Principal Component Analysis

Applications in Finance

September 2024

Guest Lecturer: Stefan Andreev

- Quantitative Research in Fixed Income
- B.A. in Physical Chemistry from Dartmouth College
- Ph.D. in Chemical Physics from Harvard University
- 18 years experience in Finance as a quant
- Email:
- Finance Experience:
  - 2006-2014: Fixed Income Desk Strat at **Morgan Stanley**
  - 2014-2019: Quantitative Researcher at Global Fixed Income Fund at **Citadel**
  - 2020- : Quantitative Researcher at Fixed Income RV Fund at **Two Sigma**

Machine Learning Algorithms *(sample)*

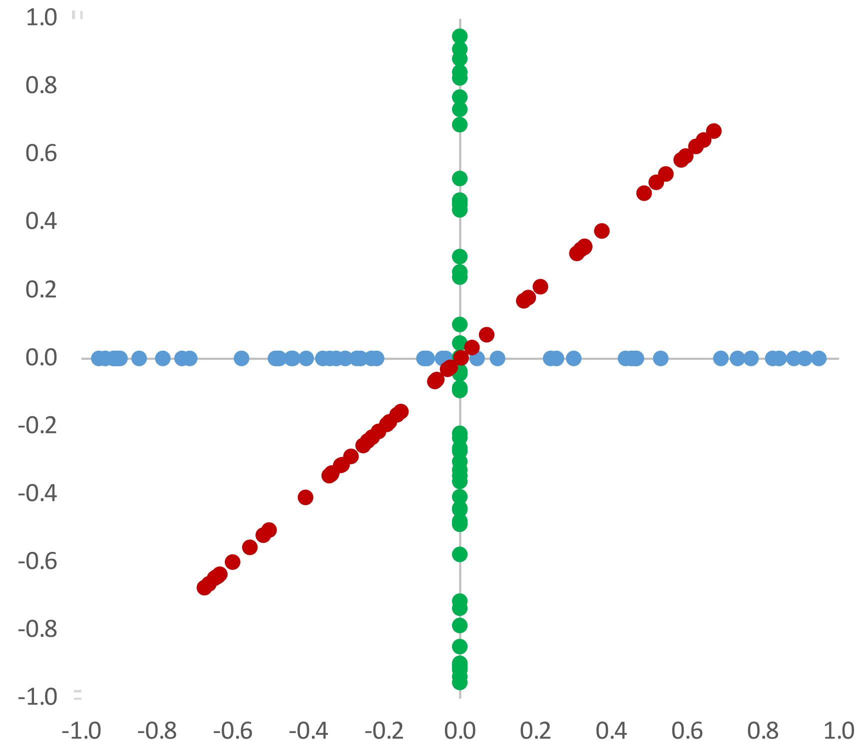
	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

Principal Component Analysis (PCA)

- Applied on large datasets of multidimensional data
- Goal: Find the linear combinations of input variables that describe most of the variance of the dataset
- It can be used to extract the main few drivers of variance in a dataset
  - This helps to make a linear model of the dataset with much fewer dimensions
  - Helps focus on the signal, not the noise
  - Can reduce complexity of analysis dramatically
- Used across data science
- Examples outside of finance:
  - Face Recognition
  - Weather modeling
  - Genetics data mining
- Examples in finance:
  - Analyzing correlations of asset returns
  - Empirical asset clustering
  - Regime change detection
  - Yield curve modeling
  - Index replication
  - Returns replications

Reducing dimensionality – Simplest Case

- Three 2d datasets
  - Visually they are 1 dimensional
    - Can be described by a single variable m
  - Green – described by y coordinate
    - $m = y$
  - Blue – described by x coordinate
    - $m = x$
  - Red – described by a linear combination of x and y
    - $m = x + y$
- Effectively the data is 1 dimensional in all three cases
- Principal Component Analysis applied to the Red dataset would tell us
  - The data is effectively 1 dimensional
  - How to find the equation for m
- Nomenclature
  - m is the first principal component (PC1)
  - The coefficients a, b in $m = ax + by$ are called the loadings of the principal component



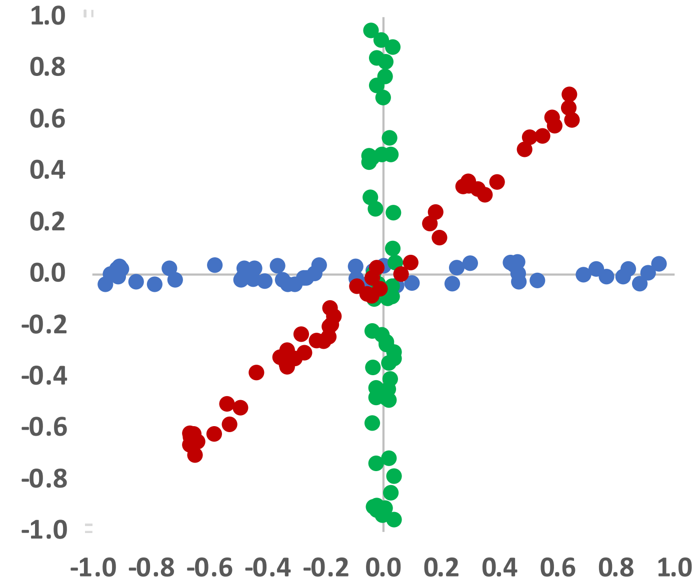
Reducing dimensionality – Mostly 1D

- Datasets in Plot A are 2D
  - Most variance is in a single direction
- PCA identifies the principal direction of the variance of the data
  - The first principal component p_1 of each dataset is the same as in Plot B

