# Supplementary
# FORKS: Finding Orderings Robustly using K-means and Steiner trees

Mayank Sharma, Huipeng Li, Debarka Sengupta, Shyam Prabhakar & Jayadeva

## 1 Supplementary

### 1.1 Correlation and run times table

| Datasets | Algorithms (mean correlation ± standard deviation) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | FORKS | kmeans_R | monocle2 | scuba | TSCAN | waterfall | dpt | GPfates | SLICER |
| Arabidopsis | 0.9256±0.027 | 0.8976±0.0422 | 0.8582±0.0389 | 0.7743±0.0422 | 0.8976±0.0422 | 0.7786±0.0157 | 0.7694±0.0301 | 0.3622±0.2902 | 0.922±0.051 |
| Deng_2014 | 0.915±0.0078 | 0.8608±0.0502 | 0.9057±0.0047 | 0.5134±0.4003 | 0.9335±0.0113 | 0.5183±0.3674 | 0.4954±0.0439 | 0.0473±0.0324 | -0.2398±0.1076 |
| Guo_2010 | 0.8755±0.0639 | 0.2921±0.0601 | 0.5706±0.2874 | 0.2474±0.0744 | 0.3046±0.0399 | 0.4035±0.2171 | 0.8162±0.0124 | 0.0421±0.0275 | 0.322±0.3181 |
| Klein | 0.9242±0.0012 | 0.93±0.009 | * | 0.4433±0.0045 | 0.6876±0.1944 | 0.4433±0.0045 | 0.8445±0.0044 | 0.0122±0.0092 | 0.7019±0.0777 |
| LPS | 0.793±0.0298 | 0.7678±0.0299 | 0.3343±0.0874 | 0.763±0.0217 | 0.7266±0.184 | 0.7852±0.0423 | 0.404±0.06 | 0.3835±0.3726 | 0.1307±0.2733 |
| Preimplant | 0.9032±0.0043 | 0.1364±0.0438 | * | 0.1543±0.0327 | 0.1394±0.0804 | 0.1432±0.0448 | 0.7309±0.0272 | 0.2256±0.2458 | 0.0237±0.0279 |

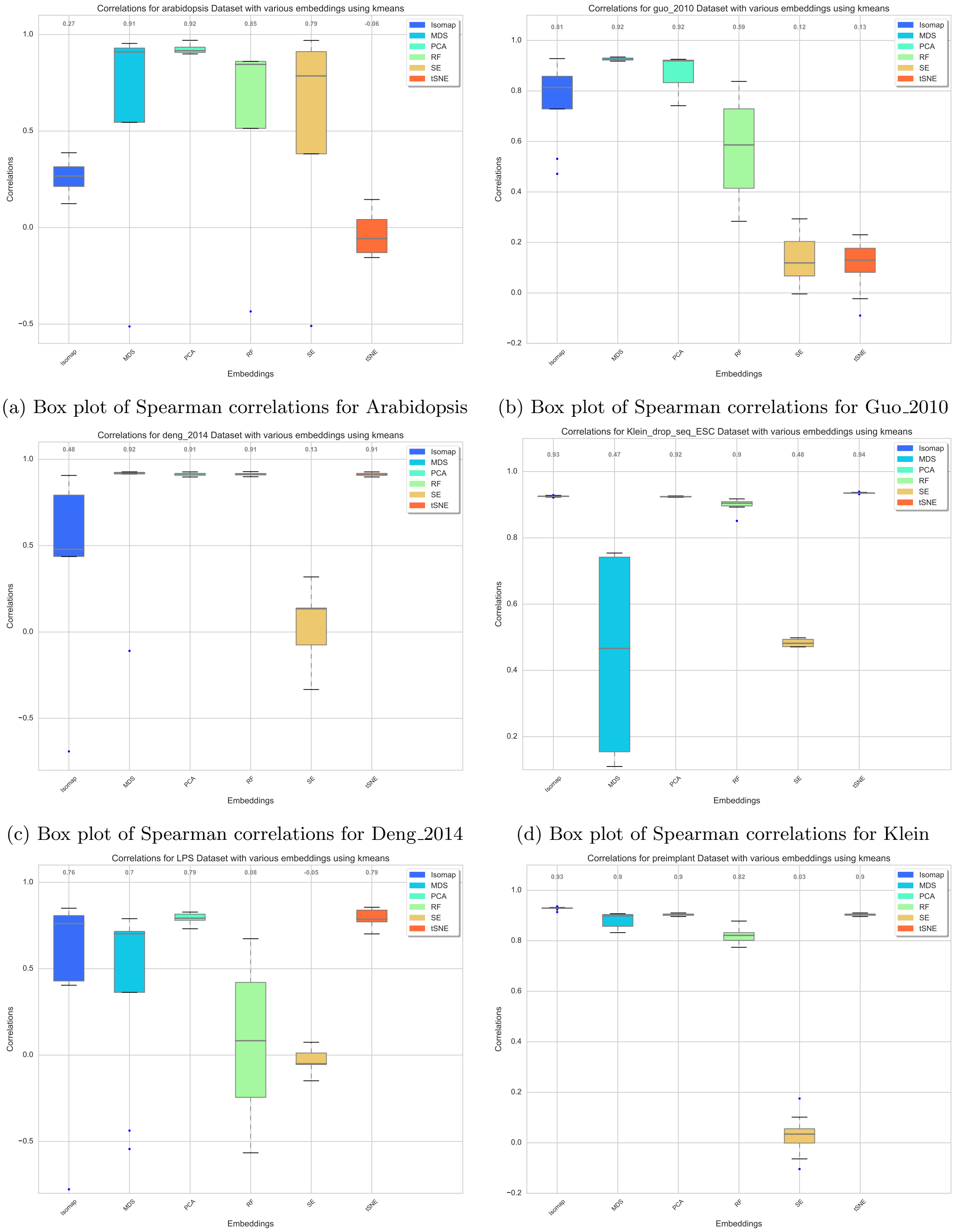
Table 1: Comparison of various algorithms showing their mean Spearman correlation with the known cell time and standard deviation

| Datasets | Algorithms (mean Runtime ± standard deviation) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | FORKS | kmeans_R | monocle2 | scuba | TSCAN | waterfall | dpt | GPfates | SLICER |
| Arabidopsis | 2.4527±0.4644 | 0.1225±0.0657 | 1.85±0.2662 | 0.0025±0.0043 | 0.0875±0.0164 | 0.295±0.3269 | 0.1834±0.2138 | 1.8656±0.8919 | 7.2525±1.6308 |
| Deng_2014 | 1.2617±0.4897 | 1.242±0.0646 | 7.578±1.6199 | 0.346±0.1122 | 1.164±0.0952 | 0.622±0.0796 | 0.2408±0.1312 | 25.8797±9.9661 | 115.412±7.5778 |
| Guo_2010 | 1.0703±0.1788 | 0.3162±0.0187 | 13.03±2.2249 | 0.925±0.7382 | 0.4913±0.0881 | 1.415±0.1287 | 0.4585±0.1024 | 0.7512±0.2665 | 38.76±1.9805 |
| Klein | 0.7114±0.1362 | 4.877±0.5638 | * | 32.511±3.8158 | 21.108±2.4441 | 6.368±0.7165 | 21.4364±0.8247 | 0.9728±0.0923 | 2482.849±307.1986 |
| LPS | 0.9429±0.3388 | 4.44±0.2023 | 248.1425±83.8347 | 3.405±1.9445 | 4.8487±0.6232 | 0.61±0.1127 | 0.2665±0.1386 | 3.4821±1.8938 | 234.228±16.3452 |
| Preimplant | 6.8219±5.0221 | 107.988±5.4183 | * | 13.212±3.1936 | 111.783±5.7511 | 2.886±0.2543 | 4.4126±0.7291 | 50.7831±42.3228 | 6184.737±223.9307 |

Table 2: Comparison of various algorithms showing their mean run time (s) and standard deviation

[1]

