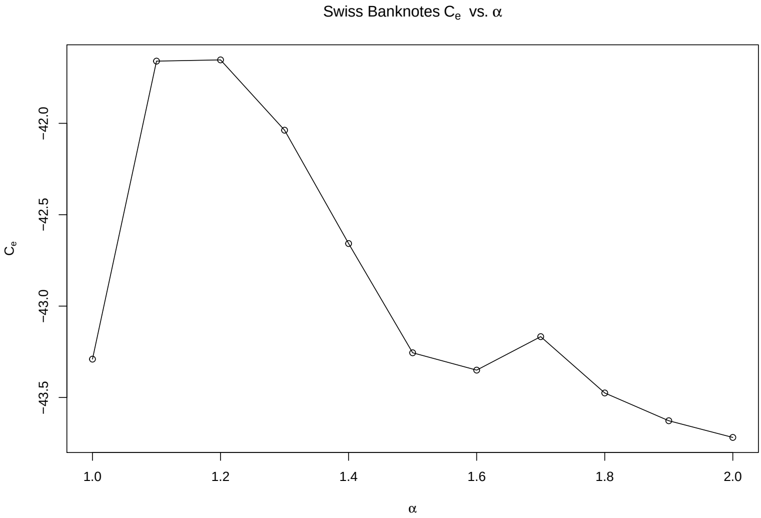
Normal is better



Cauchy is better

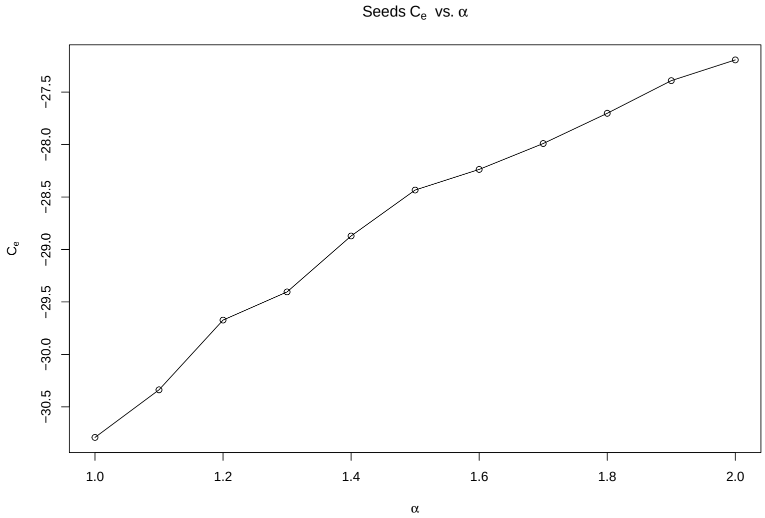


$0 < \alpha < 1$ is better

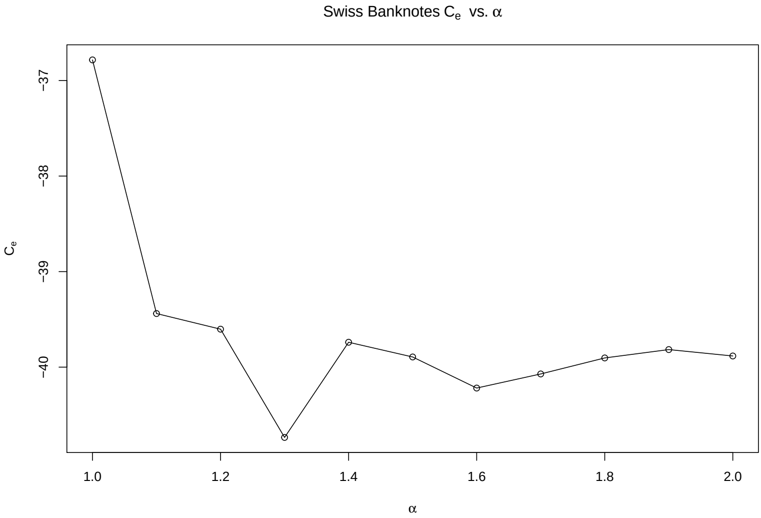


C_e versus α for $d = 3$

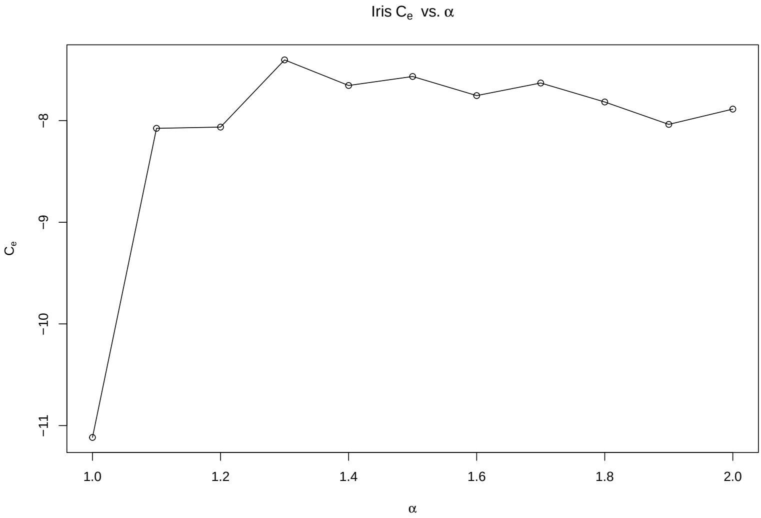
Normal is better



Cauchy is better

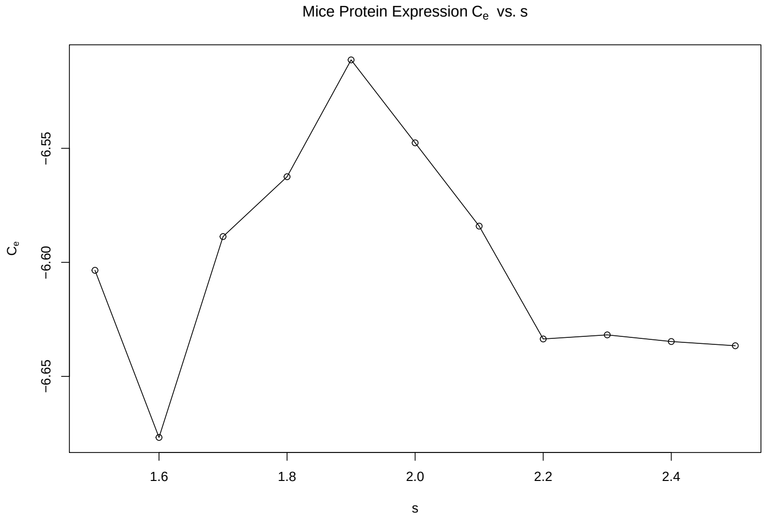


$0 < \alpha < 1$ is better



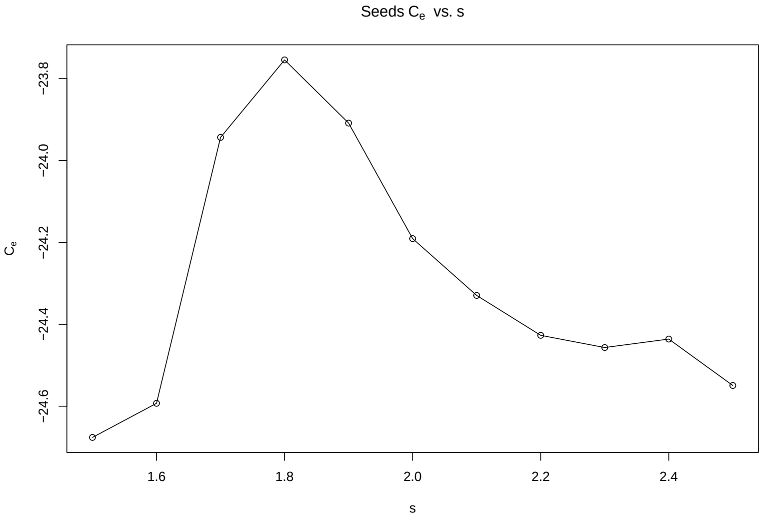
C_e versus s in Sparse for $d = 2$

MPE



C_e versus s in Sparse for $d = 3$

Seeds



Comparision $d = 2$

Comparision $d = 2$

Dataset	$\text{RP}_{\alpha=2}$	$\text{RP}_{\alpha=1}$	$\text{RP}_{s=2}$	Cov.	ρ_s	ρ'	η_p	SCV_2	FSCV_1	SCV_1
Thyroid	-18	-23	-18	-10	35	30	-6	36	37	37
Iris	-15	-11	-13	1	3	3	0	0	0	0
Diabetes	-2	-5	-3	0	22	33	8	4	38	4
Banknotes	-46	-44	-44	0	0	-97	-71	-93	0	-15
Seeds	-33	-32	-33	-14	0	-14	2	2	0	2
