20052199
Bassil Obeidi
CISC 271
March 12th 2019

Principal Components Analysis of Commodities Data

Code Summary: (done only for z as the process for z1 and z2 is the same)

The first step is to perform pcaprelim on the dataset z1. This returns sdiag, a matrix containing the singular values associated with the dataset (eigenvalues), meanvec, a vector containing the average values of every column of the original data set and uvecmat, a matrix containing the eigenvectors associated with the eigenvalues.

The second step is to choose the appropriate k value or cutoff for the data. This was done in two ways. The first was to plot the eigenvalues and visually identify the "knees" in the graph. Looking below, the knees were chosen to be certainly 4 for z1 and possibly somewhere between 2-4 for z2 (unclear from graph). The second method was to use the formula $\rho = \frac{t_k}{t_n}$ where t_k is the sum of the first k eigenvalues and t_n is the sum of all the eigenvalues to find a ratio between 0.5 and 0.6. This was confirmed computationally to be k = 4 for both z1 and z2.

The third step is to pcaapprox on every signal in the dataset, taking this signal and comparing to the original to find the difference vector. Then for every difference vector, the RMSE of the difference vector was taken and inserted in a matrix associating the error with the signal number. This step is to choose the average, best and worst reconstructed signals to perform further analysis.

The final step is to look at the plot of the RMSE for the data set, and find the minimum, maximum and entry with y value (error) closest to the normal error of the data set. And normal error is the average error for every signal reconstructed in the data set.

Performing these steps allows plots to be generated for the mean signals, principal components, reconstructions errors, best, worst, and average reconstructed data.

First 10 Singular values for z1	First 10 Singular values for z2
620.0205 369.7023 313.1078 180.0854 166.0159 146.5297 111.9309 106.6580 94.7451 71.3763	484.916850589674 364.268240691417 348.808176249812 244.592771938819 215.624077770509 161.388627090542 158.905144278216 127.188118566812 109.376034030405 104.855687846321

Analysis:

The first plots that were compared between the data are the mean vectors plots. From these plots it's evident that both mean vectors share the same general trend, however it can be seen that z2 is more delayed as it's similar trends to z1 occur slightly after they happen in z1. Also, z2 is more generally more volatile than z1, with larger and more frequent changes in direction.

Next, the eigenvalue plots were looked at to help determine the k value selected for the respective PCA. For z1, there a clear and most significant "knee" at k = 4. However, for z2 it's less clear which k value to select and thus the equation above was used.

From the RMSE plots and the normal error values printed in the console it's clear that the signals reconstructed from z1 generally have a lower error of reconstruction. This supports the conclusion made above that the data in z2 is more volatile and less correlated. Then the average, best and worst reconstructed signals were obtained from the RMSE plots using the max,

minimum and point closest to the normal error of reconstruction.

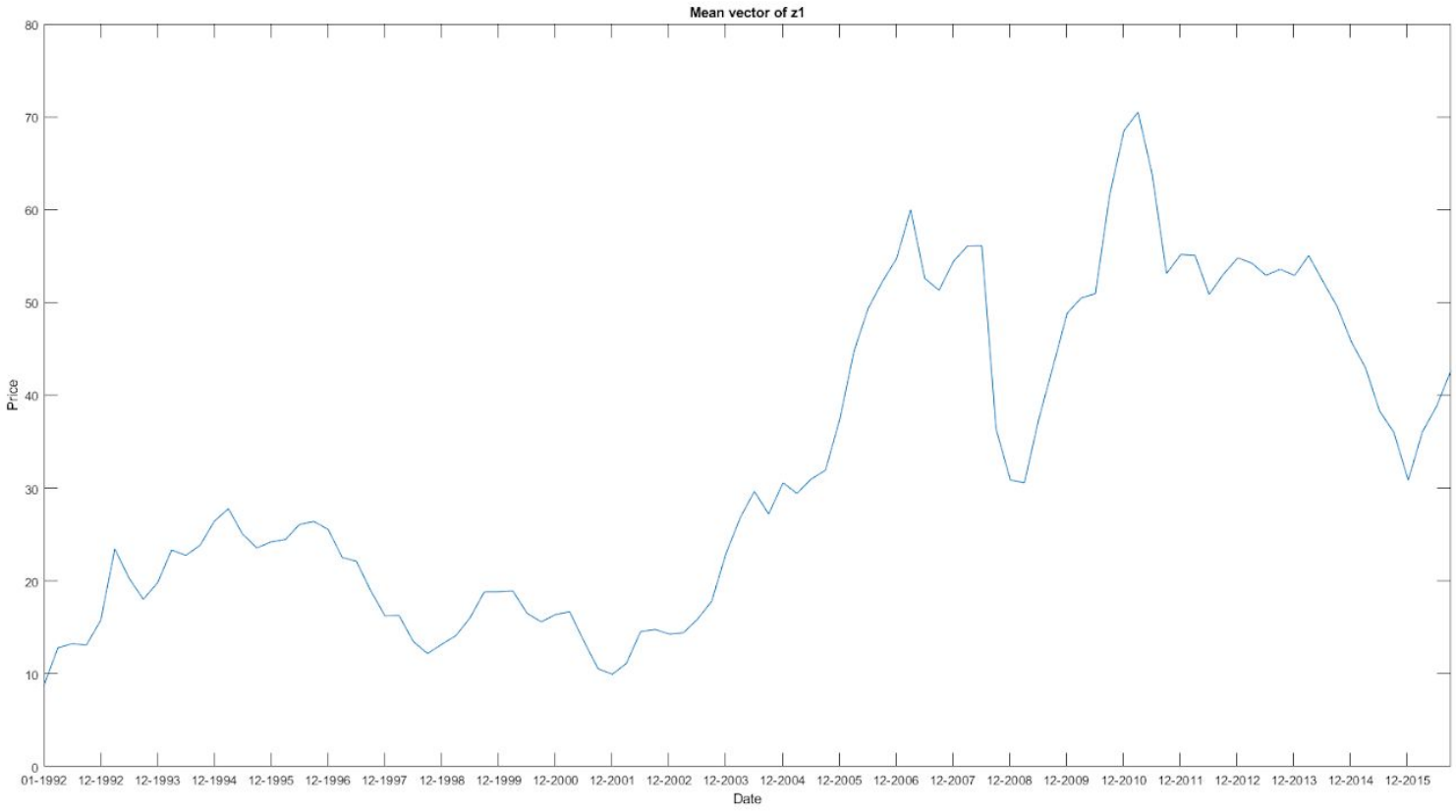
Finally, from examining the best, worst and average case reconstructed signals, some key observations can be made. Primarily, in the best case for z1, it can be seen that the plot experiences well defined and consistent ups and downs with an overall trend slightly upwards. This is in comparison to the best case for z2 in which there is much fewer and less defined peaks and troughs and a clear, strong upwards trend. Additionally, from the worst cases for both z1 and z2, these are both clear examples of counter trend commodity as it goes against the trend of the other signals and PCA has a very poor time reconstructing them