[1] * ⟹ Algorithm failed to run on the dataset



(a) Box plot of Spearman correlations for Arabidopsis

(b) Box plot of Spearman correlations for Guo_2010

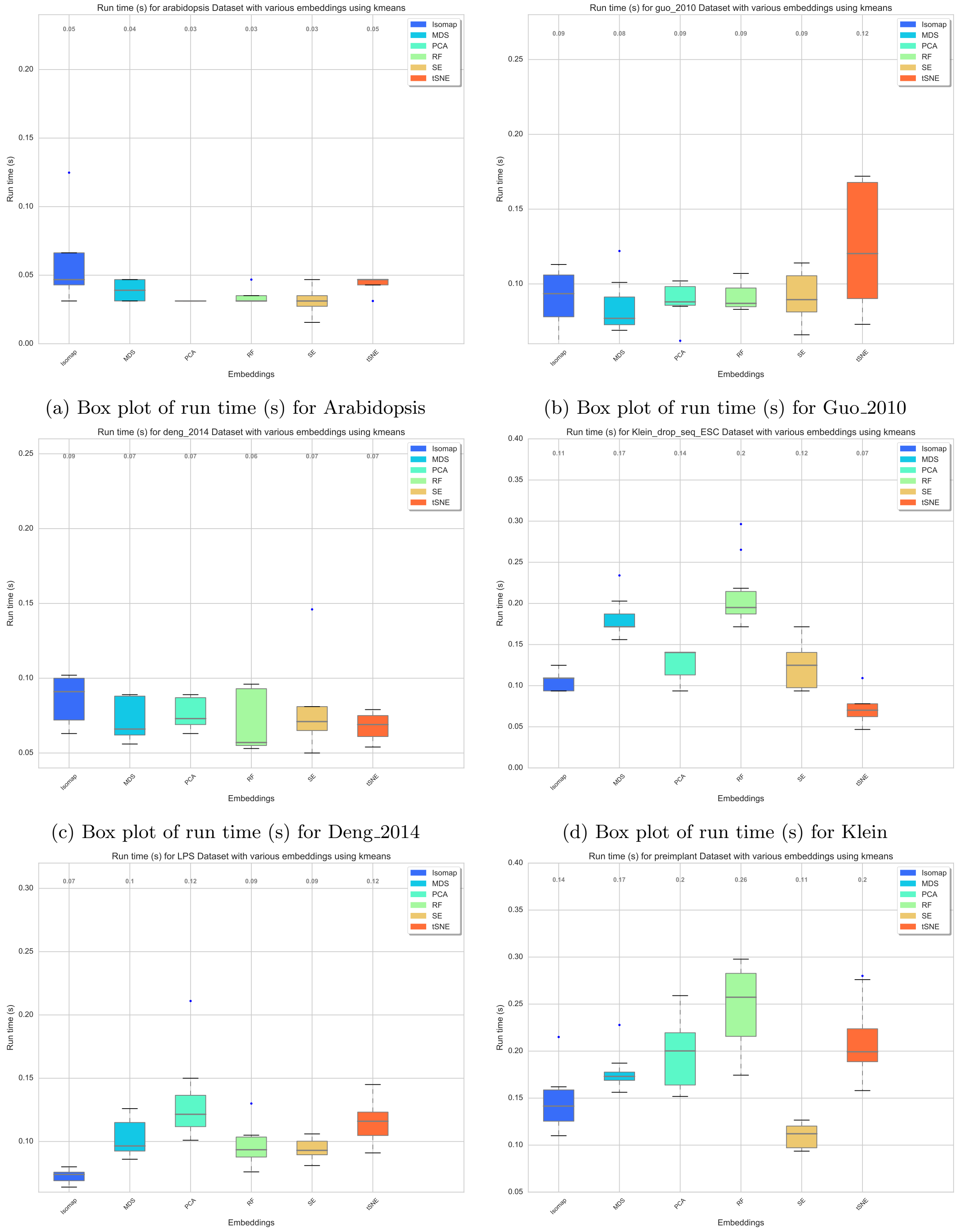
(c) Box plot of Spearman correlations for Deng_2014

(d) Box plot of Spearman correlations for Klein

(e) Box plot of Spearman correlations for LPS

(f) Box plot of Spearman correlations for Preimplant

Figure 1: Box plots showing the correlations with given cell times using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use k-means to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of run time (s) for Arabidopsis

(b) Box plot of run time (s) for Guo_2010

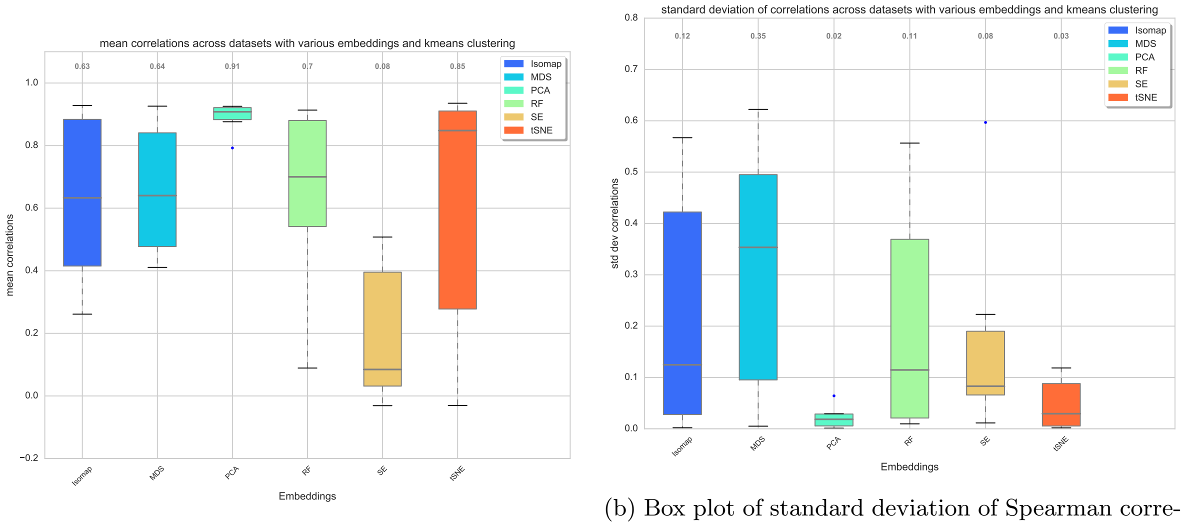
(c) Box plot of run time (s) for Deng_2014

(d) Box plot of run time (s) for Klein

(e) Box plot of run time (s) for LPS

(f) Box plot of Spearman correlations for Preimplant

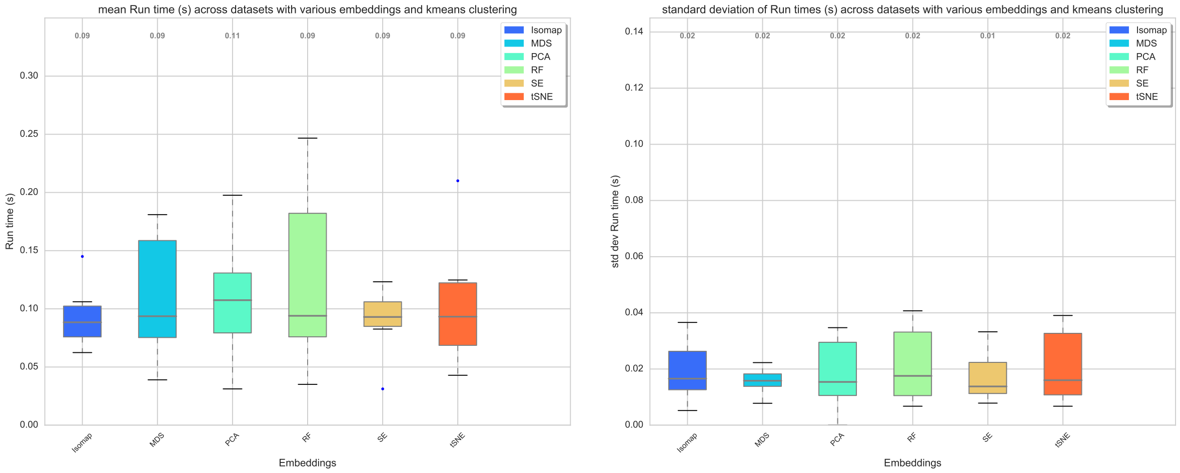
Figure 2: Box plots showing the run times (s) using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use k-means to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of mean of Spearman correlations

(b) Box plot of standard deviation of Spearman correlations

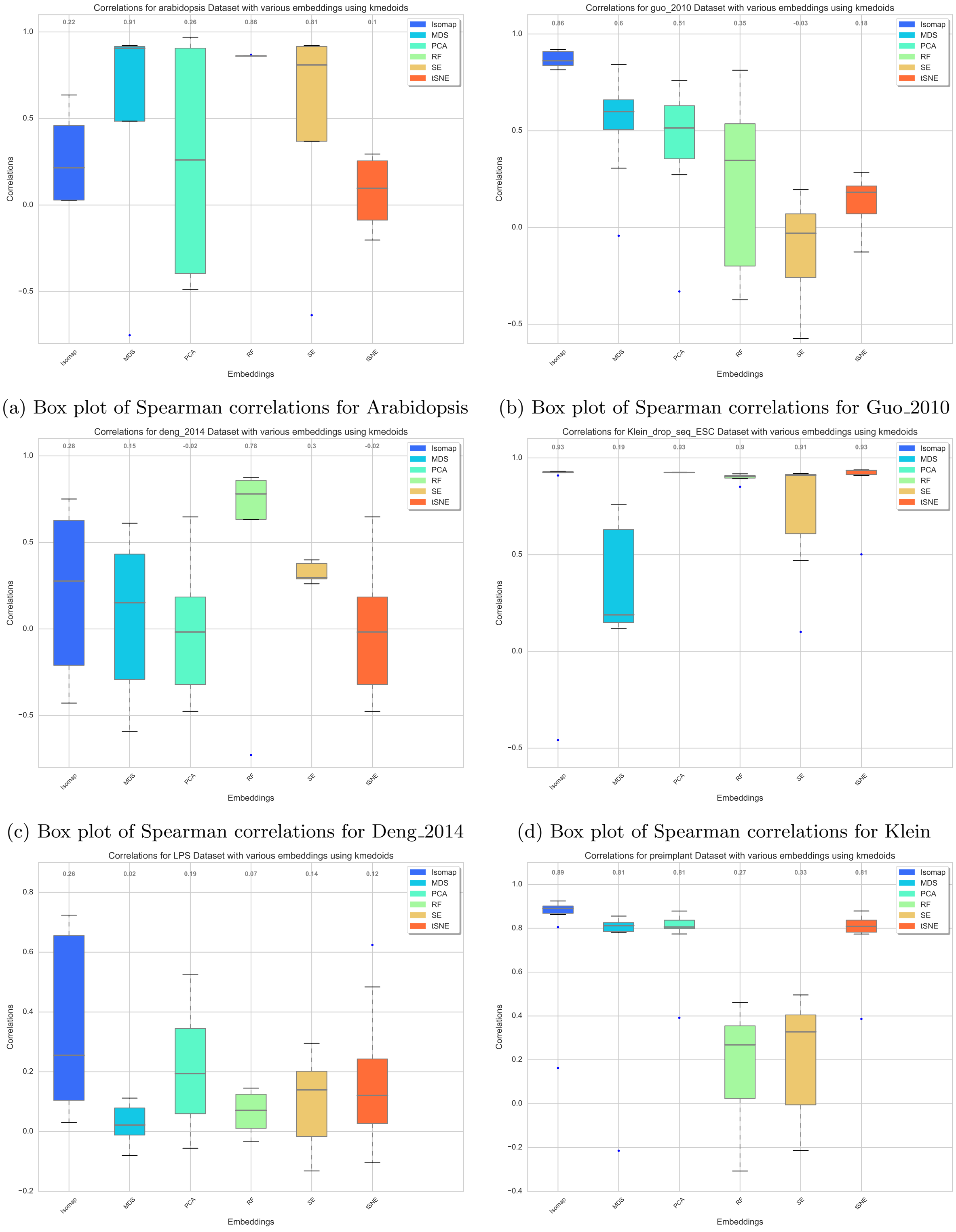
Figure 3: Box plots of means and standard deviation of Spearman correlations among all the datasets shows highly accurate and robust behavior of PCA embedding to change in folds and datasets. The clustering used here is k-means. Values at the top of each figures are the median values.



(a) Box plot of mean of run times (s)

(b) Box plot of standard deviation of run times (s)

Figure 4: Box plots of means and standard deviation of run times among all the datasets using k-means clustering. Values at the top of each figures are the median values.



(a) Box plot of Spearman correlations for Arabidopsis

(b) Box plot of Spearman correlations for Guo_2010

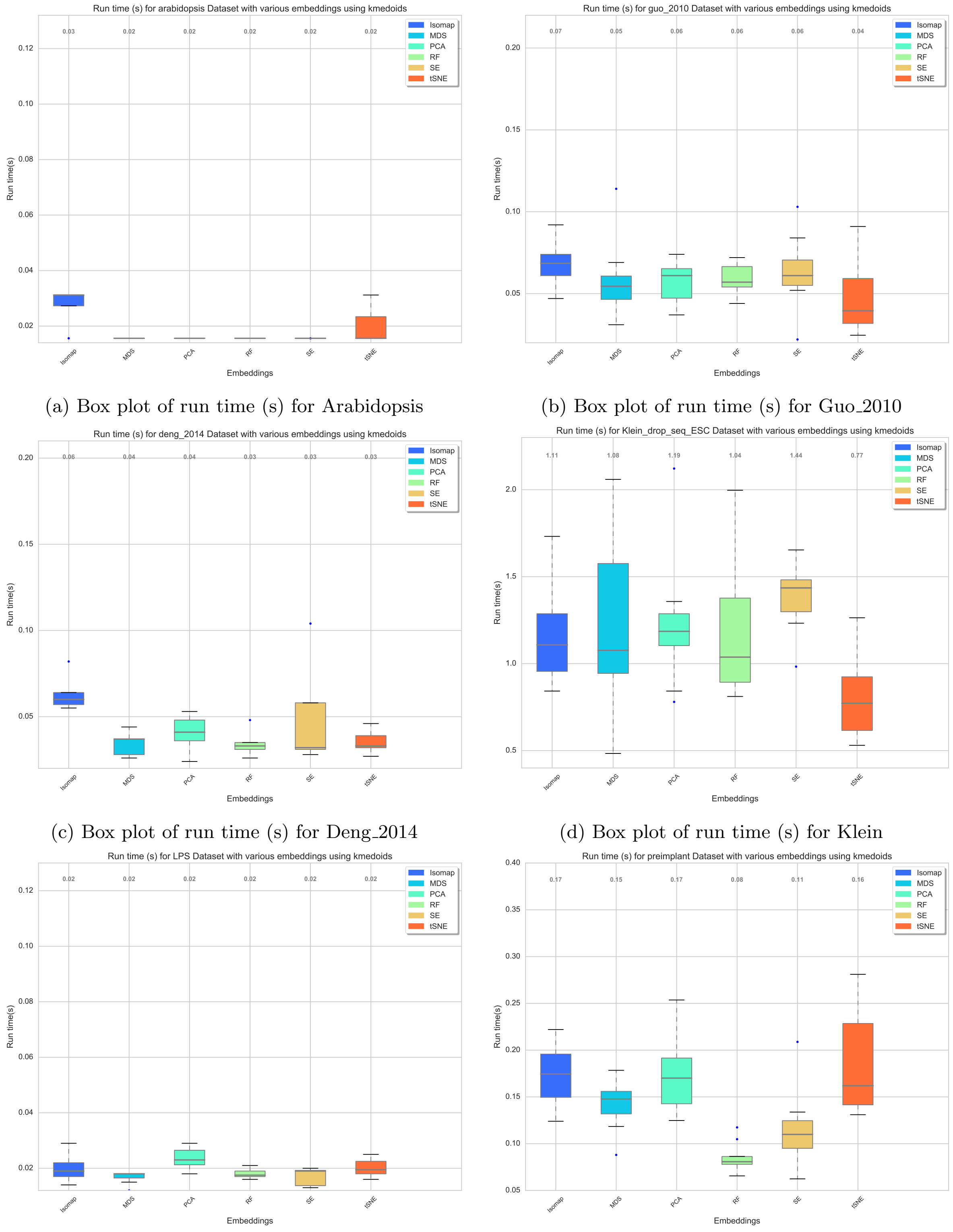
(c) Box plot of Spearman correlations for Deng_2014

(d) Box plot of Spearman correlations for Klein

(e) Box plot of Spearman correlations for LPS

(f) Box plot of Spearman correlations for Preimplant

Figure 5: Box plots showing the correlations with given cell times using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use k-medoids to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of run time (s) for Arabidopsis

(b) Box plot of run time (s) for Guo_2010

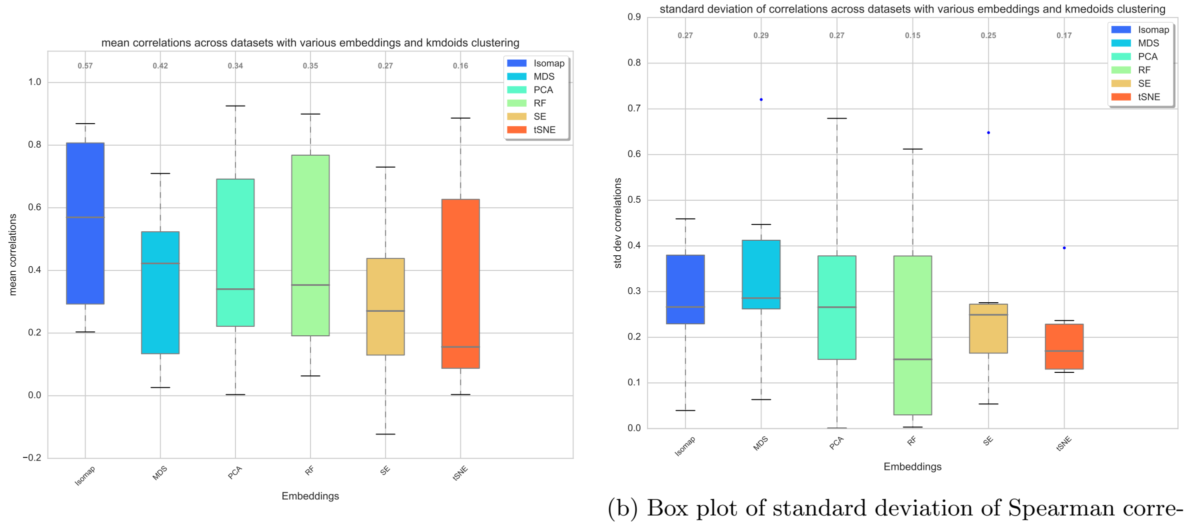
(c) Box plot of run time (s) for Deng_2014

(d) Box plot of run time (s) for Klein

(e) Box plot of run time (s) for LPS

(f) Box plot of Spearman correlations for Preimplant

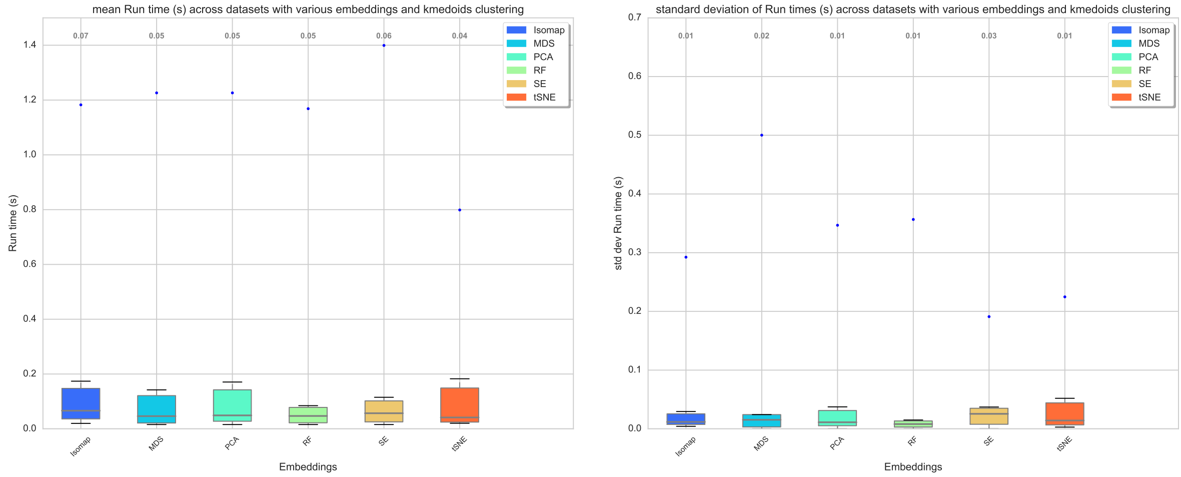
Figure 6: Box plots showing the run times (s) using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use k-medoids to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of mean of Spearman correlations

(b) Box plot of standard deviation of Spearman correlations

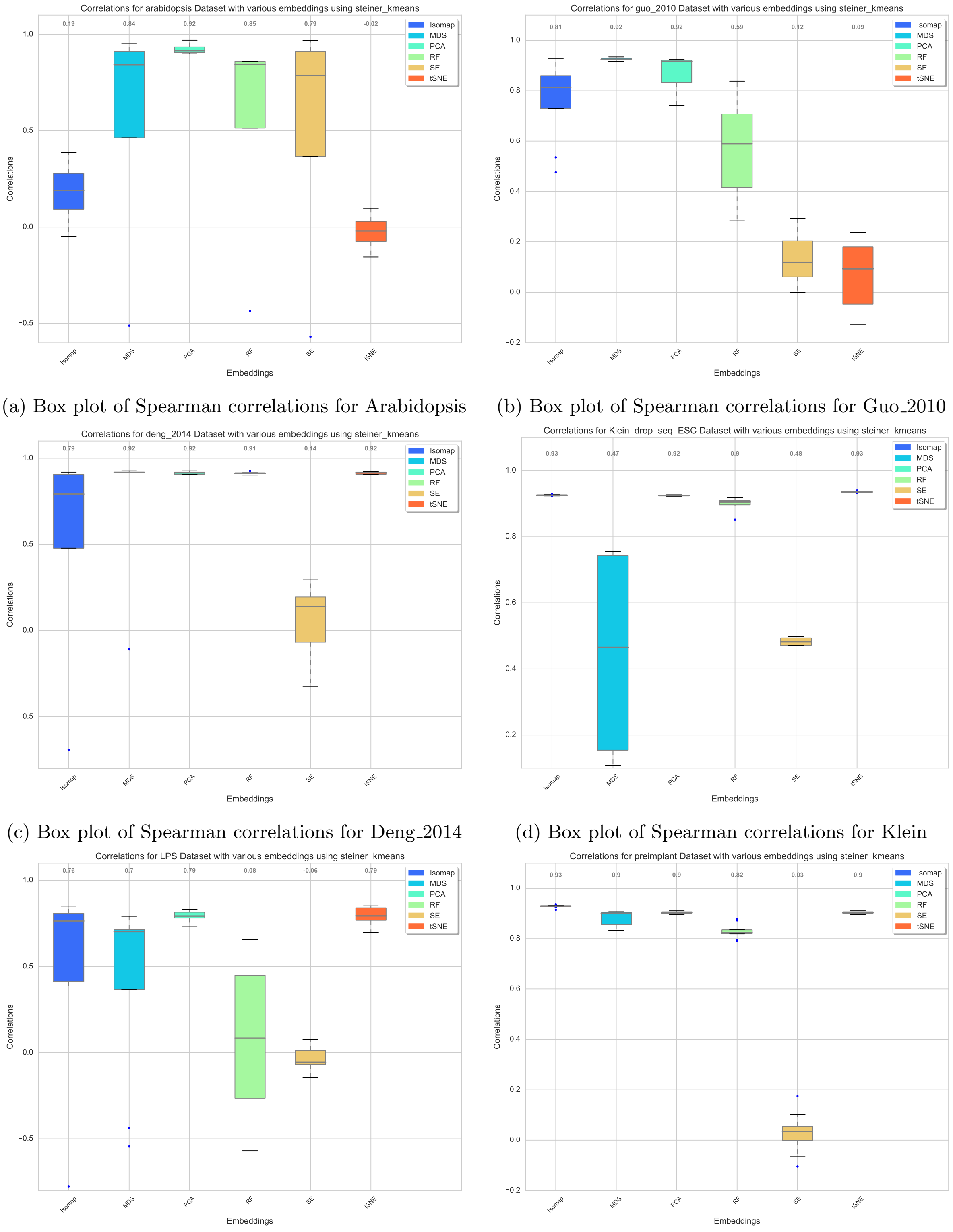
Figure 7: Box plots of means and standard deviation of Spearman correlations among all the datasets shows highly accurate and robust behavior of PCA embedding to change in folds and datasets. The clustering used here is k-medoids. Values at the top of each figures are the median values.



(a) Box plot of mean of run times (s)

(b) Box plot of standard deviation of run times (s)

Figure 8: Box plots of means and standard deviation of run times among all the datasets using k-medoids clustering. Values at the top of each figures are the median values.



(a) Box plot of Spearman correlations for Arabidopsis

(b) Box plot of Spearman correlations for Guo_2010

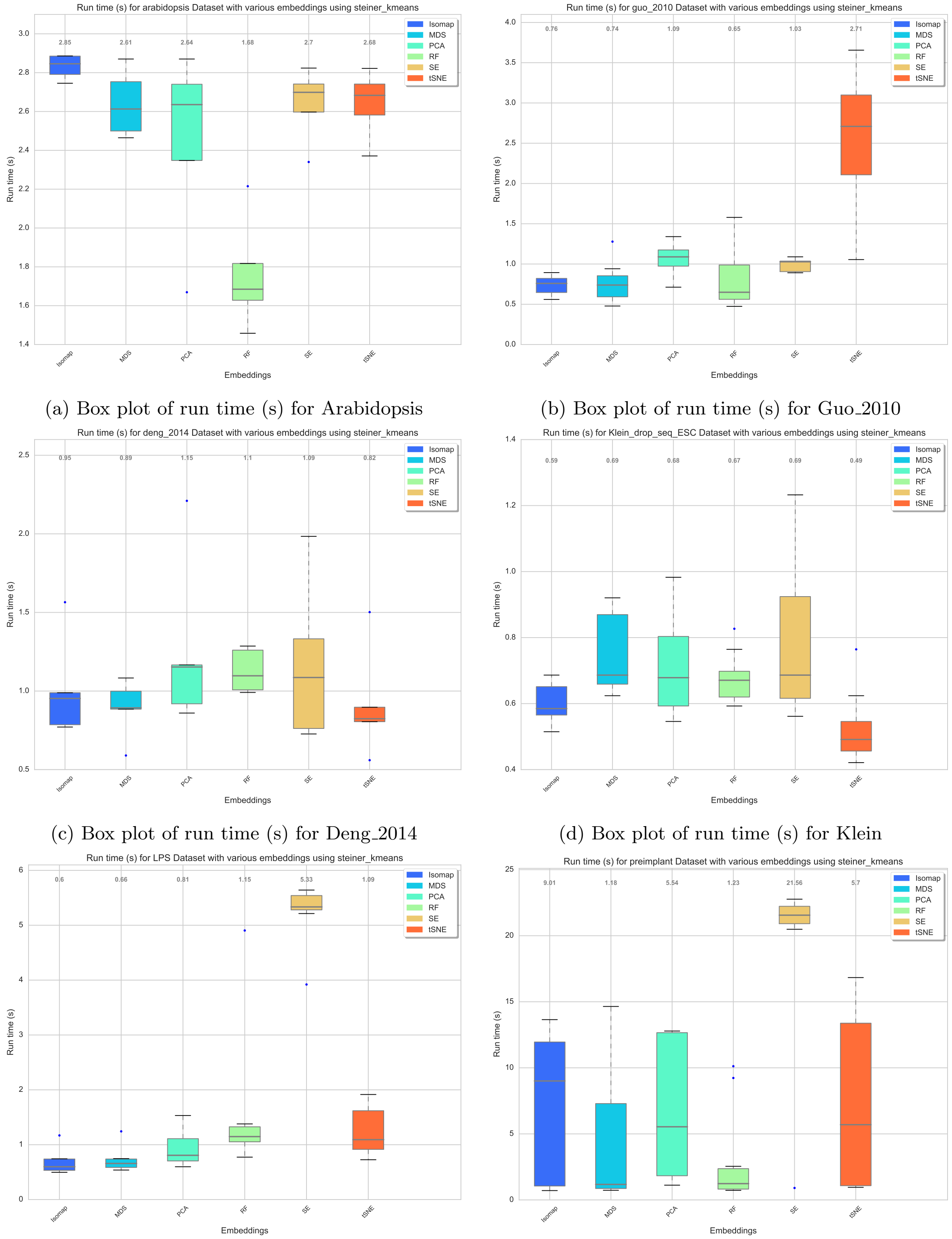
(c) Box plot of Spearman correlations for Deng_2014

(d) Box plot of Spearman correlations for Klein

(e) Box plot of Spearman correlations for LPS

(f) Box plot of Spearman correlations for Preimplant

Figure 9: Box plots showing the correlations with given cell times using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use steiner k-means to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of run time (s) for Arabidopsis

(b) Box plot of run time (s) for Guo_2010

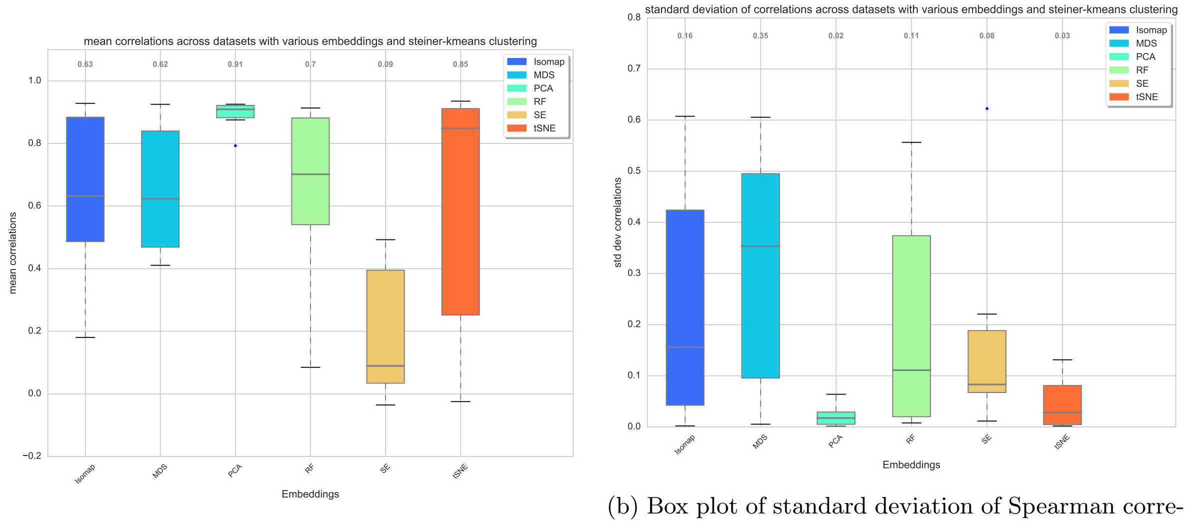
(c) Box plot of run time (s) for Deng_2014

(d) Box plot of run time (s) for Klein

(e) Box plot of run time (s) for LPS

(f) Box plot of Spearman correlations for Preimplant

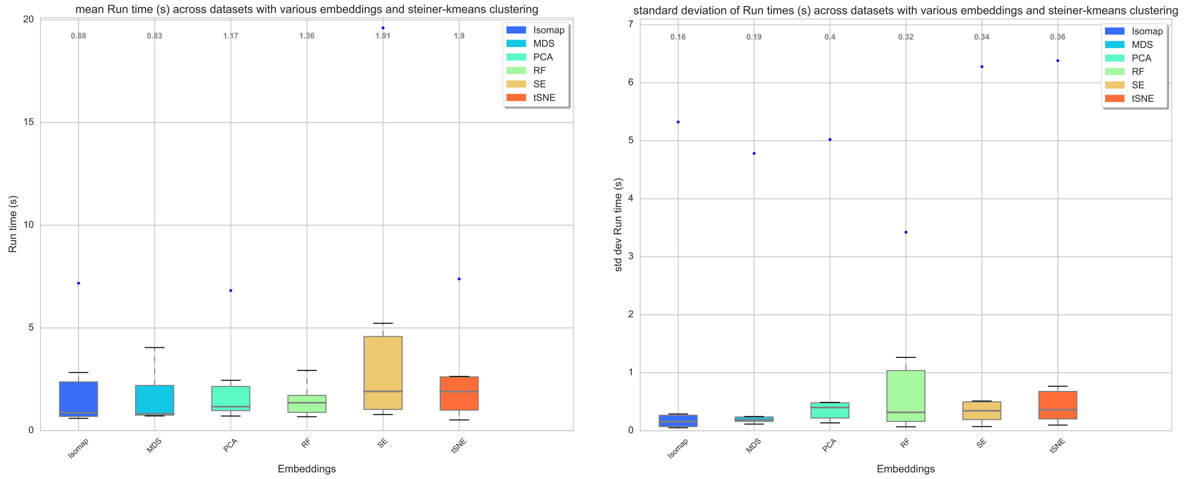
Figure 10: Box plots showing the run times (s) using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use steiner k-means to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of mean of Spearman correlations

(b) Box plot of standard deviation of Spearman correlations

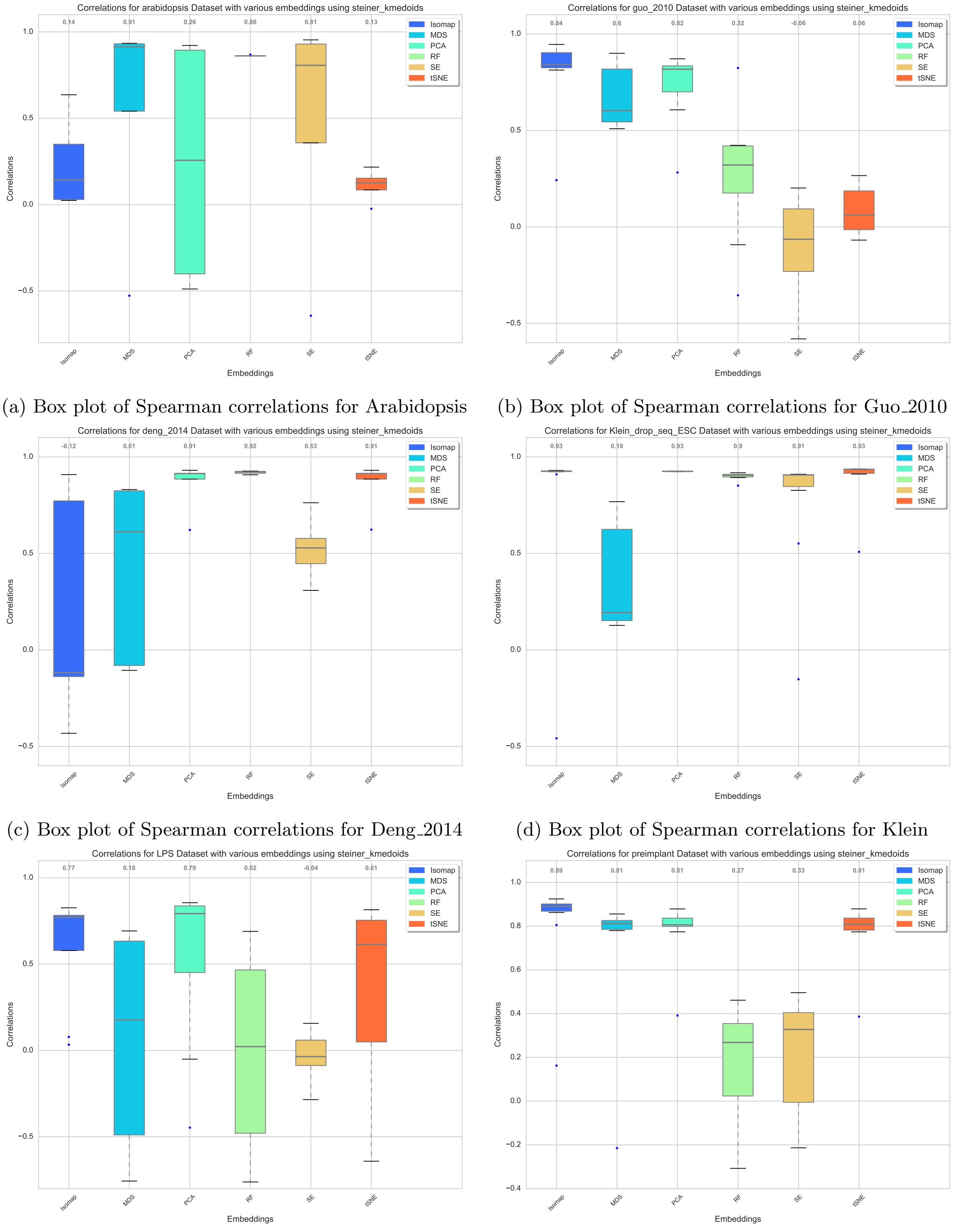
Figure 11: Box plots of means and standard deviation of Spearman correlations among all the datasets shows highly accurate and robust behavior of PCA embedding to change in folds and datasets. The clustering used here is steiner k-means. Values at the top of each figures are the median values.



(a) Box plot of mean of run times (s)

(b) Box plot of standard deviation of run times (s)

Figure 12: Box plots of means and standard deviation of run times among all the datasets using steiner k-means clustering. Values at the top of each figures are the median values.



(a) Box plot of Spearman correlations for Arabidopsis

(b) Box plot of Spearman correlations for Guo_2010

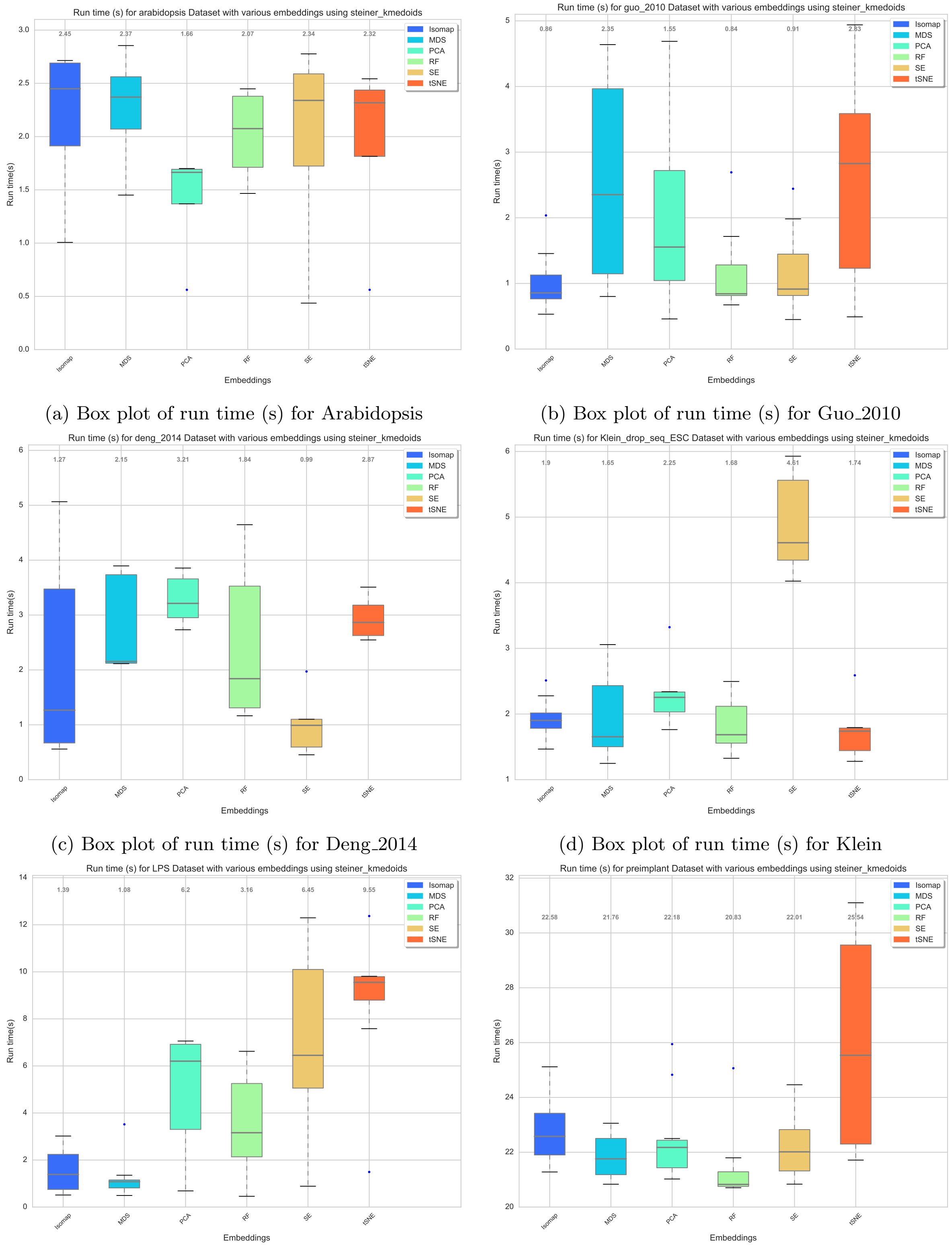
(c) Box plot of Spearman correlations for Deng_2014

(d) Box plot of Spearman correlations for Klein

(e) Box plot of Spearman correlations for LPS

(f) Box plot of Spearman correlations for Preimplant

Figure 13: Box plots showing the correlations with given cell times using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use steiner k-medoids to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of run time (s) for Arabidopsis

(b) Box plot of run time (s) for Guo_2010

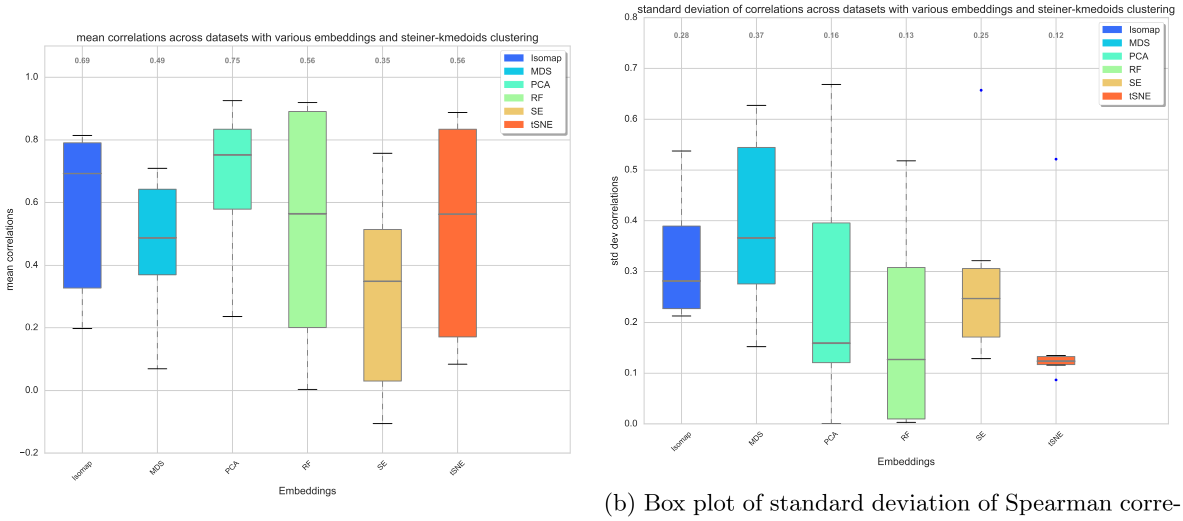
(c) Box plot of run time (s) for Deng_2014

(d) Box plot of run time (s) for Klein

(e) Box plot of run time (s) for LPS

(f) Box plot of Spearman correlations for Preimplant

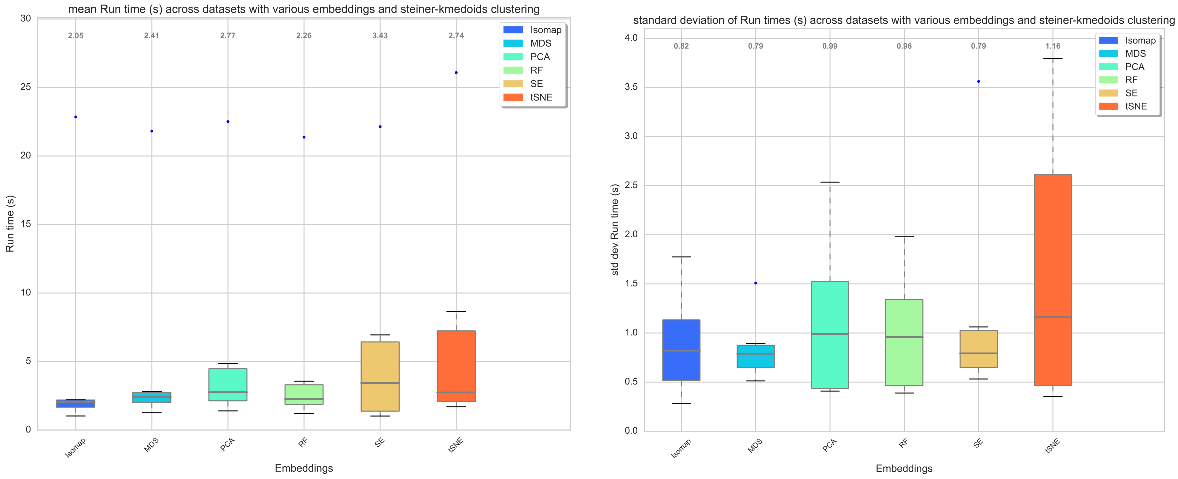
Figure 14: Box plots showing the run times (s) using various embeddings for Arabidopsis (nfolds=4), Guo_2010 (nfolds=8), Deng_2014 (nfolds=5), Klein (nfolds=10), LPS (nfolds=8) and Preimplant (nfolds=10) datasets, the legend shows the various embeddings being compared. We use steiner k-medoids to find the cluster centers. Values at the top of each figures are the median values.



(a) Box plot of mean of Spearman correlations

(b) Box plot of standard deviation of Spearman correlations

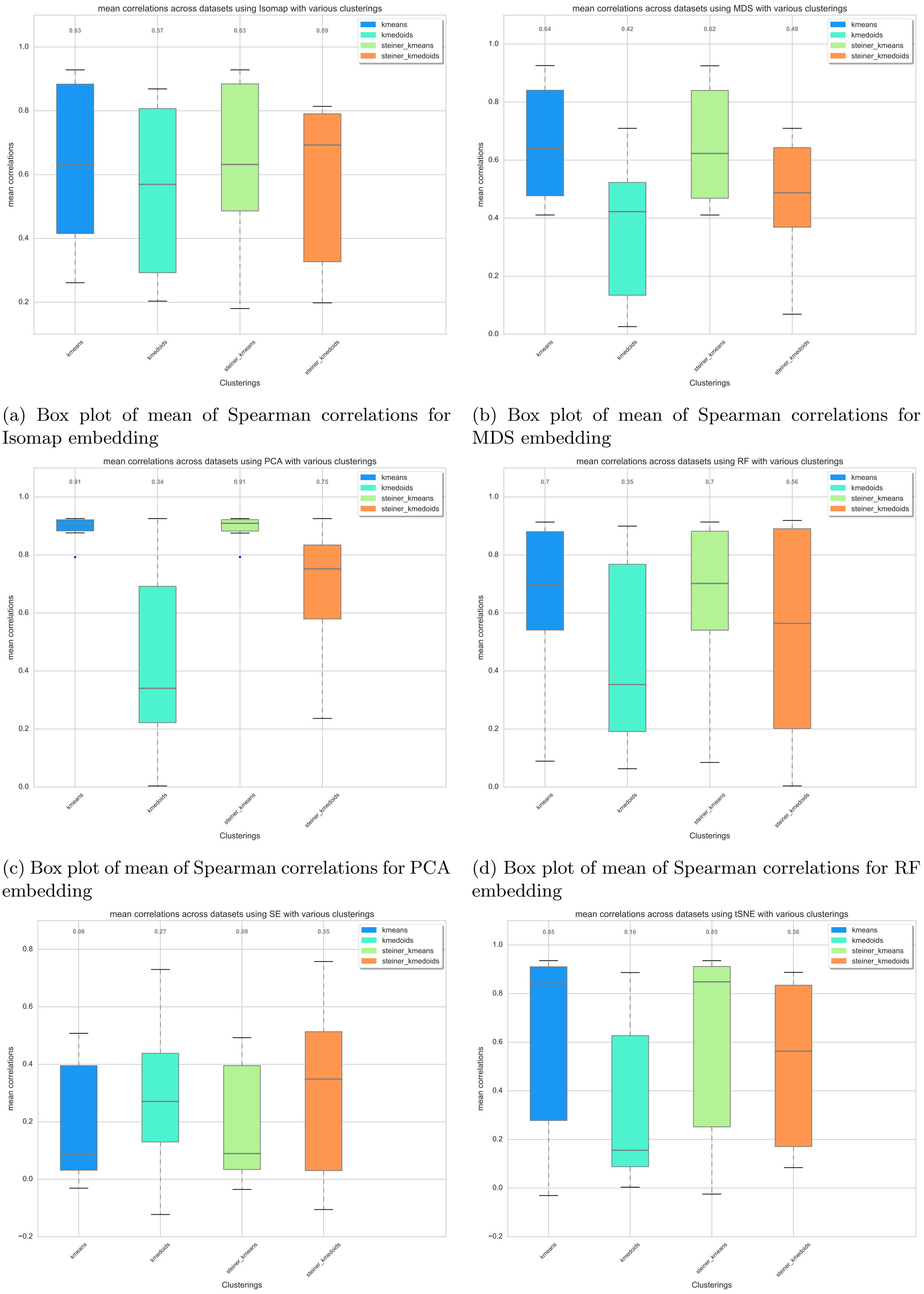
Figure 15: Box plots of means and standard deviation of Spearman correlations among all the datasets shows highly accurate and robust behavior of PCA embedding to change in folds and datasets. The clustering used here is steiner k-medoids. Values at the top of each figures are the median values.



(a) Box plot of mean of run times (s)

(b) Box plot of standard deviation of run times (s)

Figure 16: Box plots of means and standard deviation of run times among all the datasets using steiner k-medoids clustering. Values at the top of each figures are the median values.



(a) Box plot of mean of Spearman correlations for Isomap embedding

(b) Box plot of mean of Spearman correlations for MDS embedding

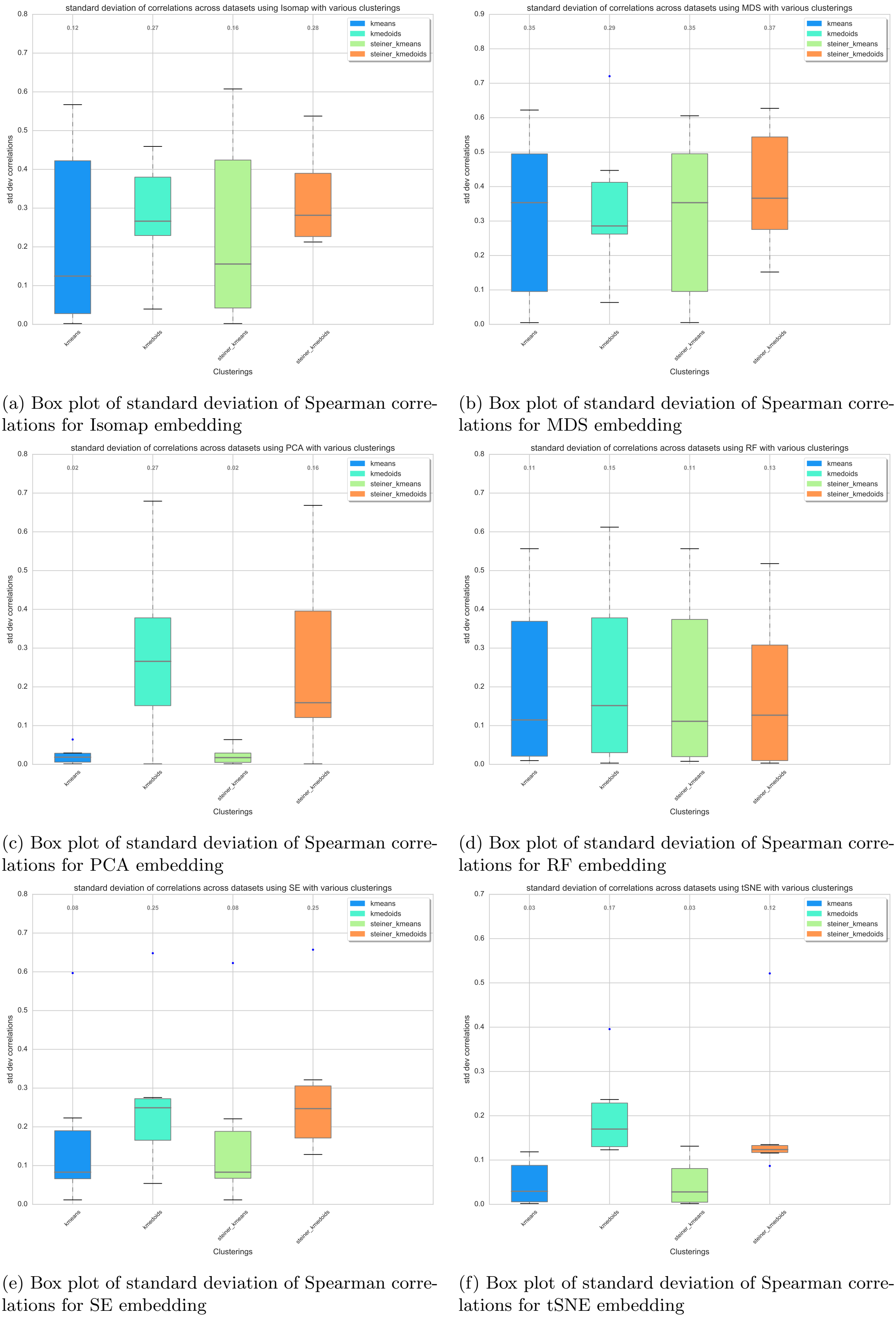
(c) Box plot of mean of Spearman correlations for PCA embedding

(d) Box plot of mean of Spearman correlations for RF embedding

(e) Box plot of mean of Spearman correlations for SE embedding

(f) Box plot of mean of Spearman correlations for tSNE embedding

Figure 17: Box plots of means Spearman correlations among all the datasets comparing the various clustering algorithms used. Values at the top of each figures are the median values.



(a) Box plot of standard deviation of Spearman correlations for Isomap embedding

(b) Box plot of standard deviation of Spearman correlations for MDS embedding

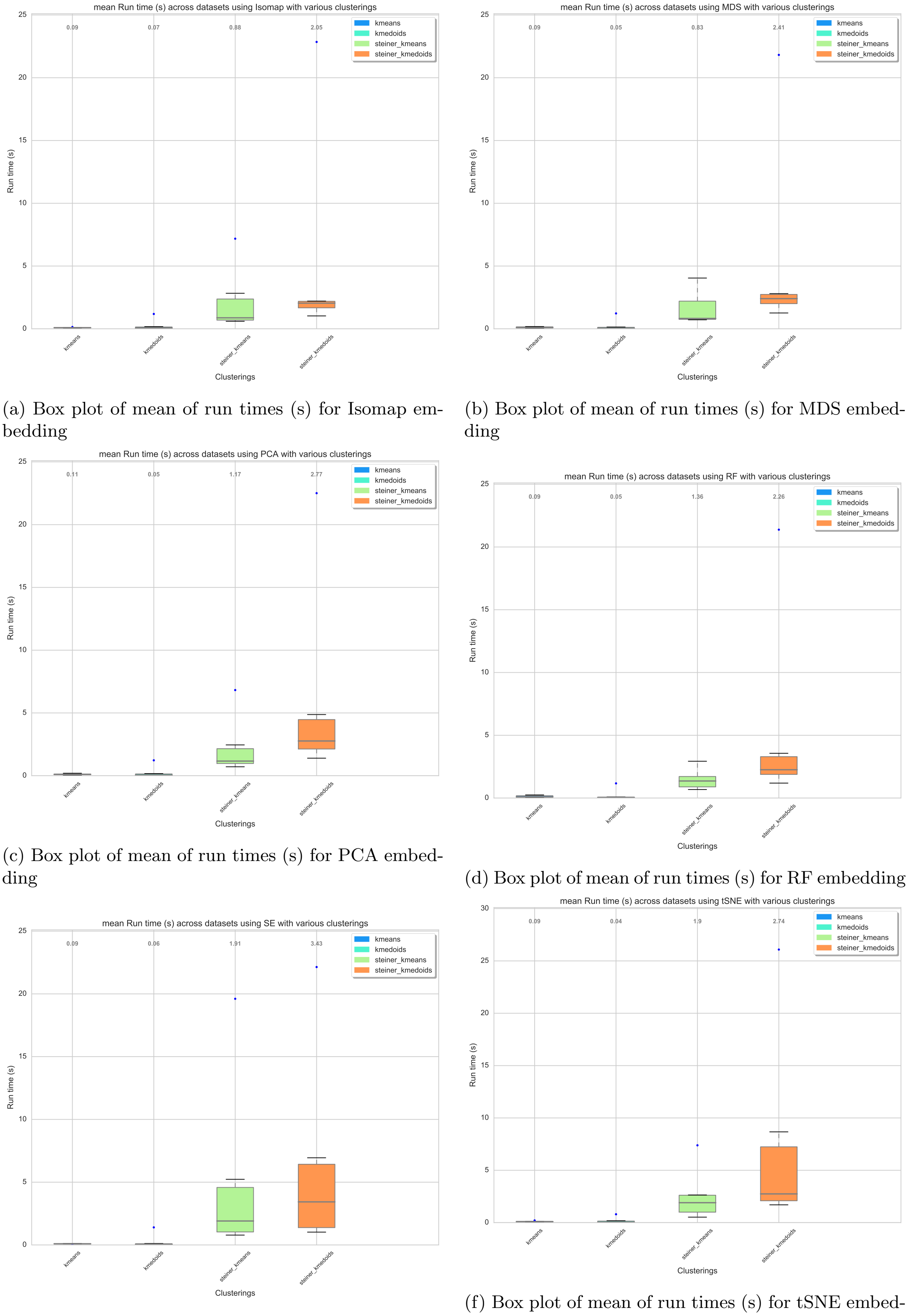
(c) Box plot of standard deviation of Spearman correlations for PCA embedding

(d) Box plot of standard deviation of Spearman correlations for RF embedding

(e) Box plot of standard deviation of Spearman correlations for SE embedding

(f) Box plot of standard deviation of Spearman correlations for tSNE embedding

Figure 18: Box plots of standard deviation of Spearman correlations among all the datasets comparing the various clustering algorithms used. Values at the top of each figures are the median values.



(a) Box plot of mean of run times (s) for Isomap embedding

(b) Box plot of mean of run times (s) for MDS embedding

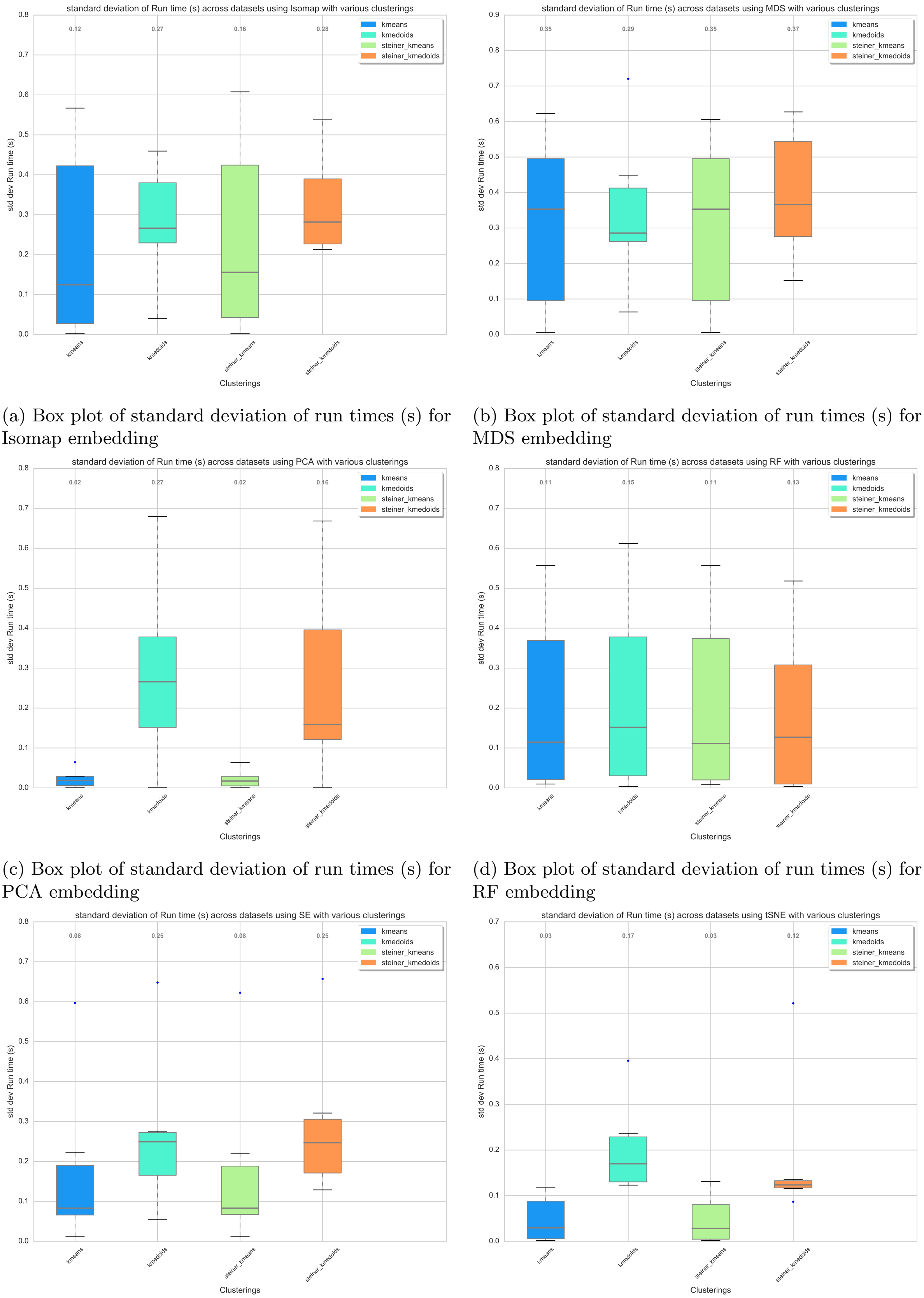
(c) Box plot of mean of run times (s) for PCA embedding

(d) Box plot of mean of run times (s) for RF embedding

(e) Box plot of mean of run times (s) for SE embedding

(f) Box plot of mean of run times (s) for tSNE embedding

Figure 19: Box plots of means run times (s) among all the datasets comparing the various clustering algorithms used. Values at the top of each figures are the median values.



(a) Box plot of standard deviation of run times (s) for Isomap embedding

(b) Box plot of standard deviation of run times (s) for MDS embedding

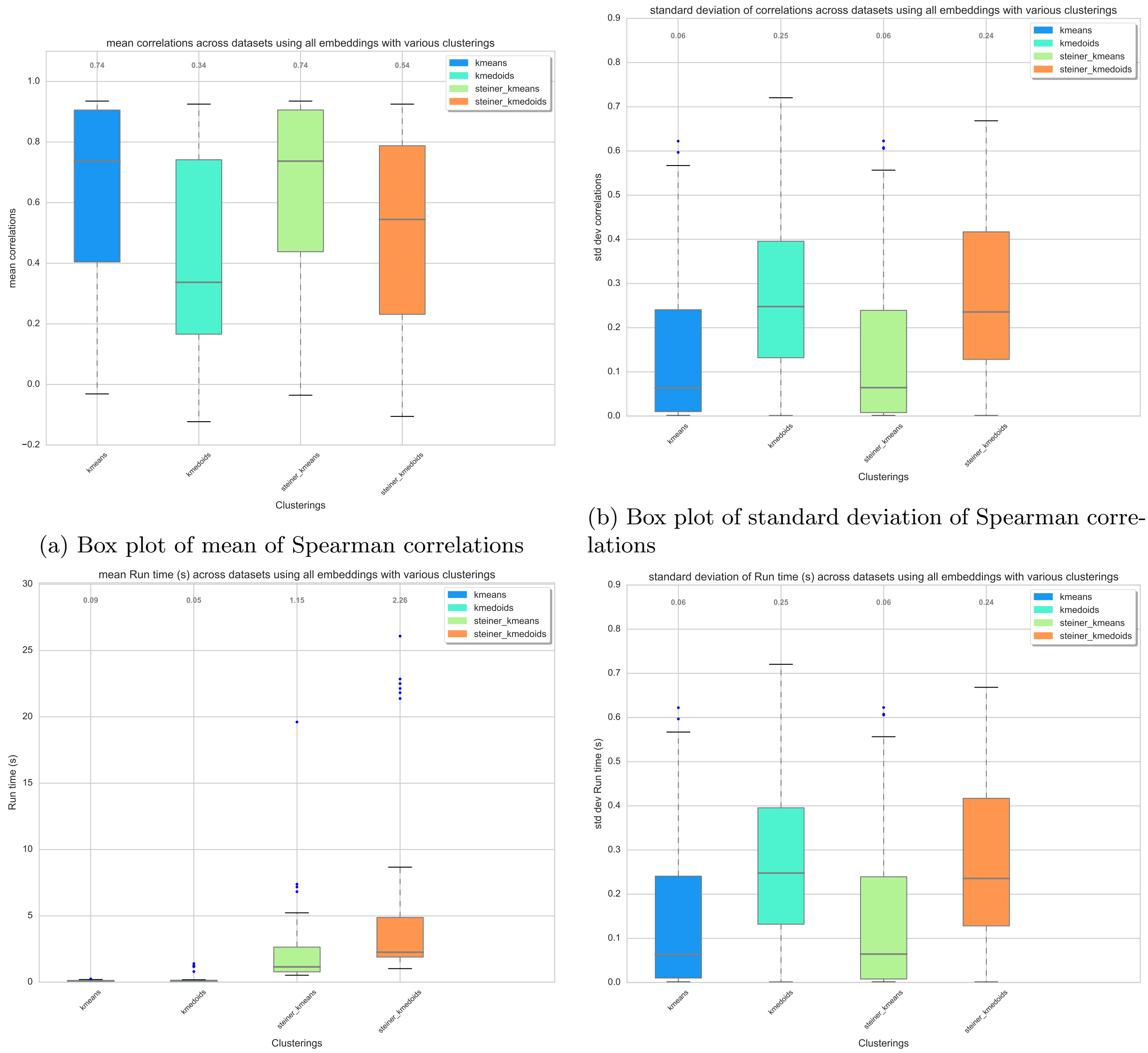
(c) Box plot of standard deviation of run times (s) for PCA embedding

(d) Box plot of standard deviation of run times (s) for RF embedding

(e) Box plot of standard deviation of run times (s) for SE embedding

(f) Box plot of standard deviation of run times (s) for tSNE embedding

Figure 20: Box plots of standard deviation of run times (s) among all the datasets comparing the various clustering algorithms used. Values at the top of each figures are the median values.



(a) Box plot of mean of Spearman correlations

(b) Box plot of standard deviation of Spearman correlations

(c) Box plot of mean run times (s)

(d) Box plot of standard deviation of run times (s)

Figure 21: Box plot of mean and standard deviation of correlations and run times (s) across various datasets, embeddings and clusterings. Values at the top of each figures are the median values.