SICE COUPLED WITH STATIC PAGE RANK FOR FEATURE SELECTION

A PREPRINT

Dr.Mansoor Rezghi, Sepideh Moradi

July 11, 2022

1. First we segment the data using slide window method (we did this without overlapping between segments.) and then cosine similarity matrix for each segment is calculated.
   for example by applying slide window on a data with 76 features and 1617029 samples, and a window of size 1000 we obtain about 1617 segments .after that we calculate cosine similarity between each two features. for instance for k-th segment cosine similarity matrix denoted by C_k is :

   $$C_{k(i,j)} = cosinesimilarity(f_{k(i)}, f_{k(j)}) \tag{1}$$

   where $C_k \in R^{76\times 76}$ for $k = 1, \ldots, 1617$
2. If we stack these matrices frontly a tensor with the following form will be obtained:

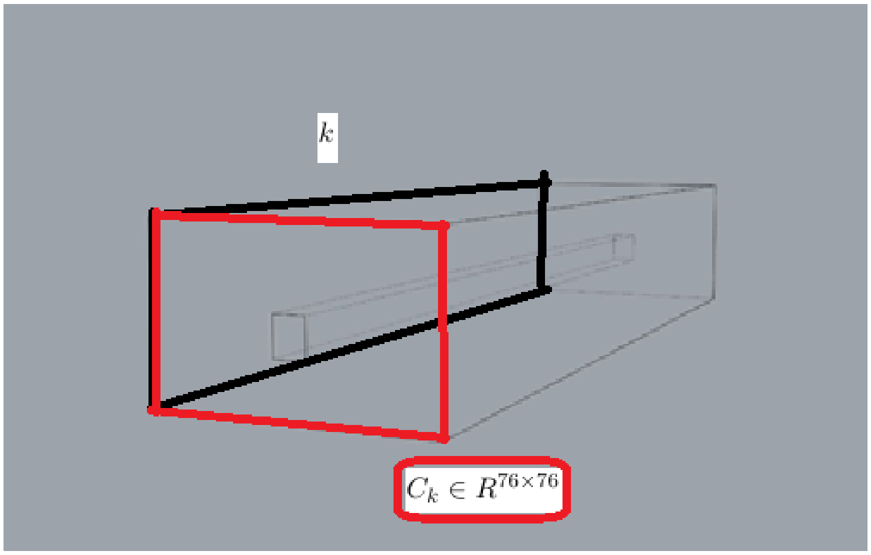


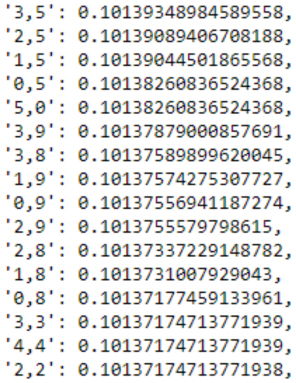
Figure 1

where the inner cube, $\mathcal{C}_{i,j,:}$ is the simlirity between features i and j over k time stamps.
Now we denote $y_{ij} = \mathcal{C}_{i,j,:}$ and construct the matrix

$$H_{ij,pq} = cosinesimilarity(y_{ij}, y_{pq}) \tag{2}$$

So matrix H contains the similarity of the relation of each feature pair over time with other pairs.
in the above example $H \in R^{5776\times 5776}$.

3. In the next step we compute Absolute value of H and substitude values which are less than 0.125 with zero and then apply page rank algorithm on it and sort the feature pairs ranks in descending order.
   for example for our example 15 top ranked pairs are :



```
'3,5': 0.10139348984589558,
'2,5': 0.10139089406708188,
'1,5': 0.10139044501865568,
'0,5': 0.10138260836524368,
'5,0': 0.10138260836524368,
'3,9': 0.10137879000857691,
'3,8': 0.10137589899620045,
'1,9': 0.10137574275307727,
'0,9': 0.10137556941187274,
'2,9': 0.1013755579798615,
'2,8': 0.10137337229148782,
'1,8': 0.101373100792904З,
'0,8': 0.10137177459133961,
'3,3': 0.10137174713771939,
'4,4': 0.10137174713771939,
'2,2': 0.10137174713771938,
```

Figure 2

4. Next for each feature which is involved in top ranked relations we sum up its ranks .for example in the above instance feature 1 is encountered in 3 rows:
   $1, 5$: 0.10139044501865568,
   $1, 9$: 0.10137574275307727,
   $1, 8$: 0.1013731007929043,
   So its total score is about 0.304138 .
   Lastly we sort out the scores and we select features which are associated with top scores and make our regression model base on them.

5. Results: number of componentsis 5.
   Results for linear regression :

   1962-07-31 : 1964-06-30
   MSE of pca is: 0.36537470702410885
   mse model is: 0.38382745154458825
   1964-07-31 : 1966-06-30
   MSE of pca is: 0.5647523137387686
   mse model is: 0.560134411633979
   1966-07-31 : 1968-06-30
   MSE of pca is: 0.7700791146028392
   mse model is: 0.7663538728389412
   1968-07-31 : 1970-06-30
   MSE of pca is: 0.8515996581394818
   mse model is: 0.8486710017375306
   1970-07-31 : 1972-06-30
   MSE of pca is: 0.6048583332950537
   mse model is: 0.6345137333712583
   1972-07-31 : 1974-06-30
   MSE of pca is: 0.6976739728163991
   mse model is: 0.6687275466633703
   1974-07-31 : 1976-06-30
   MSE of pca is: 0.617760322339892
   mse model is: 0.6271669485571683
   1976-07-31 : 1978-06-30
   MSE of pca is: 0.6856988277351984
   mse model is: 0.6745008774461432
   1978-07-31 : 1980-06-30
   MSE of pca is: 0.8079108416775255
   mse model is: 0.8122371484044416

```
1980-07-31 : 1982-06-30
MSE of pca is: 0.6184904692751453
mse model is: 0.6162358756433778
1982-07-31 : 1984-06-30
MSE of pca is: 0.728045498908772
mse model is: 0.7342851617066669
1984-07-31 : 1986-06-30
MSE of pca is: 0.9176850815306394
mse model is: 0.9226882626550906
1986-07-31 : 1988-06-30
MSE of pca is: 0.825137899570742
mse model is: 0.8241529764394406
1988-07-31 : 1990-06-30
MSE of pca is: 0.9888828432403375
mse model is: 0.9887728713371153
1990-07-31 : 1992-06-30
MSE of pca is: 1.0170821872374134
mse model is: 1.0238796814881885
1992-07-31 : 1994-06-30
MSE of pca is: 0.7689407845643
mse model is: 0.7750282223635
1994-07-31 : 1996-06-30
MSE of pca is: 1.084976603544571
mse model is: 1.0848589983661154
1996-07-31 : 1998-06-30
MSE of pca is: 1.0038638179589088
mse model is: 1.0145048325371473
1998-07-31 : 2000-06-30
MSE of pca is: 1.3127255406336134
mse model is: 1.2835498507187428
2000-07-31 : 2002-06-30
MSE of pca is: 1.0771474139139088
mse model is: 1.1091744959339862
2002-07-31 : 2004-06-30
MSE of pca is: 0.8363857575928891
mse model is: 0.8260031561748892
2004-07-31 : 2006-06-30
MSE of pca is: 0.7066219191877153
mse model is: 0.7224399090120911
2006-07-31 : 2008-06-30
MSE of pca is: 0.988641304269171
mse model is: 0.9819794865931525
2008-07-31 : 2010-06-30
MSE of pca is: 1.0328388647409406
mse model is: 0.9883551447651389
2010-07-31 : 2012-06-30
MSE of pca is: 0.7946309993640988
mse model is: 0.7965688231027187
2012-07-31 : 2014-06-30
MSE of pca is: 0.7634746145251421
mse model is: 0.7473192721489231

Mse average of our method: 0.8236896158916811
Mse average of PCA: 0.8242799881318297
```

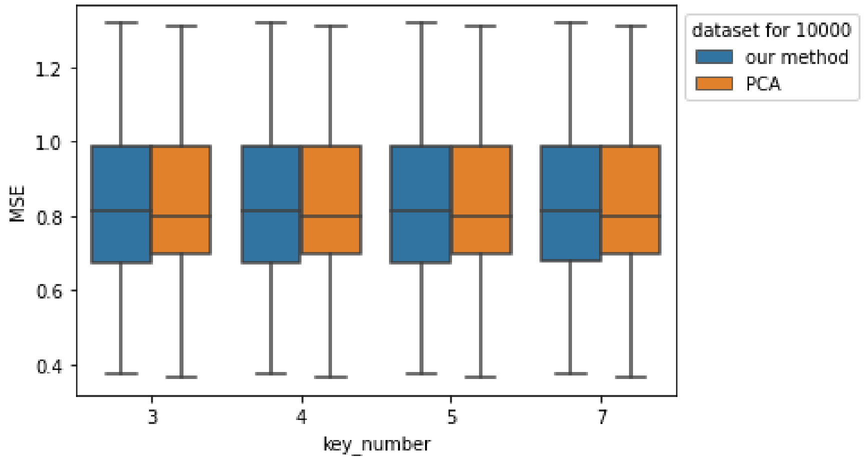


Figure 3

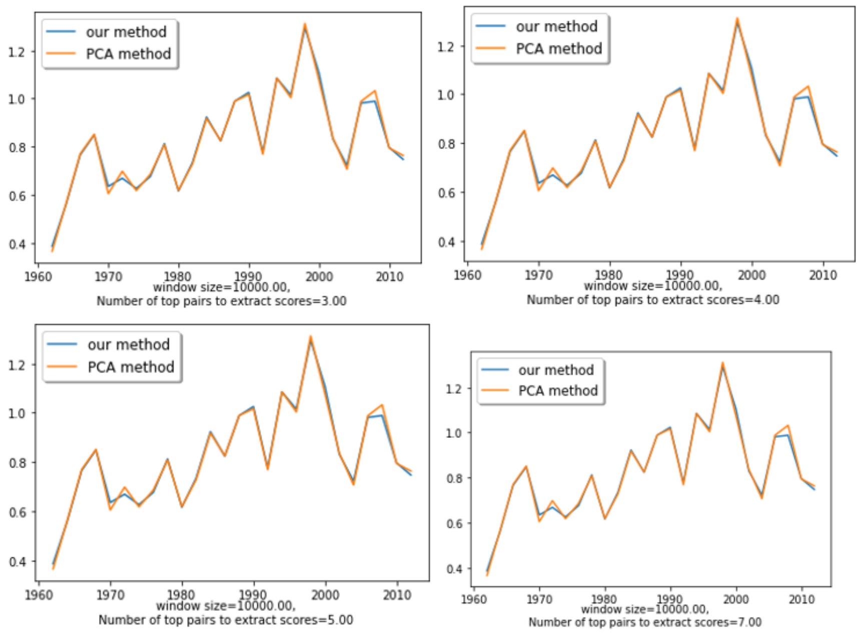


Figure 4

```
Results for xgboost regression
number of components are 5.
1962-07-31 : 1964-06-30
MSE of pca is: 0.4598841220600054
mse model is: 0.48129799291179803
1964-07-31 : 1966-06-30
MSE of pca is: 0.6107484207728083
mse model is: 0.6164241975926117
1966-07-31 : 1968-06-30
MSE of pca is: 0.8041380173062166
mse model is: 0.7916281790736376
1968-07-31 : 1970-06-30
MSE of pca is: 0.874827630650798
mse model is: 0.8766807071708603
```

```
1970-07-31 : 1972-06-30
MSE of pca is: 0.6185122502615008
mse model is: 0.6523765131651403
1972-07-31 : 1974-06-30
MSE of pca is: 0.7317763578391069
mse model is: 0.727359546712835
1974-07-31 : 1976-06-30
MSE of pca is: 0.6524737113252863
mse model is: 0.6484146896009443
1976-07-31 : 1978-06-30
MSE of pca is: 0.6970858285782782
mse model is: 0.6885148452633697
1978-07-31 : 1980-06-30
MSE of pca is: 0.8274629094417606
mse model is: 0.8400211758556159
1980-07-31 : 1982-06-30
MSE of pca is: 0.640661727860927
mse model is: 0.6222491104906764
1982-07-31 : 1984-06-30
MSE of pca is: 0.7543525061563328
mse model is: 0.7345640047770066
1984-07-31 : 1986-06-30
MSE of pca is: 0.9613384016584314
mse model is: 0.9351215310526068
1986-07-31 : 1988-06-30
MSE of pca is: 0.855609729761162
mse model is: 0.8941499506172376
1988-07-31 : 1990-06-30
MSE of pca is: 1.0157404268330783
mse model is: 1.0312235376735384
1990-07-31 : 1992-06-30
MSE of pca is: 1.092794327068058
mse model is: 1.051812295084315
1992-07-31 : 1994-06-30
MSE of pca is: 0.956445403951072
mse model is: 0.7894695195616109
1994-07-31 : 1996-06-30
MSE of pca is: 1.1145781624798077
mse model is: 1.1172714310015877
1996-07-31 : 1998-06-30
MSE of pca is: 1.0427840202151097
mse model is: 1.047640517838594
1998-07-31 : 2000-06-30
MSE of pca is: 1.3836621907856723
mse model is: 1.4572116830328237
2000-07-31 : 2002-06-30
MSE of pca is: 1.09802221113403
mse model is: 1.1368945867907096
2002-07-31 : 2004-06-30
MSE of pca is: 0.8182650777251089
mse model is: 0.8586017730197253
2004-07-31 : 2006-06-30
MSE of pca is: 0.7911743848335295
mse model is: 0.7340953213990877
2006-07-31 : 2008-06-30
MSE of pca is: 1.0246675463254222
mse model is: 0.9905893193022491
2008-07-31 : 2010-06-30
MSE of pca is: 1.0675916871752786
```

```
mse model is: 1.0144978224478467
2010-07-31 : 2012-06-30
MSE of pca is: 0.8231744117244277
mse model is: 0.8131604758276375
2012-07-31 : 2014-06-30
MSE of pca is: 0.8013141271644433
mse model is: 0.7676986521032856

Mse average of our method: 0.8584218992064366
Mse average of PCA: 0.8661186765802943
```

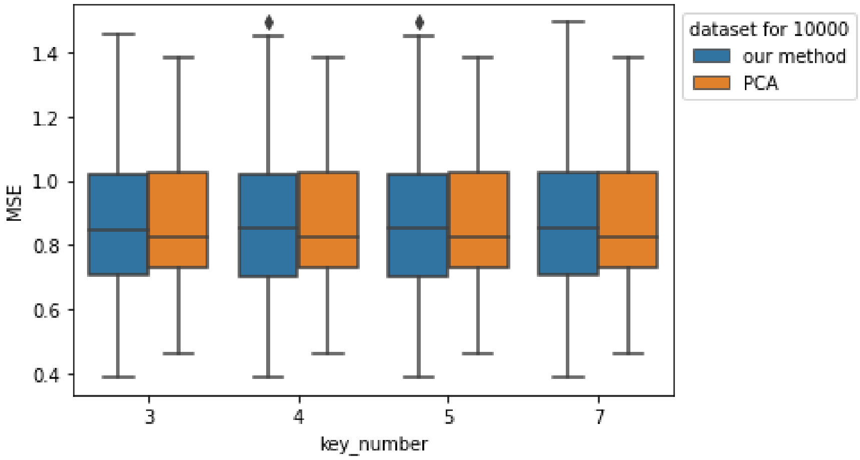


Figure 5

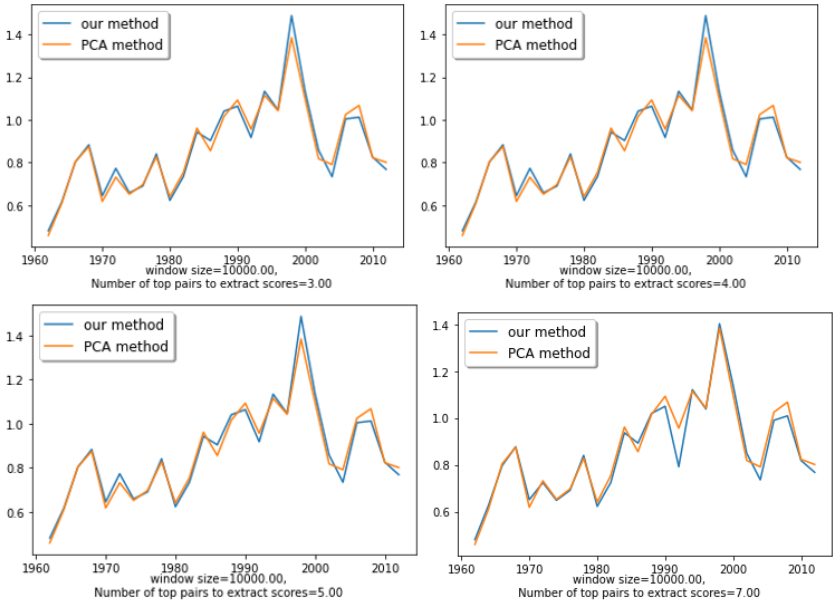


Figure 6

Results: number of components is 13.

Results for linear regression :

```
1962-07-31 : 1964-06-30
MSE of pca is: 0.973883186369144
mse model is: 0.3857728616883571
1964-07-31 : 1966-06-30
MSE of pca is: 0.5634486561923692
mse model is: 0.5601530452407651
1966-07-31 : 1968-06-30
MSE of pca is: 0.7850982582736661
mse model is: 0.766552964237521
1968-07-31 : 1970-06-30
MSE of pca is: 0.880395728338137
mse model is: 0.848922750878503
1970-07-31 : 1972-06-30
MSE of pca is: 0.6043328797862578
mse model is: 0.6348251501766982
1972-07-31 : 1974-06-30
MSE of pca is: 0.7168486199870459
mse model is: 0.6688621806205933
1974-07-31 : 1976-06-30
MSE of pca is: 0.6122040748665822
mse model is: 0.6268943975848779
1976-07-31 : 1978-06-30
MSE of pca is: 0.6940243446296017
mse model is: 0.6755214893631525
1978-07-31 : 1980-06-30
MSE of pca is: 0.8093809152694535
mse model is: 0.8119195194088319
1980-07-31 : 1982-06-30
MSE of pca is: 0.618095117534729
mse model is: 0.6159436903924871
1982-07-31 : 1984-06-30
MSE of pca is: 0.7232916203353306
mse model is: 0.734736216248754
1984-07-31 : 1986-06-30
MSE of pca is: 0.9140412214337307
mse model is: 0.9227955666043581
1986-07-31 : 1988-06-30
MSE of pca is: 0.8232191520245505
mse model is: 0.8241849498951266
1988-07-31 : 1990-06-30
MSE of pca is: 0.9867743637182256
mse model is: 0.9887413550452435
1990-07-31 : 1992-06-30
MSE of pca is: 1.0174189519823171
mse model is: 1.0238955426188383
1992-07-31 : 1994-06-30
MSE of pca is: 0.7663564235124969
mse model is: 0.7750702152043633
1994-07-31 : 1996-06-30
MSE of pca is: 1.0788763782406867
mse model is: 1.0850319039356404
1996-07-31 : 1998-06-30
MSE of pca is: 1.0031653410205286
mse model is: 1.0145773922855015
1998-07-31 : 2000-06-30
MSE of pca is: 1.4102401572282244
mse model is: 1.2846958050146686
```

```
2000-07-31 : 2002-06-30
MSE of pca is: 1.0882270566823273
mse model is: 1.10930425353744
2002-07-31 : 2004-06-30
MSE of pca is: 0.8377624278581158
mse model is: 0.8255533037190973
2004-07-31 : 2006-06-30
MSE of pca is: 0.7066670220385783
mse model is: 0.722453682331854
2006-07-31 : 2008-06-30
MSE of pca is: 0.986810200886446
mse model is: 0.9818737545547132
2008-07-31 : 2010-06-30
MSE of pca is: 1.0237280004531355
mse model is: 0.9886707003624974
2010-07-31 : 2012-06-30
MSE of pca is: 0.7948160233947403
mse model is: 0.7969314031271905
2012-07-31 : 2014-06-30
MSE of pca is: 0.7633589693793781
mse model is: 0.7473281448955912

Mse average of our method: 0.8238927784220256
Mse average of PCA: 0.8531717342859924
```

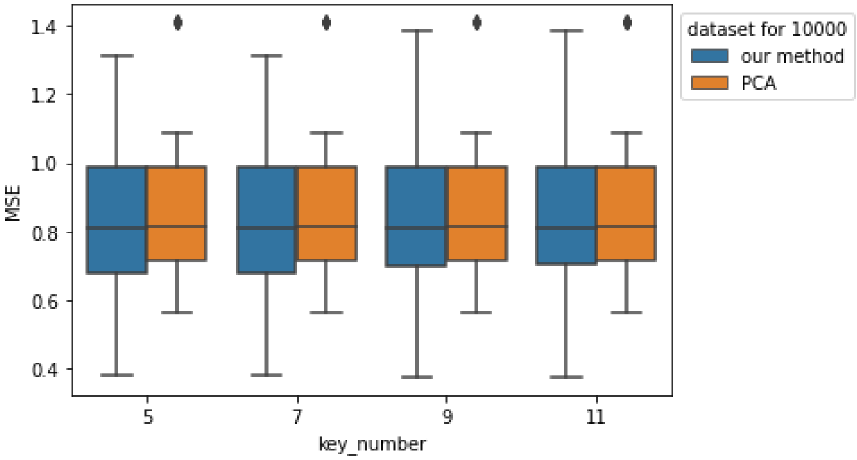


Figure 7

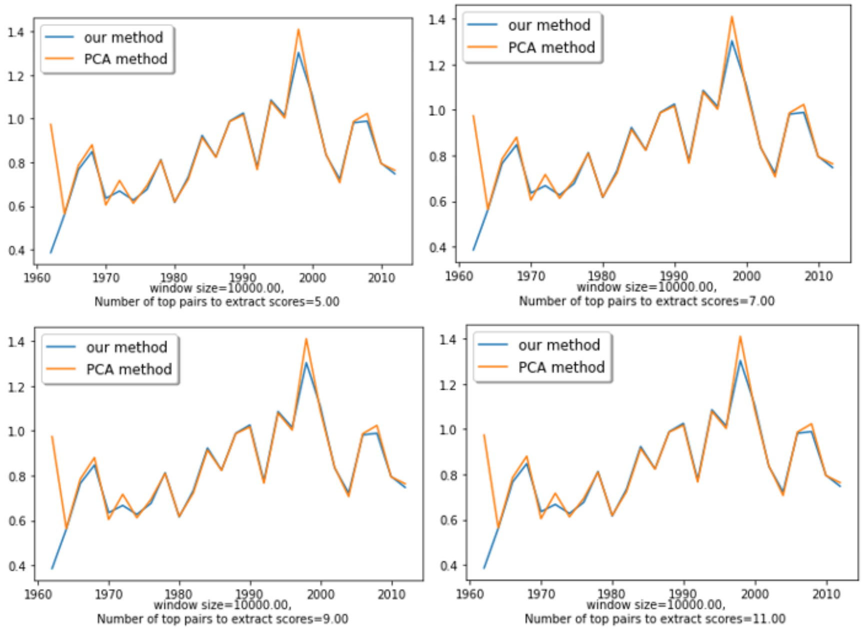


Figure 8

Results for xgboost regression :

```
1962-07-31 : 1964-06-30
MSE of pca is: 0.4283341126513111
mse model is: 0.48129799291179803
1964-07-31 : 1966-06-30
MSE of pca is: 0.6116425410827246
mse model is: 0.6206187743607128
1966-07-31 : 1968-06-30
MSE of pca is: 0.8184432657093145
mse model is: 0.8032405248653068
1968-07-31 : 1970-06-30
MSE of pca is: 0.9195191814595027
mse model is: 0.8845425970450588
1970-07-31 : 1972-06-30
MSE of pca is: 0.6295367982232336
mse model is: 0.6408049924540115
1972-07-31 : 1974-06-30
MSE of pca is: 0.7390532630541248
mse model is: 0.7733068839770103
1974-07-31 : 1976-06-30
MSE of pca is: 0.6955755717171007
mse model is: 0.646331320553191
1976-07-31 : 1978-06-30
MSE of pca is: 0.7040759885629102
mse model is: 0.6932098849314344
1978-07-31 : 1980-06-30
MSE of pca is: 0.8426946092232566
mse model is: 0.8439052619727286
1980-07-31 : 1982-06-30
MSE of pca is: 0.6372417899005514
mse model is: 0.6231043235264158
1982-07-31 : 1984-06-30
MSE of pca is: 0.7747085353849054
```

```
mse model is: 0.7328821746737061
1984-07-31 : 1986-06-30
MSE of pca is: 0.9537531720204316
mse model is: 0.9432410499669652
1986-07-31 : 1988-06-30
MSE of pca is: 0.8751308085152942
mse model is: 0.9092280268086127
1988-07-31 : 1990-06-30
MSE of pca is: 1.0433706608106441
mse model is: 1.0384525871535624
1990-07-31 : 1992-06-30
MSE of pca is: 1.1192153195113304
mse model is: 1.074143258995525
1992-07-31 : 1994-06-30
MSE of pca is: 0.793372398016508
mse model is: 0.900367278275317
1994-07-31 : 1996-06-30
MSE of pca is: 1.1029599975392002
mse model is: 1.1306063000362163
1996-07-31 : 1998-06-30
MSE of pca is: 1.0596765576529268
mse model is: 1.0450351042661772
1998-07-31 : 2000-06-30
MSE of pca is: 1.4861244776317055
mse model is: 1.5049767166621428
2000-07-31 : 2002-06-30
MSE of pca is: 1.1037558921448636
mse model is: 1.1411400608180617
2002-07-31 : 2004-06-30
MSE of pca is: 0.8394938696539889
mse model is: 0.8675225753127854
2004-07-31 : 2006-06-30
MSE of pca is: 0.8265299779154592
mse model is: 0.7353769364911464
2006-07-31 : 2008-06-30
MSE of pca is: 1.0310840794919651
mse model is: 1.0027539296861765
2008-07-31 : 2010-06-30
MSE of pca is: 1.036422454463206
mse model is: 1.0163917664840048
2010-07-31 : 2012-06-30
MSE of pca is: 0.8375563081321097
mse model is: 0.8198958726922275
2012-07-31 : 2014-06-30
MSE of pca is: 0.7996936444906205
mse model is: 0.7709587236090778

Mse average of our method: 0.8708974968665143
Mse average of PCA: 0.8734217413445843
```

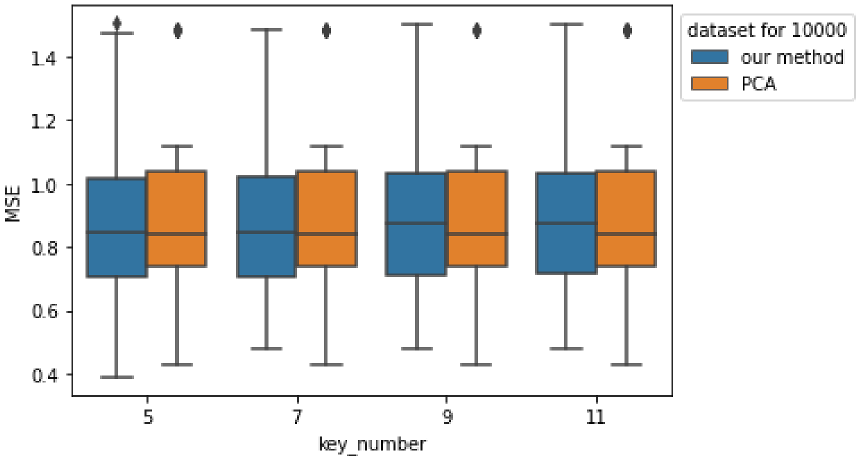


Figure 9

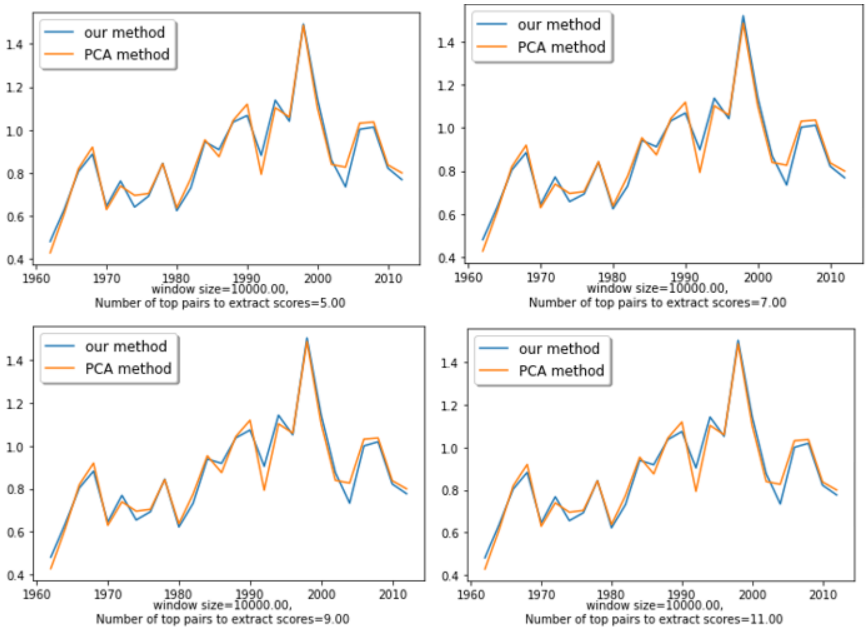


Figure 10