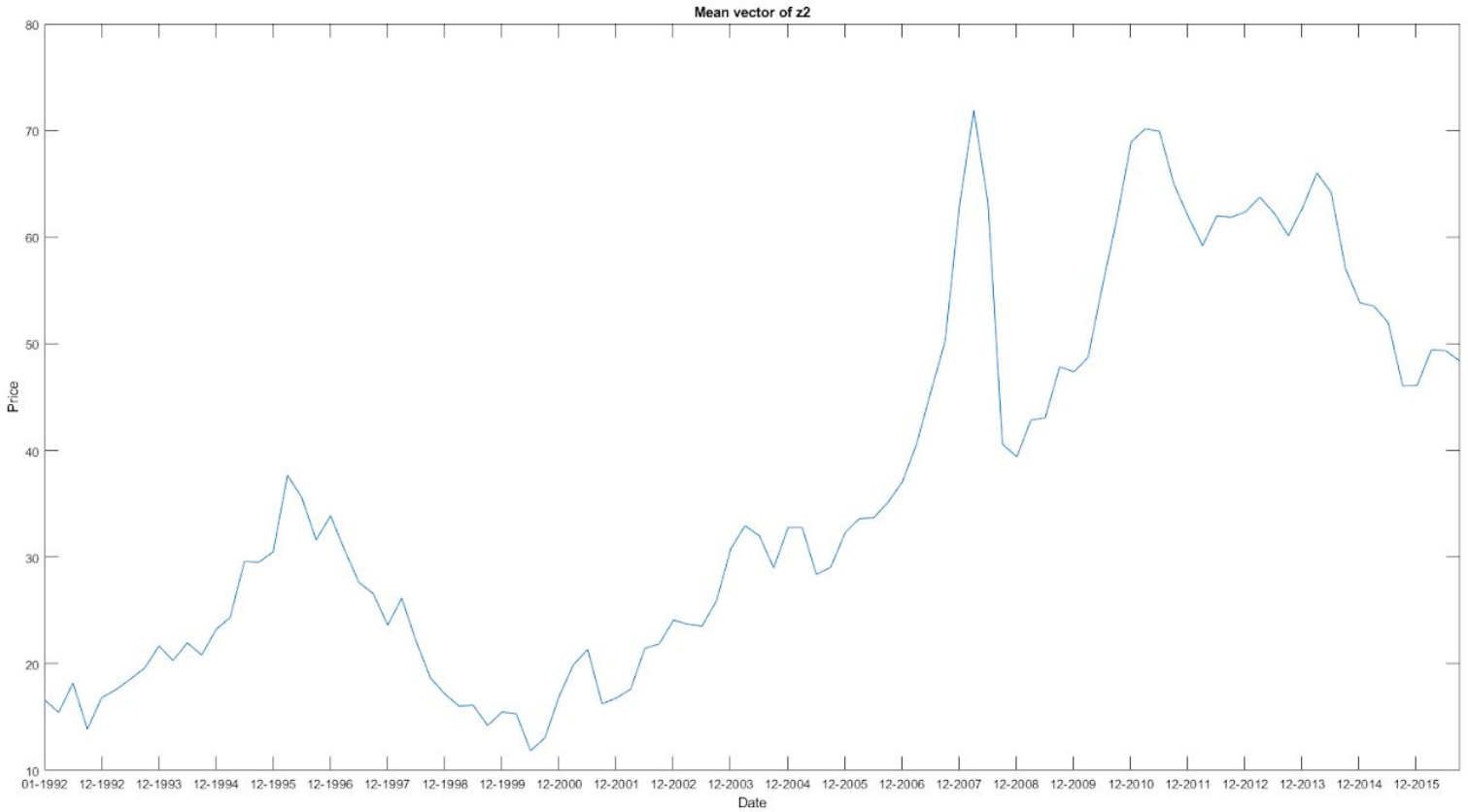
Conclusion:

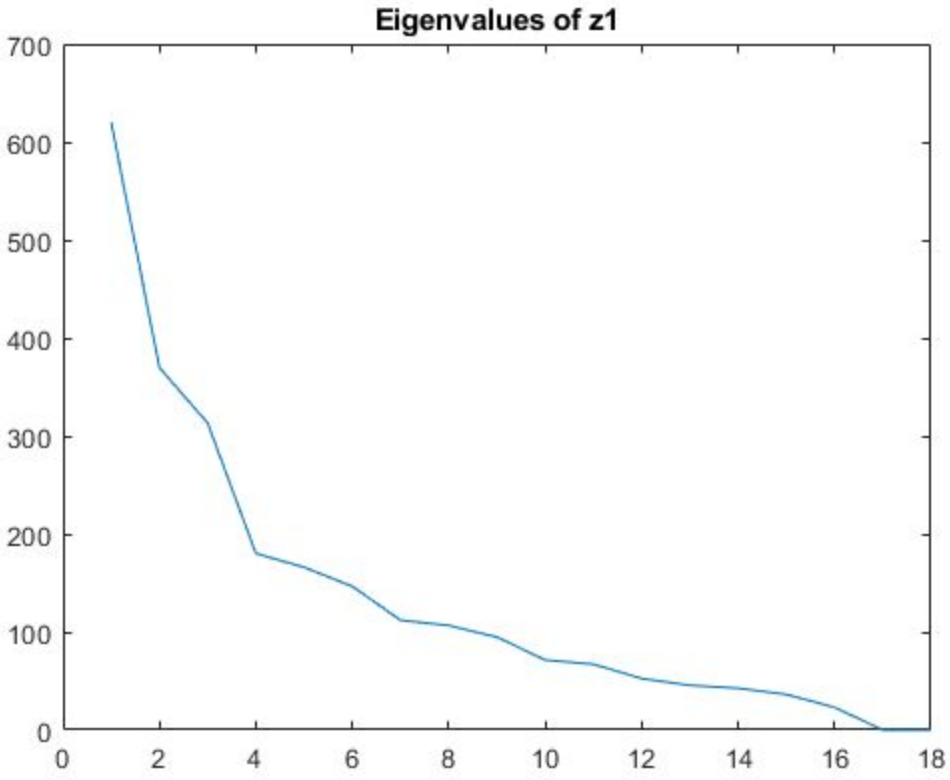
It can be concluded from computation, plotting and analysis that z1 is the set of food related commodity prices, while z2 are other basic-materials commodities. This conclusion can be supported with several reasons. The first is that in real life, raw material prices experience more volatility than food prices. Food is regrown and synthesized every year, while raw materials are taken from a finite source. Next, the best reconstructed signal in z1 displays a clear seasonality and does not trend upwards very much. Food is grown in seasons, the growing season is limited in many places and thus prices constantly change in drastic amounts. However, food prices remain stable over longer periods of times when seasonality is insignificant as food is synthesized and not taken from a finite source. Finally, the best reconstructed signal in z2 displays a very strong upwards trend with some variation but much less than z1. Raw materials can be extracted or produced constantly regardless of season in a lot of cases (not all) and is drawn for a finite source therefore prices increase significantly over time.

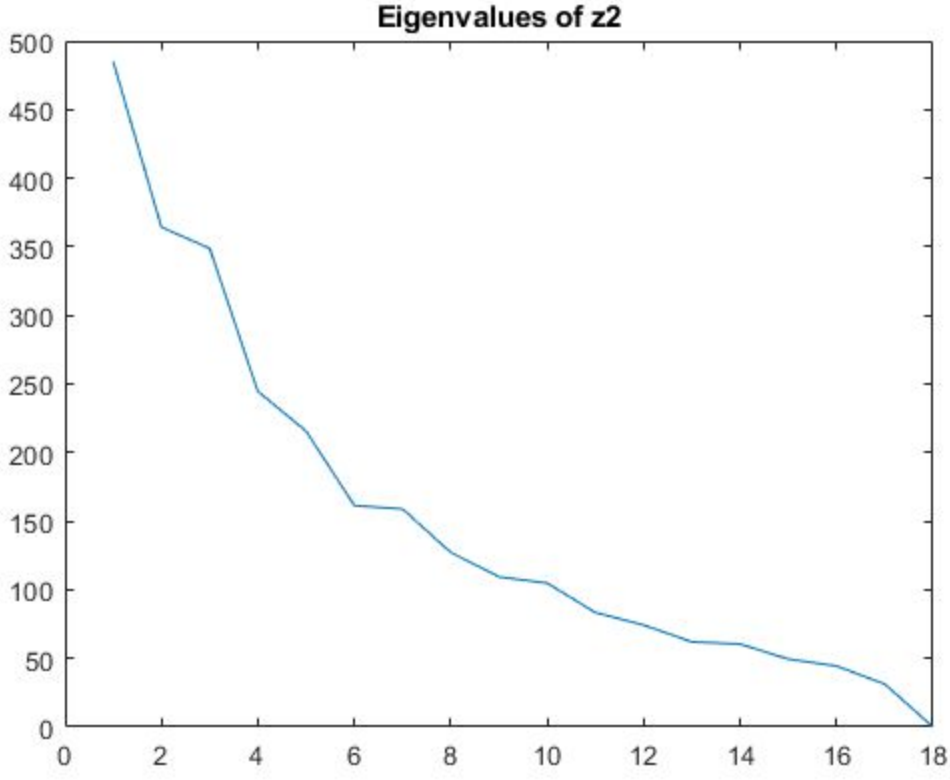
Therefore, z1 is the food price data, and z2 is the raw materials price data.

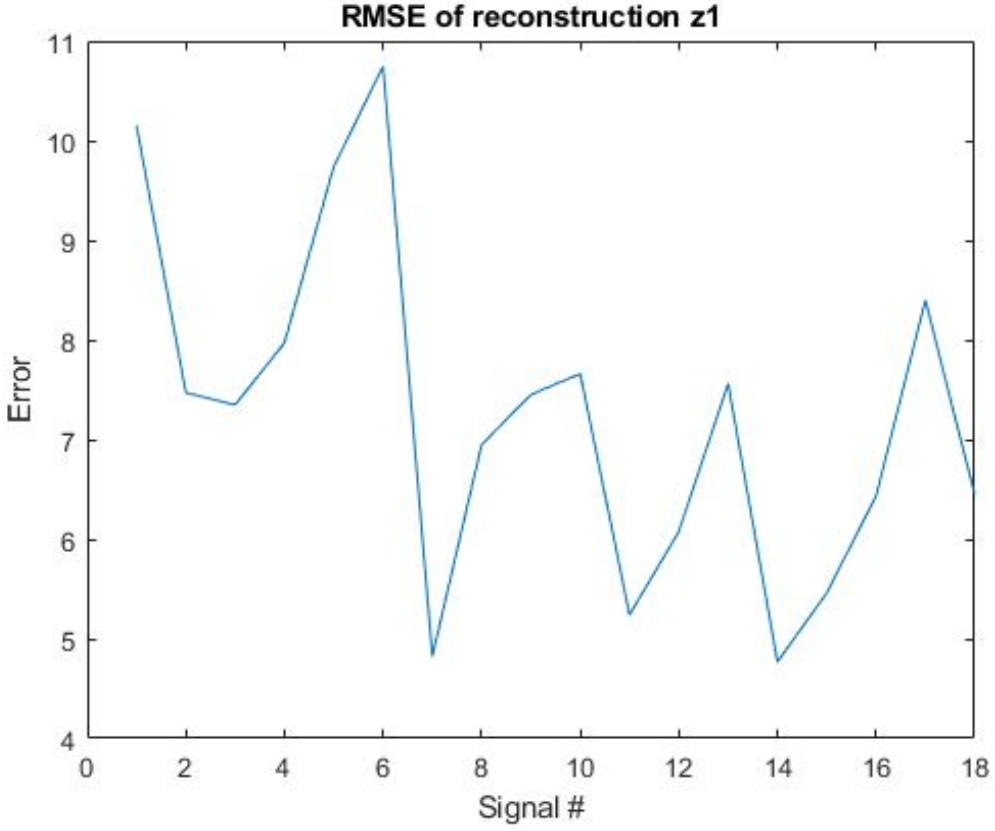
All plots are below.

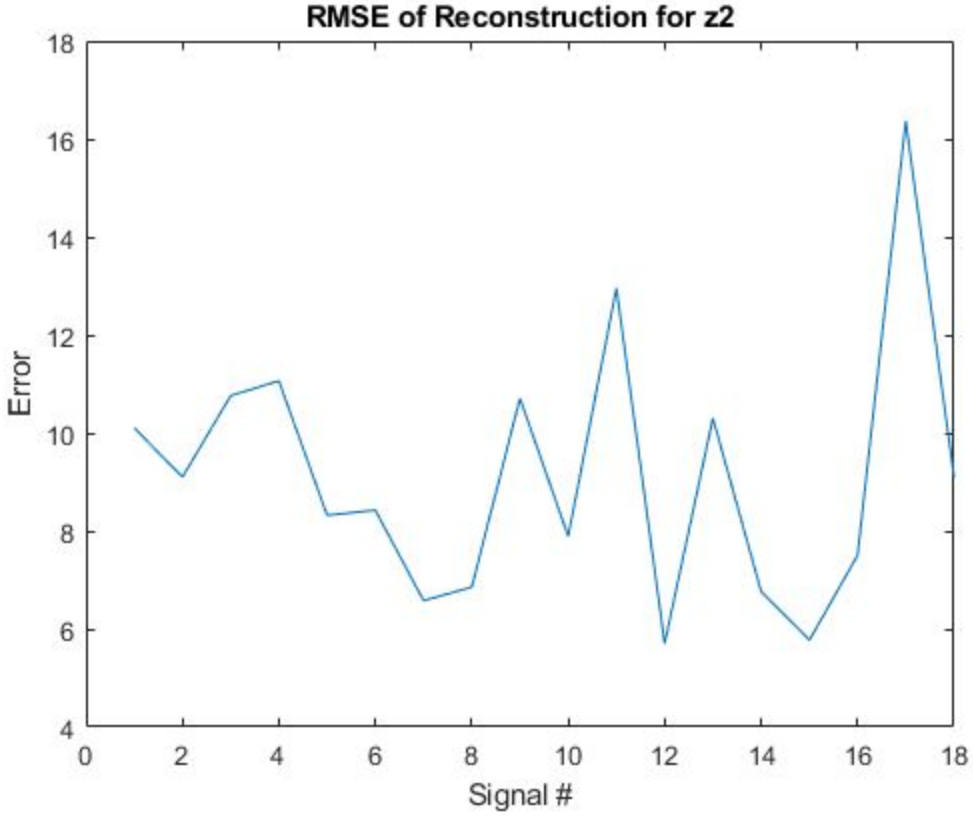

Mean vector of z1
Price
Date
0
10
20
30
40
50
60
70
80
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015