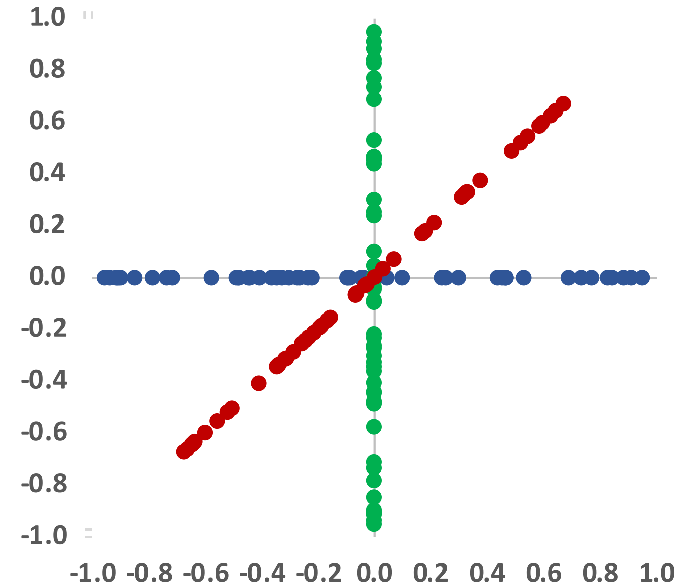
	Green	Blue	Red
PC1	y	x	$x + y$
PC2	x	y	$x - y$

- Plot B = PC1 projection of each Plot A dataset
  - The “other” direction is PC2
    - Much less important
- Used PCA to eliminate “noise” while keeping essential feature of data

A)



B)



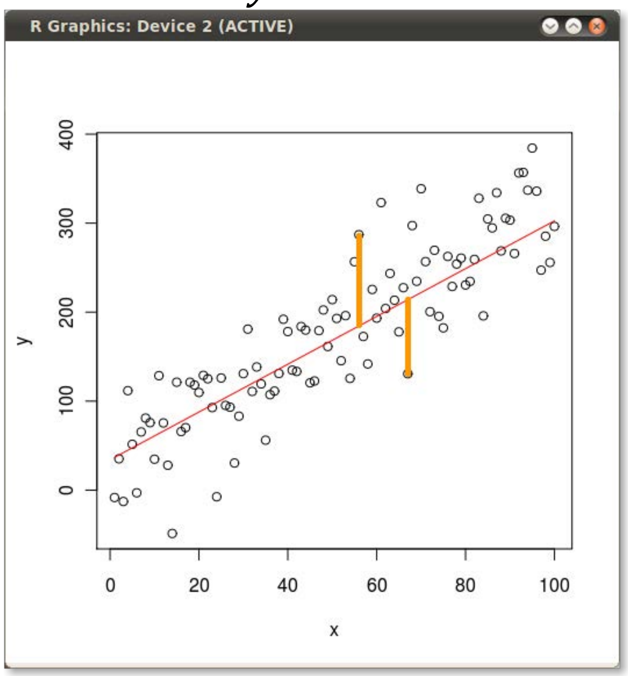
Power of PCA

- PCA can be generalized in arbitrary number of dimensions
  - Many pixels in an image
  - Time Series of returns of many stocks
- Computationally efficient
  - Efficient calculation through SVD
  - It is fast and scalable with the size of dataset and dimensionality
- PCA tells us
  - What are the principal components of a dataset
  - How much variance each principal component explains
    - Allows us to interpret the results intuitively
- Works best when a few principal components explain most of the variance of a highly dimensional dataset
  - PC1 is typically (but not always) the most important factor to analyze
- Low variance PC factors are often due to noise
  - Allows us to eliminate noise and extract a simpler picture out of a complex dataset
  - In finance, the noise is often what we are after, as it may represent trading opportunity!

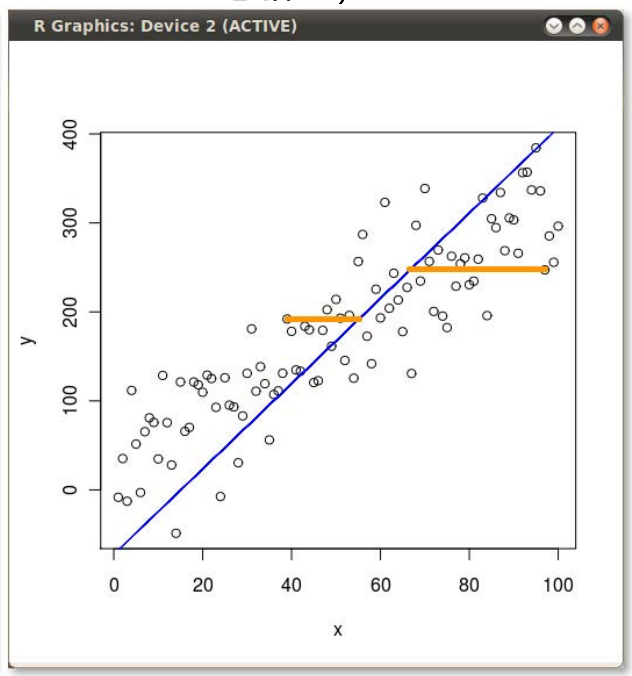
PCA vs. Ordinary Least Squares (OLS)

- OLS (aka regression) minimizes the error in predicting a dependent variable based on an independent variable
  - Necessary to choose a dependent and independent variables – prior assumption of causation!
  - $y{\sim}x$: Minimize error in $y_i = \alpha x_i + \epsilon_i$
- $y{\sim}x$ not same as $x{\sim}y$!
- Both OLS and PCA are useful, but serve a slightly different purpose
- Source/Code: http://www.cerebralmastication.com/2010/09/principal-component-analysis-pca-vs-ordinary-least-squares-ols-a-visual-explination/