MPE	-7	-7	-7	-11	-11	-19	-6	-13	-19	-10
Crabs	0	0	0	2	1	-1	0	-1	1	-2

Figure 9: d = 2

Comparision $d = 3$

Comparision $d = 3$

Dataset	$RP_{\alpha=2}$	$RP_{\alpha=1}$	$RP_{s=2}$	Cov.	ρ_s	ρ'	η_p	SCV_2	$FSCV_1$	SCV_1
Thyroid	-15	-22	-14	-12	3	-6	5	4	2	4
Iris	-8	-8	-8	0	2	1	0	3	-1	3
Banknotes	-37	-41	-36	0	-5	0	-88	0	-24	0
Seeds	-24	-30	-24	1	0	-1	0	-26	-15	-15
MPE	-5	-7	-5	-9	-8	-12	-4	-8	-7	-8
Crabs	0	0	0	-1	0	1	0	1	-1	2

Figure 10: d = 3

???

Similarity measures

تعریف ۷

X بردار تصادفی پایدار با پارامترهای α و اندازه‌ی طیفی Γ،
معیار وابستگی η_p برای $i, j = ۱, \ldots, d \quad (X_i, X_j)$:

$$\eta_p = \eta_p\left(X_i, X_j\right) = \| \gamma^{\alpha}\left(u_i, u_j\right) - \gamma_{\perp}^{\alpha}\left(u_i, u_j\right) \|_{L_p, du} \qquad (۳)$$

$\gamma_{\perp}$ تابع مقیاس تصویر توزیع پایدار دو متغیره با مولفه‌های مستقل

Figure 11:

