Principal Component Analysis in Excel ~ PART I

Posted on June 28, 2015April 4, 2016 by bquanttrading

We decided to write a series of posts on a very useful statistical technique called Principal Component Analysis (PCA). In the current post we give a brief explanation of the technique and its implementation in excel. In practice it is less important to know the computations behind PCA than it is to understand the intuition behind the results. For those who are interested to know the mathematics behind this technique we recommend any multivariate statics book. One book which we really like is Carol Alexander's Market Risk Analysis Volume 1. This book comes with a free excel addin Matrix.xla that can be used to implement PCA in excel. Alternatively the reader can download this excellent addin for free from http://excellaneous.com/Downloads.html (http://excellaneous.com/Downloads.html).

The idea of PCA is to find a set of linear combinations of variables that describe most of the variation in the entire data set. For example, we may have a time series of daily changes in interest rate swap rates for the past year. We consider changes in 2y, 3y, 4y, 5y, 7y, 10y, 15y, 20y, 30y swap tenors. Our data set has nine variables in total. With so many variables it may be easier to consider a smaller number of combinations of this original data rather than consider the full data set. This is often called a reduction in the data set's dimension. Having set the goal of reducing dimension of our data set to a smaller number of factors a simple choice would be to use the average. In our case this would be Average = 1/9*2y+1/9*3y+1/9*4y+1/9*5y+1/9*7y+1/9*10y+1/9*15y+1/9*20y+1/9*30y. Our vector of coefficients C=[1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9] is called a linear combination. Linear combinations where the sum of squared coefficients equal to 1 are called a standardized linear combinations. PCA finds a set of standardized linear combinations where each individual factor is orthogonal (meaning not correlated). There are as many principal components as there are variables in the original data set but they are ordered in such a way that only a few factors explain most of the original data. The orthogonal factors are computed from the correlation or covariance matrix of the original (sometimes standardized) data.

Let's walk through an example to gain a better understanding. We start out with daily changes in US swap rates for abovementioned tenors.

	Daily Swap Rate Change (in bps)								
Date	2y	3y	4y	5y	7y	10y	15y	20y	30y
26/06/2015	4.18	2.67	3.47	4.28	5.18	5.48	6.02	6.05	6.34
25/06/2015	-.83	2.18	2.61	3.15	3.75	4.61	4.68	4.78	4.84
24/06/2015	-1.35	-1.93	-2.51	-3.06	-3.47	-3.40	-3.23	-3.30	-3.11
23/06/2015	1.20	1.83	2.46	2.82	3.70	4.10	4.73	4.82	4.71
22/06/2015	3.60	5.50	7.20	8.69	10.29	11.46	11.55	11.43	11.35
19/06/2015	-2.75	-4.00	-5.30	-5.78	-6.60	-7.60	-8.06	-8.12	-8.30
18/06/2015	-1.25	-1.25	-.15	.20	.90	1.85	2.31	1.97	2.16
17/06/2015	-4.20	-4.05	-4.00	-3.47	-1.95	-.05	1.55	2.65	3.64
16/06/2015	-2.00	-2.90	-3.22	-3.36	-3.65	-3.89	-4.02	-4.00	-4.24
15/06/2015	-2.03	-2.93	-3.50	-3.91	-3.88	-3.49	-2.88	-2.55	-2.11
12/06/2015	1.45	2.00	2.02	2.07	1.78	1.43	1.20	1.25	1.15
11/06/2015	-.27	-2.20	-3.93	-5.63	-8.19	-10.48	-11.11	-11.39	-11.72

(https://asmquantmacro.com/wp-content/uploads/2015/06/shot1.jpg)

We would like to reduce the dimension to as few factors as possible that describe the variability in the data. To run PCA on the data we need to generate a correlation or covariance matrix. We choose to use a covariance matrix in this example.

O3 {=covarm(C4:K391)}

	2y	3y	4y	5y	7y	10y	15y	20y	30y
2y	6.07	8.28	9.46	9.79	9.41	8.53	7.57	7.11	6.46
3y	8.28	11.87	13.73	14.36	14.02	12.85	11.52	10.88	9.97
4y	9.46	13.73	16.54	17.47	17.41	16.30	14.93	14.23	13.21
5y	9.79	14.36	17.47	19.13	19.41	18.52	17.26	16.60	15.59
7y	9.41	14.02	17.41	19.41	20.43	20.15	19.30	18.79	17.93
10y	8.53	12.85	16.30	18.52	20.15	20.57	20.21	19.92	19.29
15y	7.57	11.52	14.93	17.26	19.30	20.21	20.37	20.29	19.93
20y	7.11	10.88	14.23	16.60	18.79	19.92	20.29	20.36	20.12
30y	6.46	9.97	13.21	15.59	17.93	19.29	19.93	20.12	20.08

(https://asmquantmacro.com/wp-content/uploads/2015/06/shot2.jpg)

To make the calculations of a covariance matrix easier we use below custom array function that will loop through each data column and calculate pair wise covariance using excels built in COVAR function

(General) CovarM

```
Option Explicit

Function CovarM(Data_rng As Range) As Variant
    Dim i As Integer
    Dim j As Integer

    Dim numCols As Integer: numCols = Data_rng.Columns.Count
    Dim numRows As Integer: numRows = Data_rng.Rows.Count

    Dim matrixCov() As Double
    ReDim matrixCov(numCols - 1, numCols - 1)

        For i = 1 To numCols
            For j = 1 To numCols
                matrixCov(i - 1, j - 1) = _
                    Application.WorksheetFunction.Covar(Data_rng.Columns(i), Data_rng.Columns(j))
            Next j
        Next i
    CovarM = matrixCov
End Function
```

(https://asmquantmacro.com/wp-content/uploads/2015/06/vba.jpg)

We can then use =MEigenvecPow(OurCovarianceMatrix,TRUE) function from the Matrix.xla addin to generate the eigenvector of the covariance matrix. These values are often called loadings. Below are the results for our example.

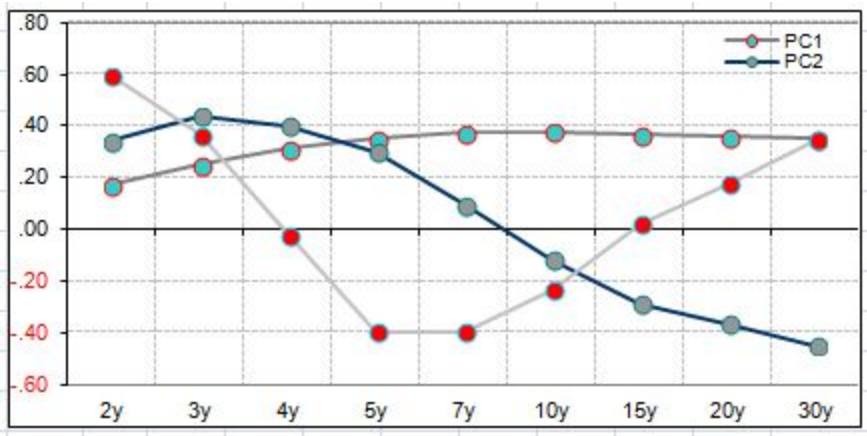
N16 fx {=MEigenvecPow(N3:V11,TRUE)}

		2y	3y	4y	5y	7y	10y	15y	20y	30y
2y		6.07	8.28	9.46	9.79	9.41	8.53	7.57	7.11	6.46
3y		8.28	11.87	13.73	14.36	14.02	12.85	11.52	10.88	9.97
4y		9.46	13.73	16.54	17.47	17.41	16.30	14.93	14.23	13.21
5y		9.79	14.36	17.47	19.13	19.41	18.52	17.26	16.60	15.59
7y		9.41	14.02	17.41	19.41	20.43	20.15	19.30	18.79	17.93
10y		8.53	12.85	16.30	18.52	20.15	20.57	20.21	19.92	19.29
15y		7.57	11.52	14.93	17.26	19.30	20.21	20.37	20.29	19.93
20y		7.11	10.88	14.23	16.60	18.79	19.92	20.29	20.36	20.12
30y		6.46	9.97	13.21	15.59	17.93	19.29	19.93	20.12	20.08
EigVal		141	13.57	0.747	0.222	0.142	0.114	0.02	0.019	0.011
		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
2y		.17	.34	.59	.50	.21	.45	.08	.02	.01
3y		.25	.44	.37	-.04	-.29	-.72	-.04	-.02	-.02
4y		.32	.40	-.02	-.69	-.21	.47	.00	.01	.00
5y		.36	.30	-.39	.05	.66	-.14	-.36	-.17	-.06
7y		.38	.10	-.40	.22	-.03	-.06	.67	.43	.01
10y		.38	-.12	-.23	.33	-.46	.13	-.16	-.47	.45
15y		.37	-.29	.02	.11	-.21	.07	-.12	-.09	-.83
20y		.36	-.36	.17	-.05	.02	-.01	-.45	.65	.27
30y		.35	-.45	.35	-.31	.36	-.12	.41	-.36	.16