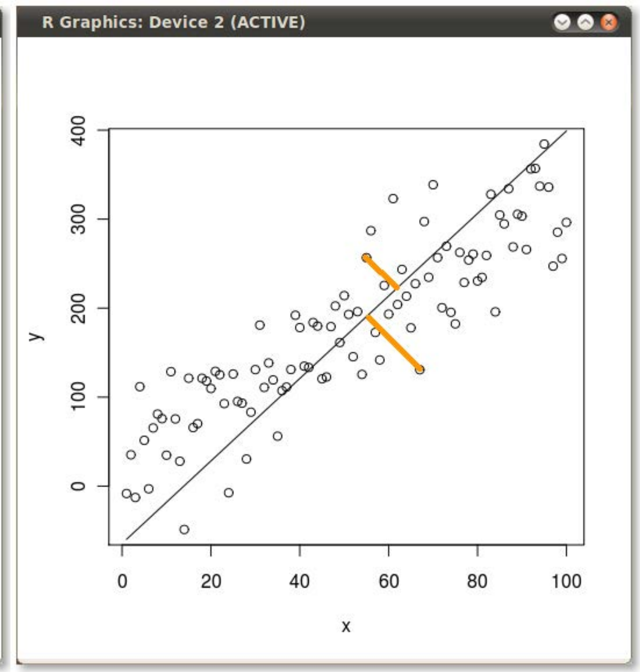
A: $y{\sim}x$



B: $x{\sim}y$



C: PC1 line



PCA Intuition

- PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.
- Intuitively, PCA is a change of coordinates, or a coordinate rotation
- Visual interactive demonstration in 2D at: http://setosa.io/ev/principal-component-analysis/

PCA Derivation: Math Overview

- Notation
  - X: $N \times p$ matrix containing N observation of p-dimensional data
    - $X = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$, where x_i is a p-dimensional vector corresponding to the i^{th} observation
    - $\bar{x} = \frac{1}{N}\sum_i x_i = 0$, i.e. the data is demeaned in all p dimensions
    - $\Sigma = X^T X$ is the $p \times p$ covariance matrix of the observed data. Should be replaced with sample covariance matrix when X is a sample of a larger set
  - w_k: The p-dimensional orthonormal vector representing the direction of principal component k
  - $x_i \cdot w_k = \sum_{j=1}^{p} x_{ij} w_{kj}$ is the projection of the i^{th} observation onto the k^{th} principal component
  - $Var(Xw_k) = w_k^T X^T X w_k = w_k^T \Sigma w_k$ is the variance of X along the k^{th} principal component
  - W: eigenvector matrix whose columns are w_k
  - Λ: pxp diagonal matrix with eigenvalues on the diagonal
- Efficient Computation
  - Computation of covariance matrix for a sample X is $\mathrm{O}(N^2 p^2)$.
  - SVD: $X = UMW^T$ where U is $N \times p$ and $U^T U = I$ is $O(Np^2)$
  - Relation to PCA: $\Sigma = X^T X = W\mathrm{M}^T U^T U M W^T = W M^T M W^T = W \Lambda W^T$
    - M is diagonal with $\sqrt{\lambda_k}$ on the diagonal. W is the eigenvectors of X
    - Much more efficient when $N \gg p$, which is usually the case for meaningful results in any case.

PCA Math: properties

- $W^{-1} = W^T$ -- orthonormality of eigenvectors
- $\Sigma = W\Lambda W^T$ -- SVD decomposition
- Total variance: $Var(X) = tr(\Sigma) = tr(\Lambda)$
- % Variance explained (in-sample): $\dfrac{W^T \Sigma W}{tr(\Sigma)} = \dfrac{\Lambda}{tr(\Lambda)}$
- Out of sample variance explained
  - Σ was estimated over a sample X. How well does the eigendecomposition of Σ explain the variance of a new sample?
  - New sample Y with covariance $\Sigma^Y = Y^T Y = W^Y \Lambda^Y (W^Y)^{\mathrm{T}}$
    - The new eigenvectors may be different. Define $D = (W^Y)^{\mathrm{T}}\mathrm{W}$
  - the out of sample variance explained on eigenvectors W is:

$$\frac{Var(YW)}{tr(\Sigma^Y)} = \frac{\mathrm{W}^{\mathrm{T}}\Sigma^Y W}{tr(\Sigma^Y)} = \frac{W^T W^Y \Lambda^Y (W^Y)^{\mathrm{T}}\mathrm{W}}{\mathrm{tr}(\Lambda^{\mathrm{Y}})} = \frac{D^T \Lambda^Y D}{tr(\Lambda^Y)}$$

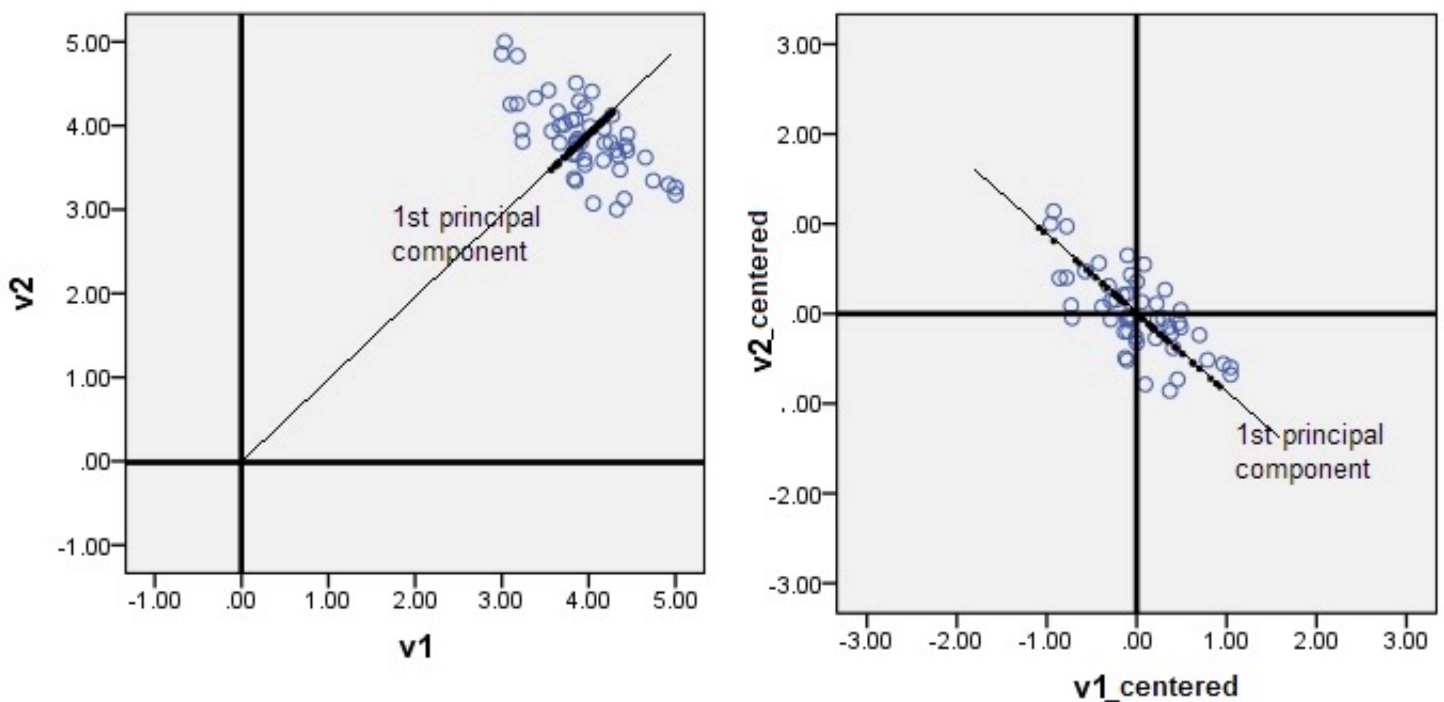
  - To the extent the W^Y are different from W, D will distribute variance to lower ranked PC factors, i.e. the top factors of W won't be as efficient in explaining variance of Y

Summary of typical workflow

- Transform p-dimensional data to a state that is as close to a multivariate normal distribution as possible
  - For a multivariate normal distribution, the mean and the covariance matrix fully characterize the whole distribution
- Subtract the mean from each axis and (maybe) normalize the variance in each direction to 1
  - How to, if at all, do the variance normalization is one of the hyperparameters of PCA
- Compute the covariance matrix C of size $p\times p$
  - It is symmetric, real, and positive semi-definite by construction
    - Can use SVD directly on the data for eigendecomposition as well
- Calculate the eigenvectors E_i and the eigenvalues λ_i of C
  - Order by decreasing eigenvalue: $\lambda_i \geq \lambda_{i+1}$
  - The eigenvector E_i defines the i coordinate of the new rotated principal component coordinate system
  - E_i contains the loadings of the i_{th} principal component
  - $\frac{\lambda_i}{\Sigma\lambda_i}$ is the percentage of total variance explained by the i_{th} principal component

Data Preparation: Demeaning the data

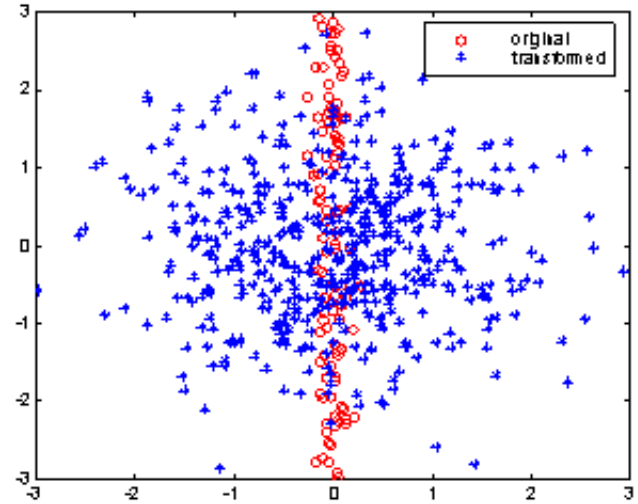
- PCA is a coordinate rotation – rotates data around the origin!
- It is crucial to demean data as the first step:
- $x^{c}_{i_{(k)}} = x_{i_{(k)}} - \overline{x_i}$
- If data is not centered at the origin, the results would not capture the maximum variance directions in the first principal component!
- Potential problem in time series analysis if rolling mean is unstable!



Source: http://stats.stackexchange.com/questions/22329/how-does-centering-the-data-get-rid-of-the-intercept-in-regression-and-pca

Data Preparation: Standardize Variance

- This is an optional but important step
- Standardize variance: scale each coordinate to unit variance
  - $x^c_{i(k)} = \frac{x_{i(k)}}{\sigma(x_i)}$
- Without standardization of variance, PCA is done on covariance matrix
  - Useful if all dimensions have the same units
  - Resulting PC1 maximizes actual variance explanation
  - If there is a large variance mismatch, PC1 will be dominated by the large variance dimension
- With standardization of variance, PCA is done on correlation matrix
  - Used when units of the dimensions are different
  - Resulting PC1 maximizes correlation explanation
- Choice is made depending on the context of the problem



Principal Component calculations

- Eigenvector and Eigenvalue calculation
  - Many software implementations. See accompanying notebook for Python implementation.
- Not supported natively in Excel
  - Can use the matrix.xla excel add-in to add support
- Supported in many numerical platforms
  - R
  - Python (numpy)
  - Matlab
- Analysis of data structure:
  - How fast do eigenvalues decay – how many components do you need to capture most of variance

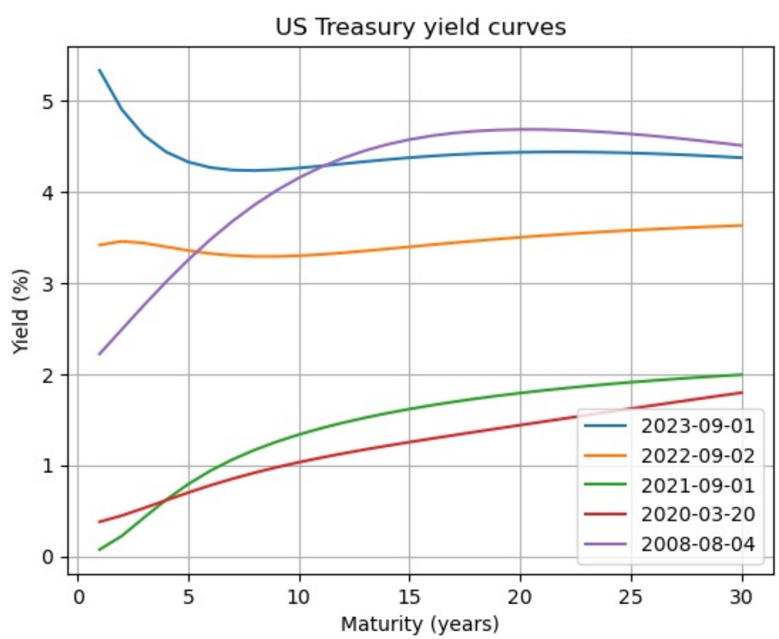
PCA on Time Series

- In Finance, we often apply PCA on time series of financial data
- Typical examples are PCA on financial asset returns or tradeable indicators
- Important questions on PCA for time series
  - Each observation is a data point
  - Dimensionality is the number of observed variables at each time stamp
  - Choosing the correct time window is important
  - Typically we want to use historical covariances to estimate future covariances
    - How robust are the answers w.r.t. the time window choice
  - Can perform the PCA on a rolling time window
    - PC1 loadings will vary through time
    - Variability can point to instability, making predictions less accurate
    - Sharp changes through time coupled with stable periods can be indications of regime change

Example: US bond market

- US Treasuries available at a wide range of maturities of up to 30 years
- Very liquid instruments – easily tradeable in large size
- Long history
- Bond yield – the average return you get holding a bond to maturity
  - The lower the price, the higher the yield
- Bond yields as a function of maturity are highly correlated
- Returns are proportional to changes in yields
- A very common are of application of PCA in practice

Liquid Points on the curve:
- New issues (monthly)
  - 2Y, 3Y, 5Y, 7Y, 10Y, 20Y, 30Y
- Bond Futures (CME)
  - TU (1.5Y)
  - FV (4Y)
  - TY (7Y)
  - UXY (10Y)
  - US (15Y)
  - WN (25Y)

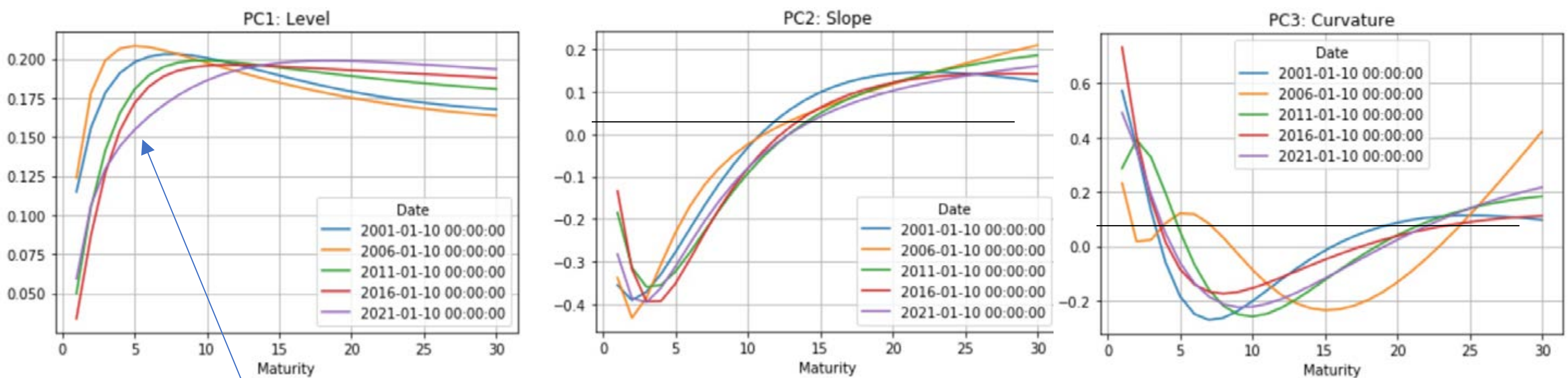


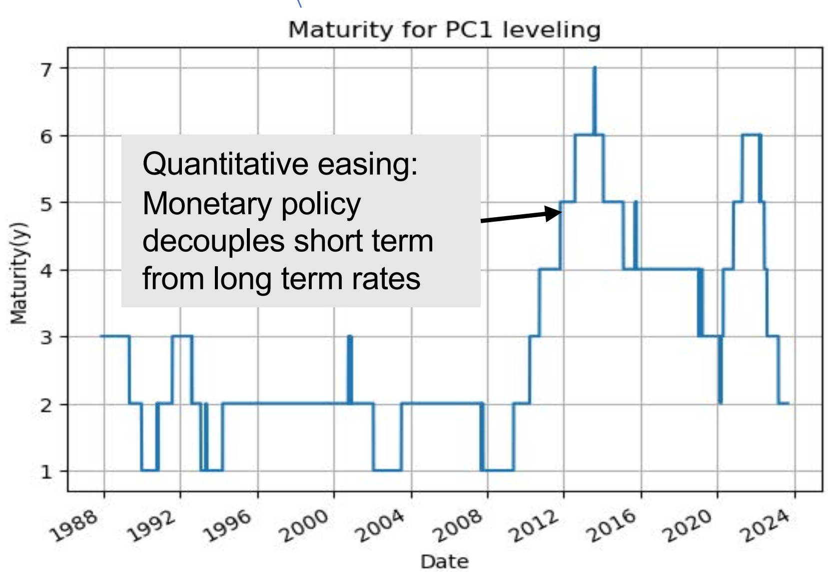
tenor	2	3	5	7	10	30
tenor						
2	100.0	98.4	92.7	87.2	81.0	68.3
3	98.4	100.0	97.3	92.8	87.0	74.8
5	92.7	97.3	100.0	98.6	94.6	83.8
7	87.2	92.8	98.6	100.0	98.5	88.9
10	81.0	87.0	94.6	98.5	100.0	92.4
30	68.3	74.8	83.8	88.9	92.4	100.0

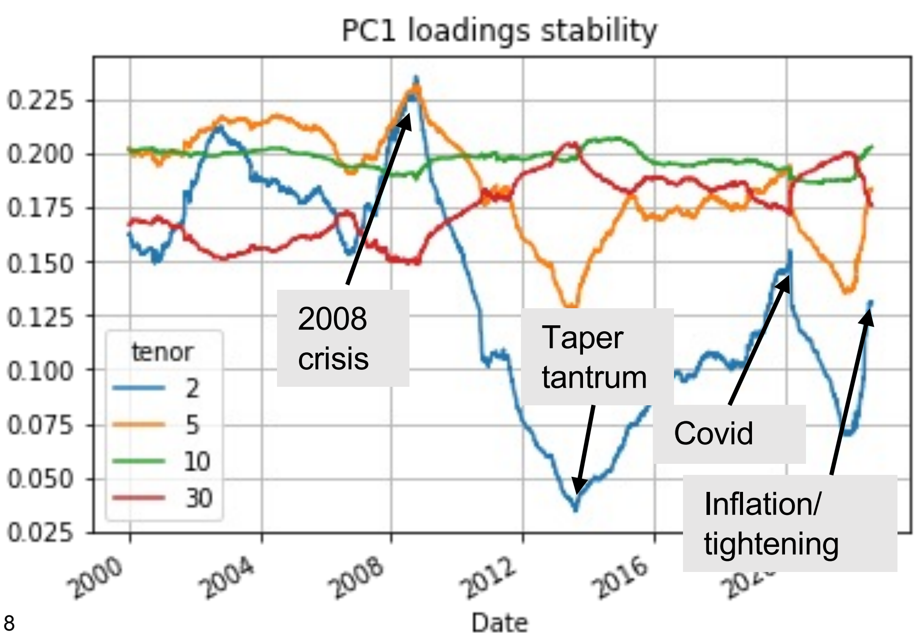
Data from Federal Reserve:
https://www.federalreserve.gov/data/nominal-yield-curve.htm

PCA of yield changes: Macroeconomic meaning

- PCA on yield changes of all maturities
- Rolling 2y historical lookback for covariance estimation – Choice of window and weighting scheme is crucial in practice
- 3 Curve regions: Front-end (0-5), Belly(5-15), Long end (15-30)







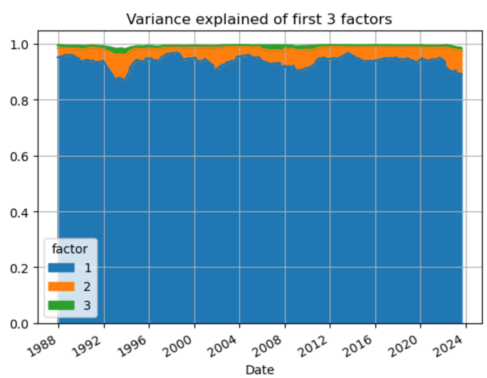
Variance explained in and out of sample

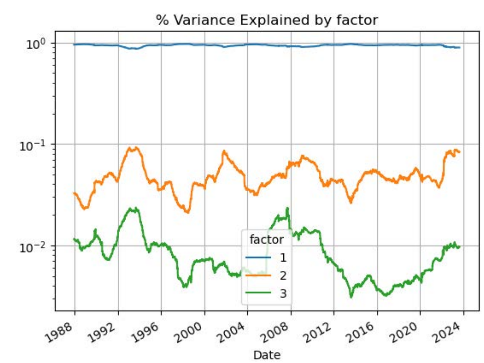
- Top 3 factors explain 99% of variance in-sample
- Out of sample variance explain can drop especially around crises
  - Represents pc-coordinate rotation
  - Represents trading behavior / yield curve dynamics change!

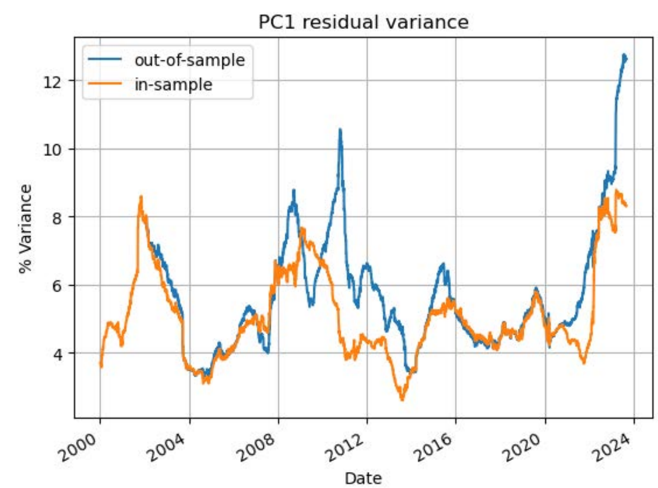
PC1/PC2 out of sample correlation is substantial!

Out of sample PC factor covariance

	1	2	3
1	100.000000	7.080478	-1.594103
2	7.080478	100.000000	-7.292835
3	-1.594103	-7.292835	100.000000







PC Portfolios

Bond Math Notation
Buy $1 dv01 of bond =
buy notional N s.t. if the bond yield goes up by 1 basis point (0.01%), the position earns $1

- PCs as portfolios:
  - Eigenvectors are weighted portfolios of the columns of X
  - $P = XW$, or $X = PW^T$ is the transformation into principal component space
- Trading pc residuals:
  - **PC1 residuals** through PC neutral curves: given 2 features $[a, b]$, find α_b s.t. the portfolio $\mathrm{P} = a - \alpha b$ has 0 weight on the first principal components. Solution is $\alpha_b = W_{a,0}/W_{b,0}$
    - *Example*: Curve 2s10s steepener: sell $1 dv01 of 10y bond, buy α dv01 of 2y bond. Portfolio profits when the curve steepens / in pc2 moves, and is neutral to pc1 moves
  - **PC2 residuals** through PC neutral flies: given 3 features $[a, b, c]$, find α_a, α_b s.t. the portfolio $P = \alpha_a a + \alpha_b b - c$ has 0 weight on first two principal components.
  - $\begin{bmatrix} W_{a,0} & W_{b,0} \\ W_{a,1} & W_{b,0} \end{bmatrix} \begin{matrix} \alpha_a \\ \alpha_b \end{matrix} = \begin{matrix} W_{c,0} \\ W_{c,1} \end{matrix}$
    - *Example*: belly outperformance macro 2s10s30s fly: buy $1 dv01 of 10y bond, sell α_2 of 2y bond and α_{30} dv01 of 30y bond
  - Constructing pc-neutral portfolios is very common application in fixed income
    - Betting on pc residual portfolios

PC residual portfolios

Pros

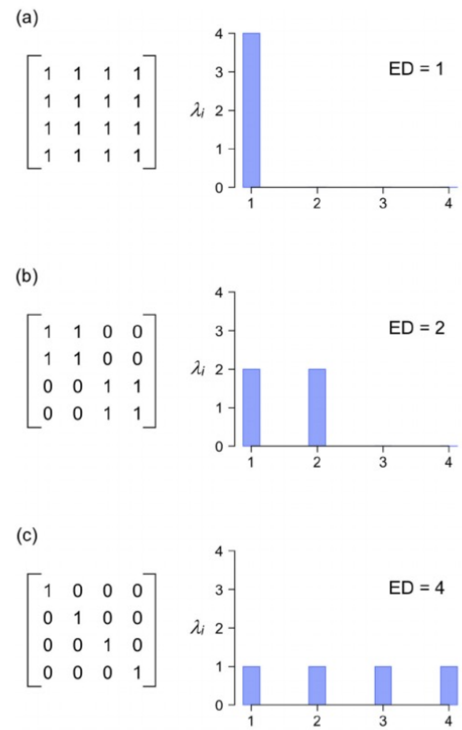
- Allows for trading portfolios unrelated to the main market move direction (PC1)
- Adds diversity to the range of possible effective uncorrelated bets
- Volatility per unit notional is much lower – any mispricing / relative value opportunity might have bigger signal/noise ratio

Cons

- Transaction cost / volatility ratio is much higher
- Portfolio weights are not constant and are model dependent
- Leverage/balance sheet: Because volatility per notional is much less, need much higher notional traded to achieve the same target volatility
  - Requires financing of the bond and long positions, i.e. borrowing money from banks, which is typically done in overnight markets
  - This can cause friction and lending ability can be severely curtailed in times of high stress

PC Portfolio Diversity: Definition[1]

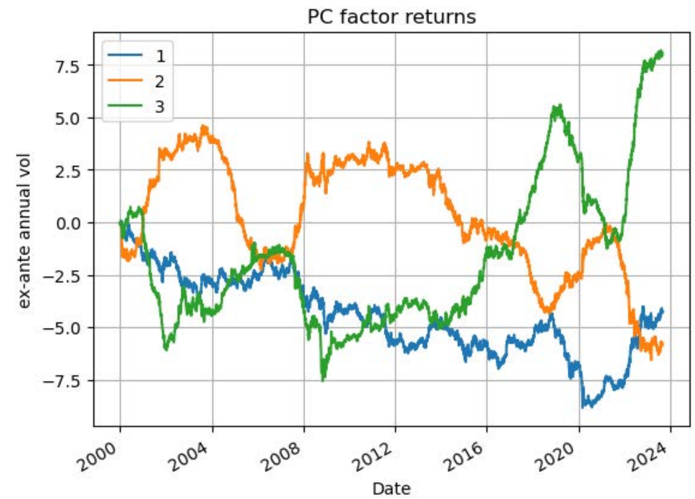
- *Goal*: Calculate number of effective orthogonal directions (ED) in a correlation matrix
  - Eigenvalues are $\lambda_i,\ p_i = \frac{\lambda_i}{\sum \lambda_i}$
- Quadratic diversity: $\mathrm{ED} = \left(\sum_i \lambda_i\right)^2 \Big/ \sum_i {\lambda_i}^2$
  - Corresponds to "effective number of independent measurements"
  - Derivation based on measuring the quadratic entropy of a distribution