Similarity measures

تعریف ۸

(X,Y) بردار تصادفی پایدار با پارامتر اندازه‌ی طیفی Γ_{XY}،

$$\rho_s(X,Y) = \left(\int_{\mathbb{S}^{2}} \left(\Gamma_{XY}(u) - \Gamma_{\perp}(u)\right)^{2} \mathrm{d}u\right)^{1/2} \qquad (6)$$

که در آن $\Gamma_{\perp}$ اندازه طیفی بردار پایدار با متغیرهای مستقل

Figure 12:

Similarity measures

معیار وابستگی هم‌پراکنشی متقارن

$$SCV_{۱} = \frac{[X_{۱}, X_{۲}]_{\alpha} + [X_{۲}, X_{۱}]_{\alpha}}{۲}.$$

Figure 13:

Similarity measures

معیار وابستگی هم‌پراکنشی متقارن

$$SCV_{۲}(X_i, X_j) = \kappa_\alpha(X_i, X_j)\left|[X_i, X_j]\,[X_j, X_i]\right|^{\frac{۱}{۲}} \quad i, j = ۱, \ldots, p,$$

که درآن

$$\kappa_\alpha(X_i, X_j) = \begin{cases} sign([X_i, X_j]_\alpha), & sign([X_i, X_j]_\alpha) = sign([X_j, X_i]_\alpha), \\ -۱, & sign([X_i, X_j]_\alpha) = -sign([X_j, X_i]_\alpha). \end{cases}$$

Figure 14: