Problem Description

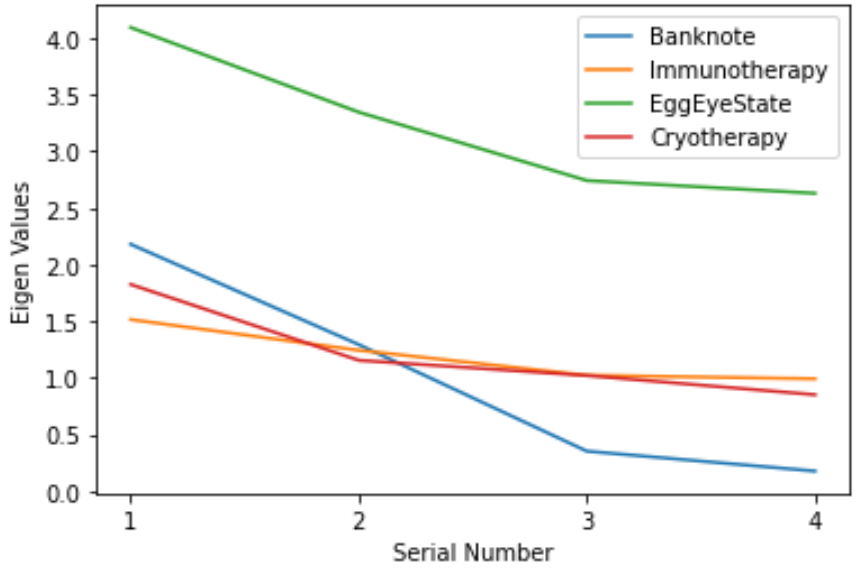
Dimensionality Reduction is one of the major sub problems from Data Mining's 6 historical problems. Principal Component Analysis is one of the mathematical tools which is used widely to project high dimensional data into lower dimensional data. There is no validated choice of selecting the number of principle components to represent the original data.

Heuristics to select number of principal components

There are three main heuristics to select number of informative (useful, meaningful, etc.) component

- Kaiser
- Broken Stick
- Conditional Number

Databases used for study



Databases	Cases	Attributes	PCA-K	PCA-BS	PCA-CN
Banknote	1372	4	2	2	3
blood	748	4	2	2	3
breastCancer	569	30	6	3	5
climate	540	18	10	1	18
Cryotherapy	90	6	3	1	6

We used 3 classifiers for performance measurement:

- KNN
- Logistic Regression
- Fisher Discriminant Analysis

For selecting K for each dataset, we used leave one out cross validation technique.

Databases	Best K
Banknote	17
blood	23
breastCancer	5
climate	5
Cryotherapy	7

For logistic Regression, we use threshold selection technique to select the best decision boundary during prediction at each fold and each iteration.

Steps:

- Define a set of unique values with all possible points.
- Add all borders
- Select threshold which gives the least error rate

Performance Metrics:

- Balanced Accuracy
- Balanced accuracy was calculated for 10 times repeated 10 fold cross validation.

Ranking protocols

We used 2 ranking protocols to rank each dimensionality reduction method.

- Average Ranking
- Ranking based on test of statistical significance of differences of performances

For Example

For Fisher,

						Ranking based on average				Ranking based on p value		
Databases	BA	BA-kaiser	BA-brok	BA-cond	Full dim	Kaiser	Broken	Condition	Full dim	Kaiser	Broken	Condition
Banknote.mat	0.9877	0.7848	0.7848	0.9137	1	3.5	3.5	2	1	-1	-1	0
blood.mat	0.5675	0.6017	0.6017	0.5675	3.5	1.5	1.5	3.5	-1	1	1	-1
breastCancer.mat	0.9587	0.9507	0.9293	0.9501	1	2	4	3	1	0	-1	0
climate.mat	0.8025	0.6354	0.499	0.8025	1.5	3	4	1.5	1	0	-1	1
Cryotherapy.mat	0.9025	0.785	0.805	0.9025	1.5	4	3	1.5	1	-1	0	1

For Example

For KNN,

							Ranking based on average				Ranking based on p value			
Databases	Best K	BA	BA-kaiser	BA-brok	BA-cond		Full dim	Kaiser	Broken	Condition	Full dim	Kaiser	Broken	Condition
Banknote.mat	17	0.9967	0.859	0.859	0.9618		1	3.5	3.5	2	1	-1	-1	0
blood.mat	23	0.5953	0.5689	0.5689	0.5953		1.5	3.5	3.5	1.5	0	0	0	0
breastCancer.mat	5	0.9589	0.9584	0.9293	0.9466		1	2	4	3	1	1	-1	0
climate.mat	5	0.637	0.517	0.499	0.637		1.5	3	4	1.5	1	0	-1	1
Cryotherapy.mat	7	0.925	0.8075	0.8075	0.925		1.5	3.5	3.5	1.5	1	-1	-1	1

For Example

For Logistic,

						Ranking based on average				Ranking based on p value			
Databases	BA	BA-kaiser	BA-brok	BA-cond		Full dim	Kaiser	Broken	Condition		Full dim	Kaiser	Broken	Condition
Banknote.mat	0.9884	0.7725	0.7725	0.9175		1	3.5	3.5	2		1	-1	-1	0
blood.mat	0.5642	0.6036	0.6036	0.5642		3.5	1.5	1.5	3.5		-1	1	1	-1
breastCancer.mat	0.9744	0.9684	0.944	0.9728		1	3	4	2		1	0	-1	1
climate.mat	0.8334	0.6355	0.499	0.8334		1.5	3	4	1.5		1	0	-1	1
Cryotherapy.mat	0.8875	0.81	0.805	0.8875		1.5	3	4	1.5		1	-1	-1	1

For p-value based ranking

- 1 means statistically insignificantly different from the best
- 0 means statistically significantly different from the best and the worst
- -1 means statistically insignificantly different from the worst

Conclusion

- There is no one universal answer which of intrinsic dimension is better.
- In average, dimensionality reduction on base of conditional number is better.