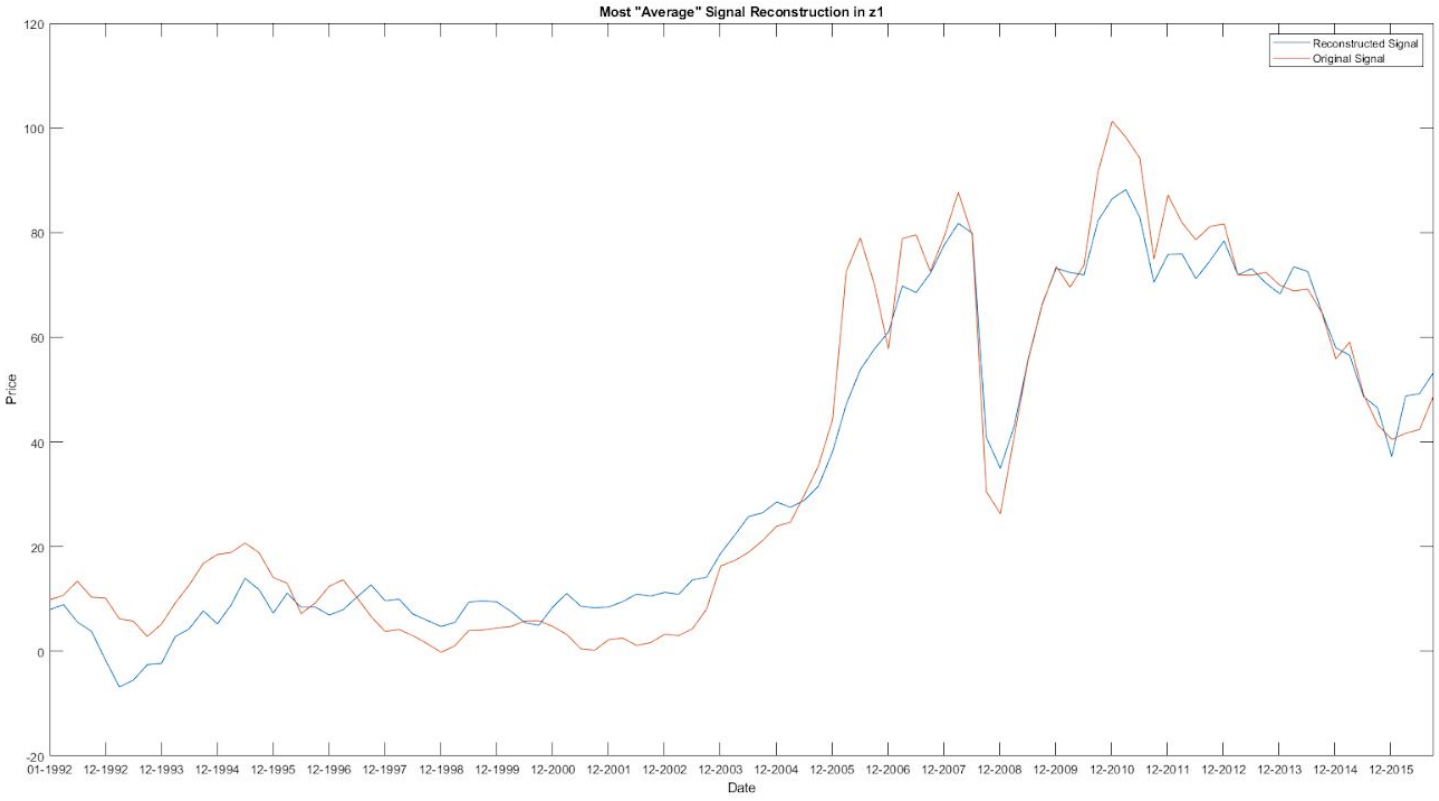

Mean vector of z2
Price
Date
10
20
30
40
50
60
70
80
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015

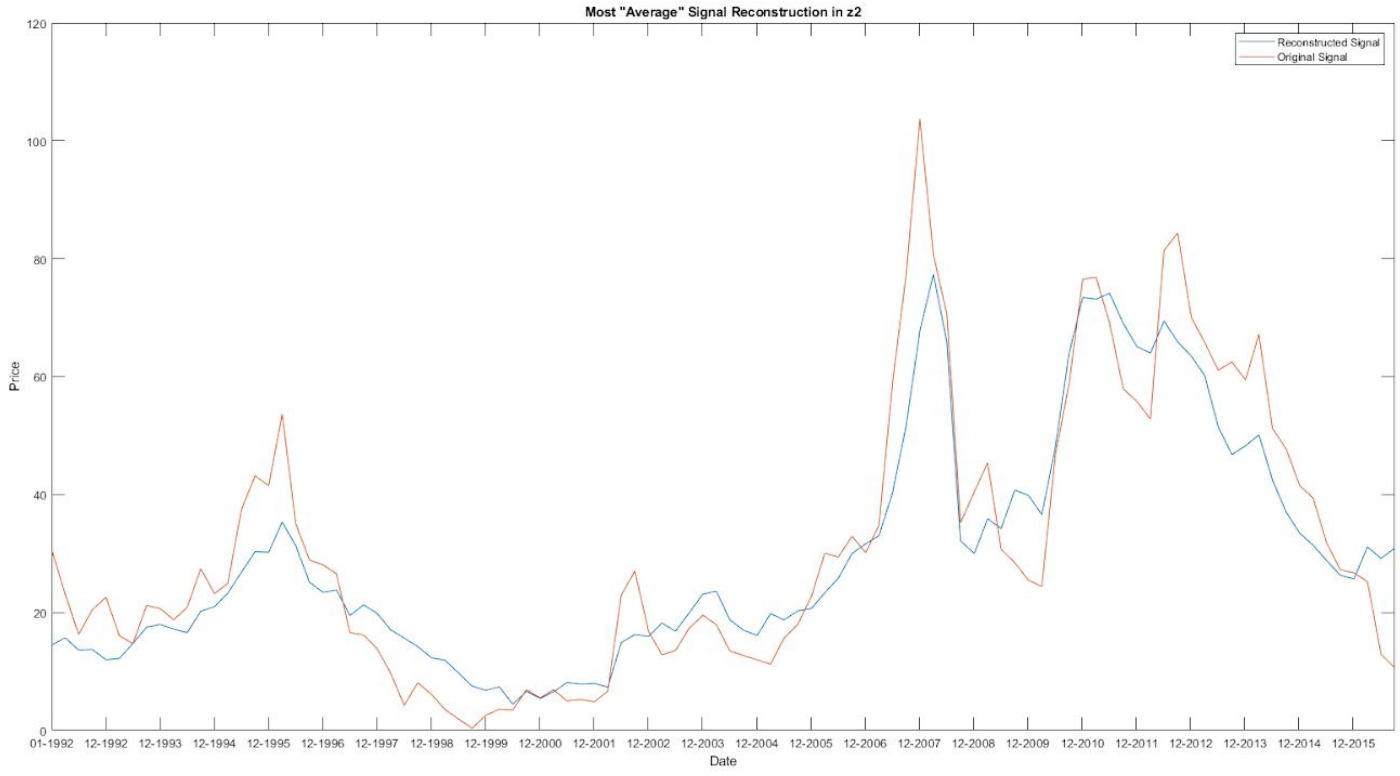

Eigenvalues of z1
700
600
500
400
300
200
100
0
0
2
4
6
8
10
12
14
16
18

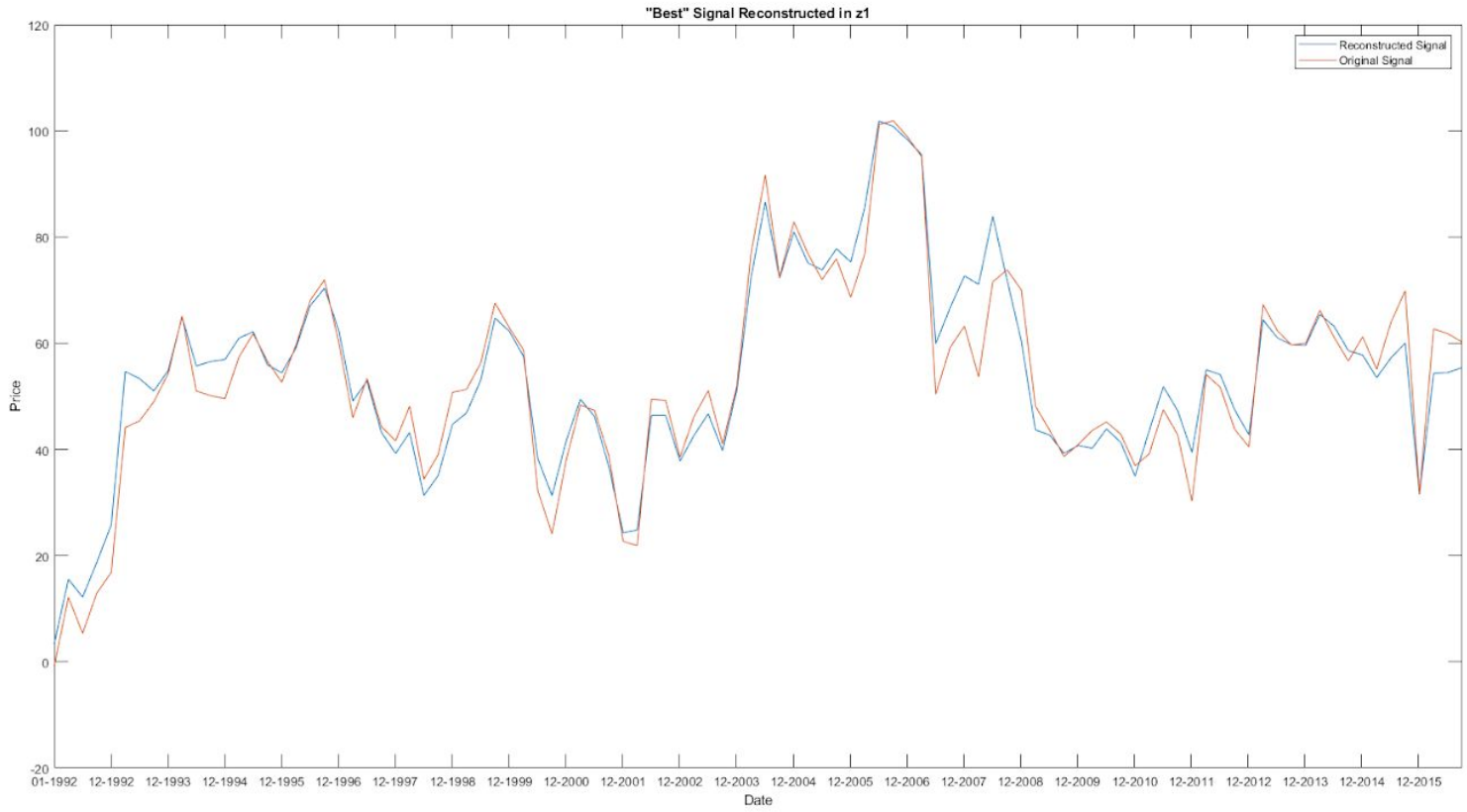

Eigenvalues of z2
500
450
400
350
300
250
200
150
100
50
0
0
2
4
6
8
10
12
14
16
18

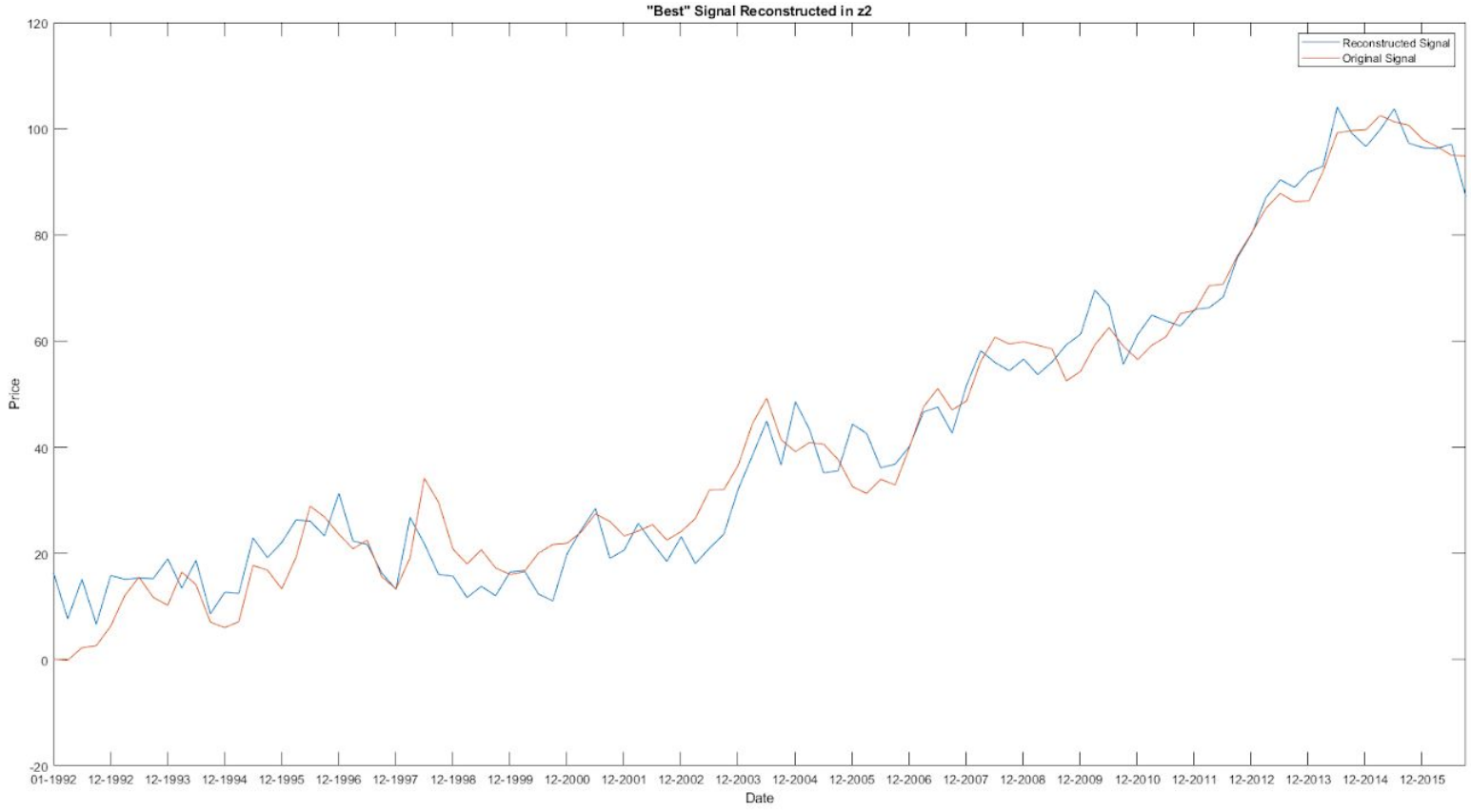

RMSE of reconstruction z1
Error
Signal #
11
10
9
8
7
6
5
4
0
2
4
6
8
10
12
14
16
18

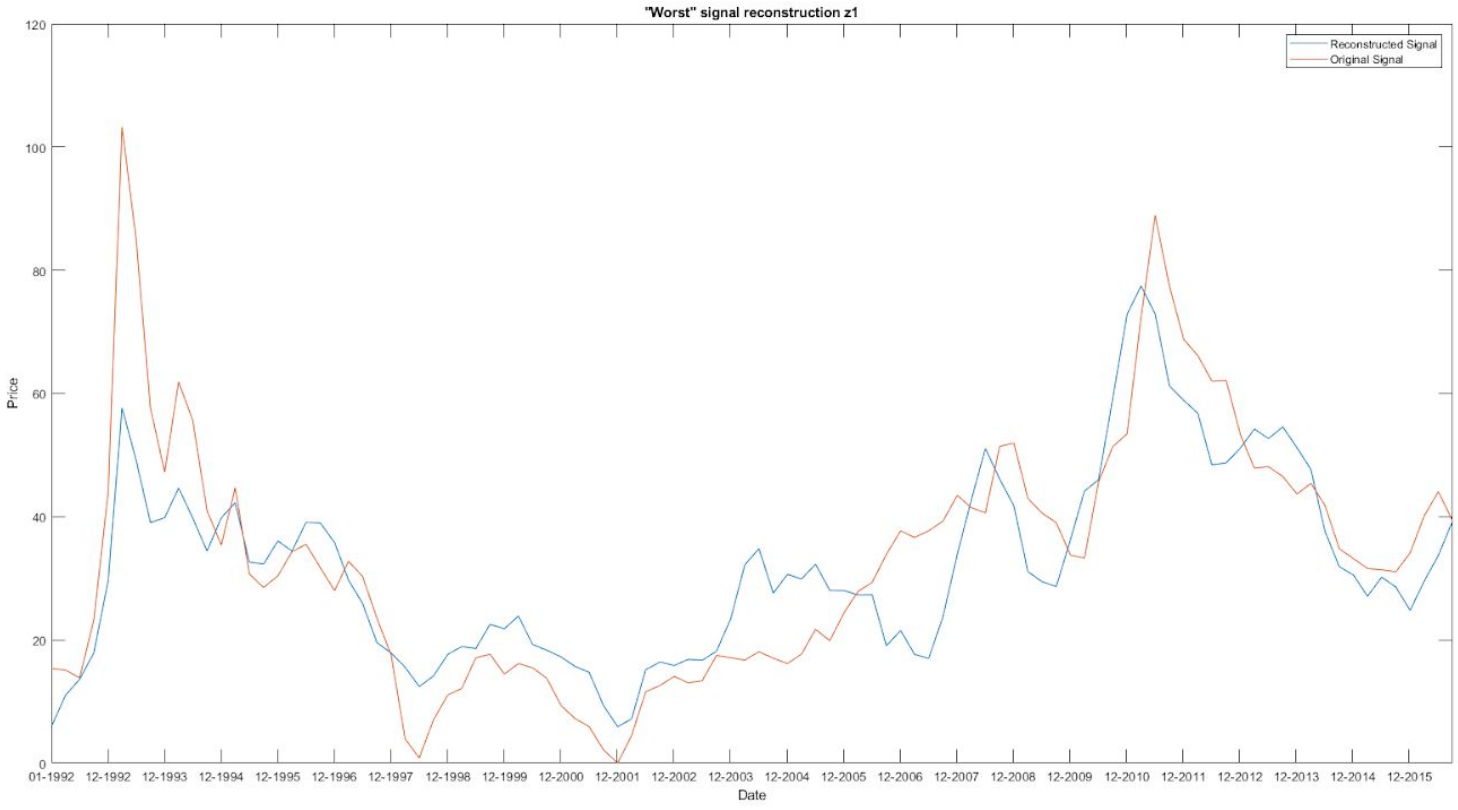

RMSE of Reconstruction for z2
Error
Signal #
18
16
14
12
10
8
6
4
0
2
4
6
8
10
12
14
16
18

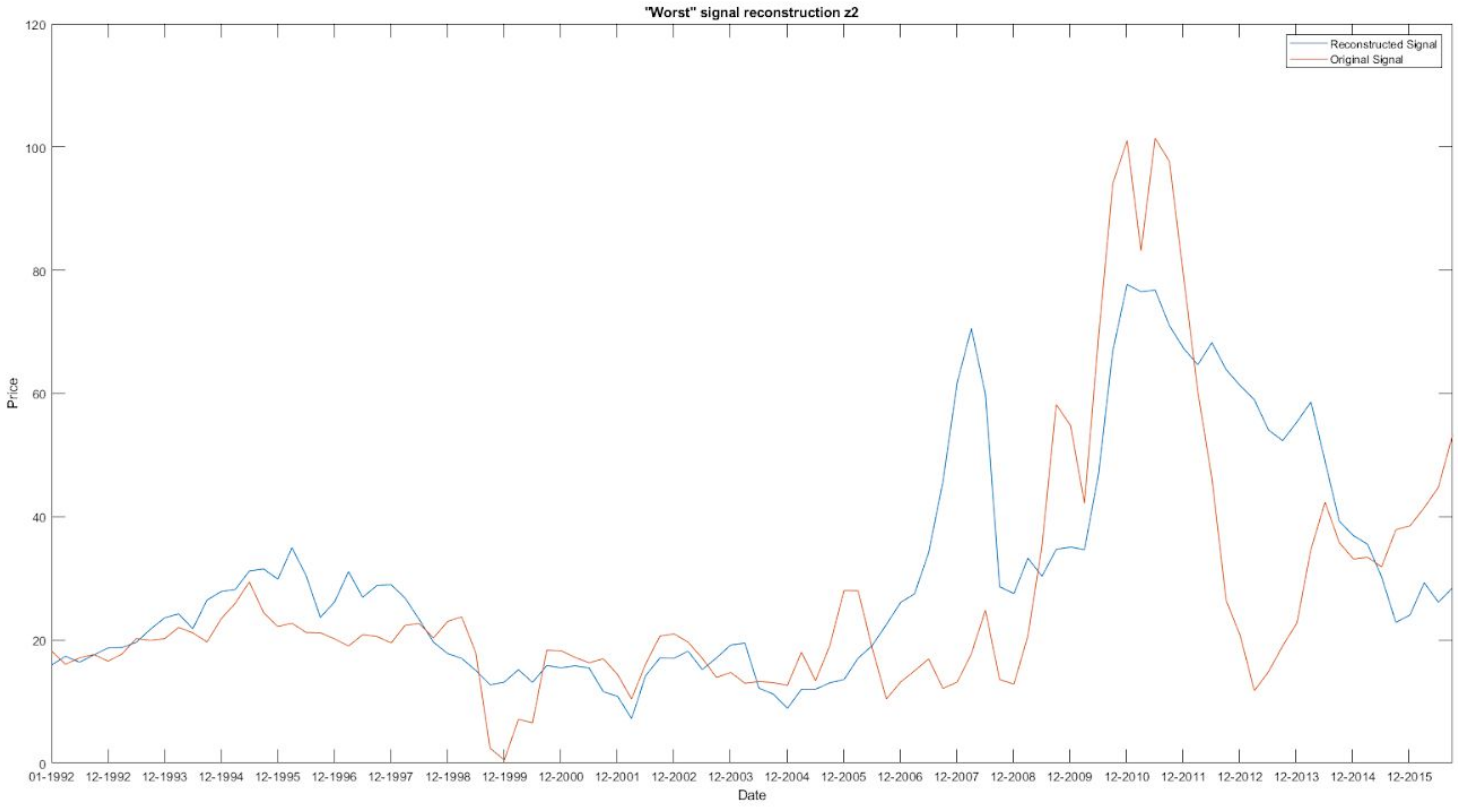

Most "Average" Signal Reconstruction in z1
Reconstructed Signal
Original Signal
Price
Date
120
100
80
60
40
20
0
-20
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015


Most "Average" Signal Reconstruction in z2
Reconstructed Signal
Original Signal
Price
Date
120
100
80
60
40
20
0
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015


"Best" Signal Reconstructed in z1
Reconstructed Signal
Original Signal
Price
Date
120
100
80
60
40
20
0
-20
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015


"Best" Signal Reconstructed in z2
Reconstructed Signal
Original Signal
Price
Date
120
100
80
60
40
20
0
-20
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015


"Worst" signal reconstruction z1
Reconstructed Signal
Original Signal
Price
Date
0
20
40
60
80
100
120
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015


"Worst" signal reconstruction z2
Reconstructed Signal
Original Signal
Price
Date
0
20
40
60
80
100
120
01-1992
12-1992
12-1993
12-1994
12-1995
12-1996
12-1997
12-1998
12-1999
12-2000
12-2001
12-2002
12-2003
12-2004
12-2005
12-2006
12-2007
12-2008
12-2009
12-2010
12-2011
12-2012
12-2013
12-2014
12-2015