(https://asmquantmacro.com/wp-content/uploads/2015/06/shot3.jpg)

Now it is time for the interpretation of the results. The loading for each factor give us the sensitivity of a particular variable to a 1 unit change in a given factor (principal component). For example, in the above, if the first principal component goes up by 1 then the 2yr swap rate will change by .17 bps, the 5yr will go up but .36bps, and 30yr swap will increase by.35 bps (this is the first column of the matrix). The second column gives us the loadings for the second factor (principal component). In this case, when the second principal component increases by 1, the short end of the curve will increase while the longer end will decrease. This just means that the curve flattens as the second principal component increases. Finally, when the third principal component increases, the short and long end of the curve increases while the middle points of the curve decrease.

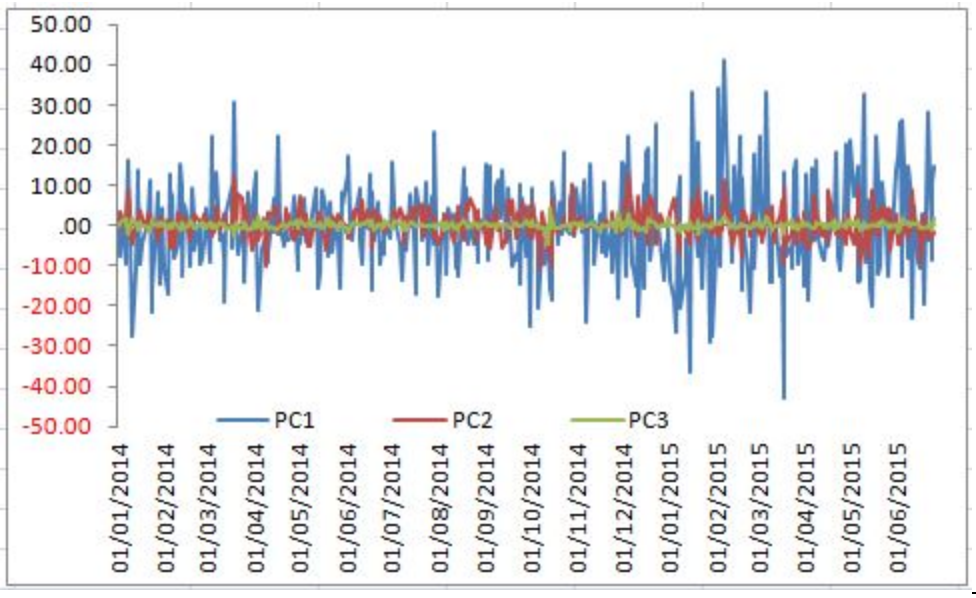
When we plot the loadings we can see the data better.



(https://asmquantmacro.com/wp-content/uploads/2015/06/shot4.jpg)

This shows us that the first component captures mostly parallel yield curve moves, the second captures the slope, while the third captures the curvature (butterfly).

So far we spoke about changes in principal components. We would like to know what value they actually take. This is easy; each principal component is a linear combination of the original data and the loadings. So for example, using above data, on 26 Jun2015 the first principal component is equal to 14.70 [.17*4.18 +.25*2.67+.32*3.47 +.36*4.28+.38*5.18+ .38*5.48 +.37*6.02+.36*6.05+.35*6.34]. Calculating a time series of the first three principal components we can see that they are indeed uncorrelated (orthogonal)



(https://asmquantmacro.com/wp-content/uploads/2015/06/shot5.jpg)

Now we would like to answer the obvious question, why did we stop at three principal components in our discussion above. The answer is that three components account for 99.7% of the variation in the data. How can we compute that number? We can use the eigenvalues of our covariance/correlation matrix. To compute these we use MEigenvalPow(OurCovarianceMatrix) from the matrix.xla addin. Below (green row) presents our results.

O13 =O14/SUM($N$14:$V$14)

	L	M	N	O	P	Q	R	S	T	U	V
1											
2		Covar	2y	3y	4y	5y	7y	10y	15y	20y	30y
3		2y	6.07	8.28	9.46	9.79	9.41	8.53	7.57	7.11	6.46
4		3y	8.28	11.87	13.73	14.36	14.02	12.85	11.52	10.88	9.97
5		4y	9.46	13.73	16.54	17.47	17.41	16.30	14.93	14.23	13.21
6		5y	9.79	14.36	17.47	19.13	19.41	18.52	17.26	16.60	15.59
7		7y	9.41	14.02	17.41	19.41	20.43	20.15	19.30	18.79	17.93
8		10y	8.53	12.85	16.30	18.52	20.15	20.57	20.21	19.92	19.29
9		15y	7.57	11.52	14.93	17.26	19.30	20.21	20.37	20.29	19.93
10		20y	7.11	10.88	14.23	16.60	18.79	19.92	20.29	20.36	20.12
11		30y	6.46	9.97	13.21	15.59	17.93	19.29	19.93	20.12	20.08
12		CumVar	90.4%	99.2%	99.7%	99.8%	99.9%	100.0%	100.0%	100.0%	100.0%
13		Var	90.4%	8.7%	0.5%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%
14		EigVal	140.57	13.57	0.747	0.222	0.1422	0.11355	0.02048	0.01855	0.01102
15			PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
16		2y	.17	.34	.59	.50	.21	.45	.08	.02	.01
17		3y	.25	.44	.37	-.04	-.29	-.72	-.04	-.02	-.02
18		4y	.32	.40	-.02	-.69	-.21	.47	.00	.01	.00
19		5y	.36	.30	-.39	.05	.66	-.14	-.36	-.17	-.06
20		7y	.38	.10	-.40	.22	-.03	-.06	.67	.43	.01
21		10y	.38	-.12	-.23	.33	-.46	.13	-.16	-.47	.45
22		15y	.37	-.29	.02	.11	-.21	.07	-.12	-.09	-.83
23		20y	.36	-.36	.17	-.05	.02	-.01	-.45	.65	.27
24		30y	.35	-.45	.35	-.31	.36	-.12	.41	-.36	.16

(https://asmquantmacro.com/wp-content/uploads/2015/06/shot6.jpg)

To assign meaning to these values and compute the percentage of variation that each principal component explains we need to do the following; Take the sum of all eigenvalues. In our example the sum across the green row is 155.41. We can now divide the first eigenvalue by 155.41 to get 90.4%. This means the first principal component explains 90.4% of the variation in the data. The second component captures 8.7% [13.57/155.41]. We can see that in total the first three principal components explain approximately 99.7% of the variation in the data. Adding more factors doesn't add to our understanding of the data.

We wish to come back to our main point that we mentioned at the start. PCA is used to represent the original data as a function of a reduced number of factors. In our case that means each change in yield for a chosen swap tenor is a function of three factors. So, for example, on any given day the change in 30yr swap is a given by its loadings times the principal components. On 26 June 2015 the first principal component was 14.70, the second principal component was -1.65 and the third was 1.71. From above table of loadings we see that the loadings of 30yr tenor for the first three principal components are .35,

-.45, .35. Taking the sum of products we get 6.48 [(14.7*.35)+(-1.65*-.45)+.(1.71*.35)]. This means that we can expect the 30yr swap rate to increase by 6.48 bps given the change in the first three principal components that we witnessed. The actual change on June 26 2015 was 6.34bps.

In this post we tried to present an intuitive explanation of Principal Component Analysis. In follow up posts we will discuss the many uses of PCA in managing risk, modelling asset prices, and trading.

Some useful resources:

1) Market Risk Analysis Volume 1 by Carol Alexander: http://www.amazon.com/Market-Analysis-Quantitative-Methods-Finance/dp/0470998008/ref=sr_1_2?s=books&ie=UTF8&qid=1435483909&sr=1-2&keywords=market+risk+analysis (http://www.amazon.com/Market-Analysis-Quantitative-Methods-Finance/dp/0470998008/ref=sr_1_2?s=books&ie=UTF8&qid=1435483909&sr=1-2&keywords=market+risk+analysis)

Posted in Excel, VBA

---

12 thoughts on “Principal Component Analysis in Excel ~ PART I”

1. Pingback: Principal Component Analysis in Excel ~ PART III |
2. ***tim*** says: March 29, 2016 at 3:15 am
   This very helpful for a project I’m working on. I’ve a simple question: is there a quick way to calculate the time series for each of the first three principal components or is it the tedious process of calculating the covariance matrix and eigenvectors for each date? Thank you.

   Reply

   bquanttrading says: March 29, 2016 at 3:33 am
   hey tim, there sure is. take the matrix of all the swap rate changes (size NxP) where N is the number of observations and P is the number of tenors. Multiply that by the first eigenvector (Px1) and you will have a time series of the first principal component (size Px1). in excel you can use MMULT(rate_change_matrix,eigenvector). for the first three principal components just include the first three eigenvectors MMULT(rate_change_matrix,3_eigenvectors). hope that helps.

   Reply
3. ***Daniel Wyczolkowski*** says: April 3, 2016 at 3:20 pm
   Near the end of this article, ” On 26 June 2015 the first principal component was 14.70, the second principal component was -1.65 and the third was 1.71.” Could you please explain the method by which you arrived at these values. I can’t for the life of me see it in the snips of excel sheets that you have included.

   Reply

   bquanttrading says: April 3, 2016 at 10:04 pm

earlier in the post i mention that "each principal component is a linear combination of the original data and the loadings." i also gave an example of the calculation just below that line. exact same approach was used to calculate PC value for 26June. sum the product of range n16:n24 and c4:k4 to get 1st pc for 26june

Reply

4. ***Daniel*** says: April 4, 2016 at 4:28 pm
   Thanks for the quick reply. I was thrown off by the calculation in the middle of the text because it stated the PC for "Jun 28th" and the data ended on Jun 26th. I now see that this was just a typo.

   Reply

   bquanttrading says: April 4, 2016 at 5:03 pm
   fat fingers. thats fixed now. thanks for spotting the typo

   Reply

5. ***Ole*** says: November 14, 2017 at 9:06 pm
   The link http://excellaneous.com/Downloads.html is no longer active.
   Would you post it again, please? 🙂
   Cheers

   Reply

6. ***Allan*** says: August 4, 2019 at 5:40 am
   I find the add-in here: https://www.bowdoin.edu/~rdelevie/excellaneous/#downloads

   Reply

7. ***Rob B. (@rballant)*** says: November 14, 2019 at 3:26 pm
   Is there a reason the CovarM function can't be dragged down/over after Ctrl+Shift+Enter? With the range locked, I'm getting the VarCov(1,1) element. With the range unlocked, I get #VALUE!.
   Thoughts?

   Reply

8. ***AO*** says: March 13, 2021 at 10:42 am
   Hello,

   This was one of the most useful and practical articles I found on PCAs. I'm curently working on a project for my PhD thesis and what I need is to calculate VaR for a fixed income portofolio taking into account yield curve scenarios built using PCA (that are historically plausible and of plausible magnitude). I read Golub (Risk Management for fixed income markets) who provides some valuable insights, but I still find it hard to put in practice, like in designing excel worksheets. Can you please help me?

   Reply

9. ***Anonymous*** says: October 1, 2024 at 11:39 pm
   Hi profesor

   i used he downloading link for PCA xls but the link do"t work. can you send me another?

   thank you

   Reply