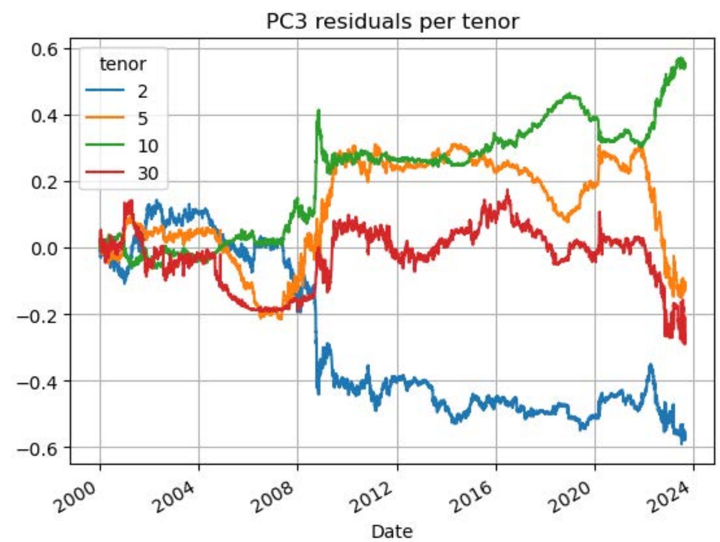
1. Marco Del Giudice, March 2020: Effective Dimensionality: A Tutorial

PC Portfolio Diversity: US yield curve

- PC residuals increase correlation matrix ED
- Higher order PC residuals tend to have mean reverting properties

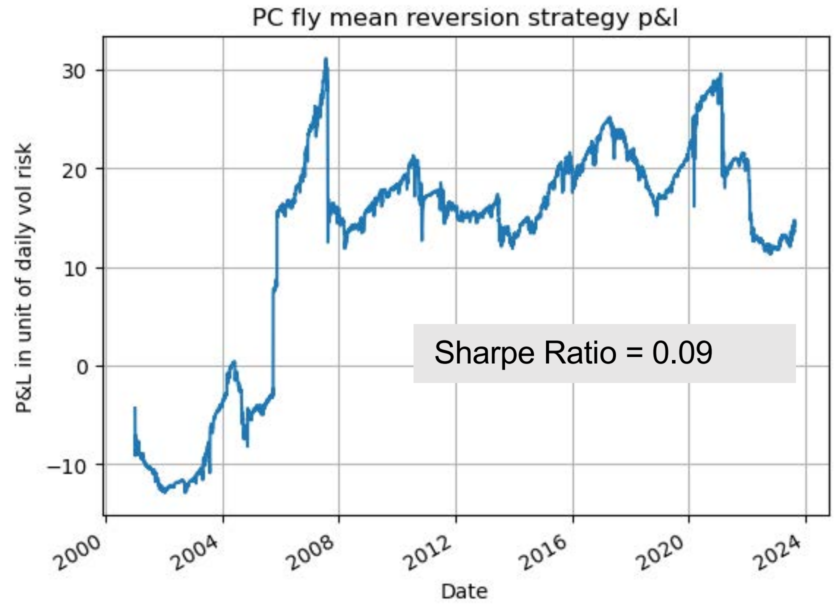


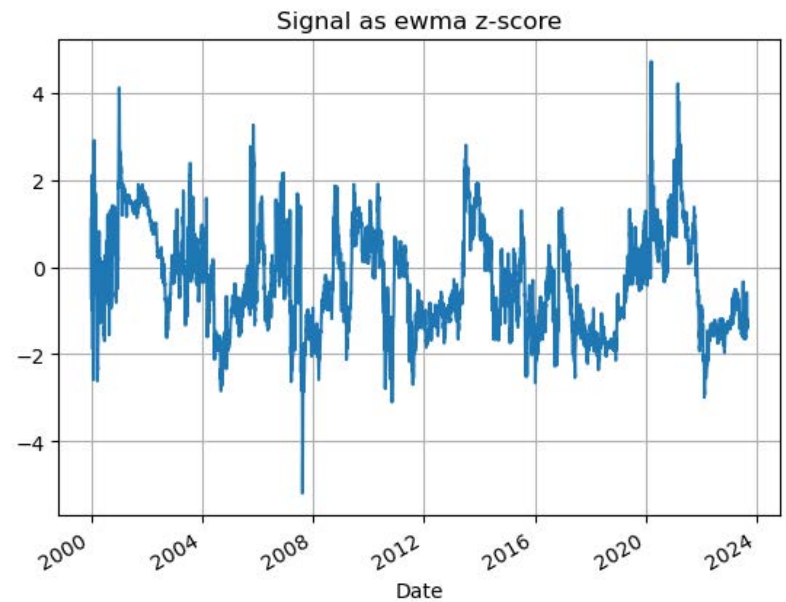
Portfolio	Effective Diversity (ED)
PC1	1
PC1+PC2	1.99
PC1+PC2+PC3	2.97
PC1+PC2+PC3+Residuals	5.28



PC-Neutral fly mean reversion strategy

- Construct pc-neutral fly based on PC 1 and PC 2 loadings
- Notice that returns are very mean reverting around a long term trend
- Construct signal: s = z-score of cumulative returns vs recent EWMA mean and std
  - Signal is delayed by a day to ensure actionability
- Position = -signal





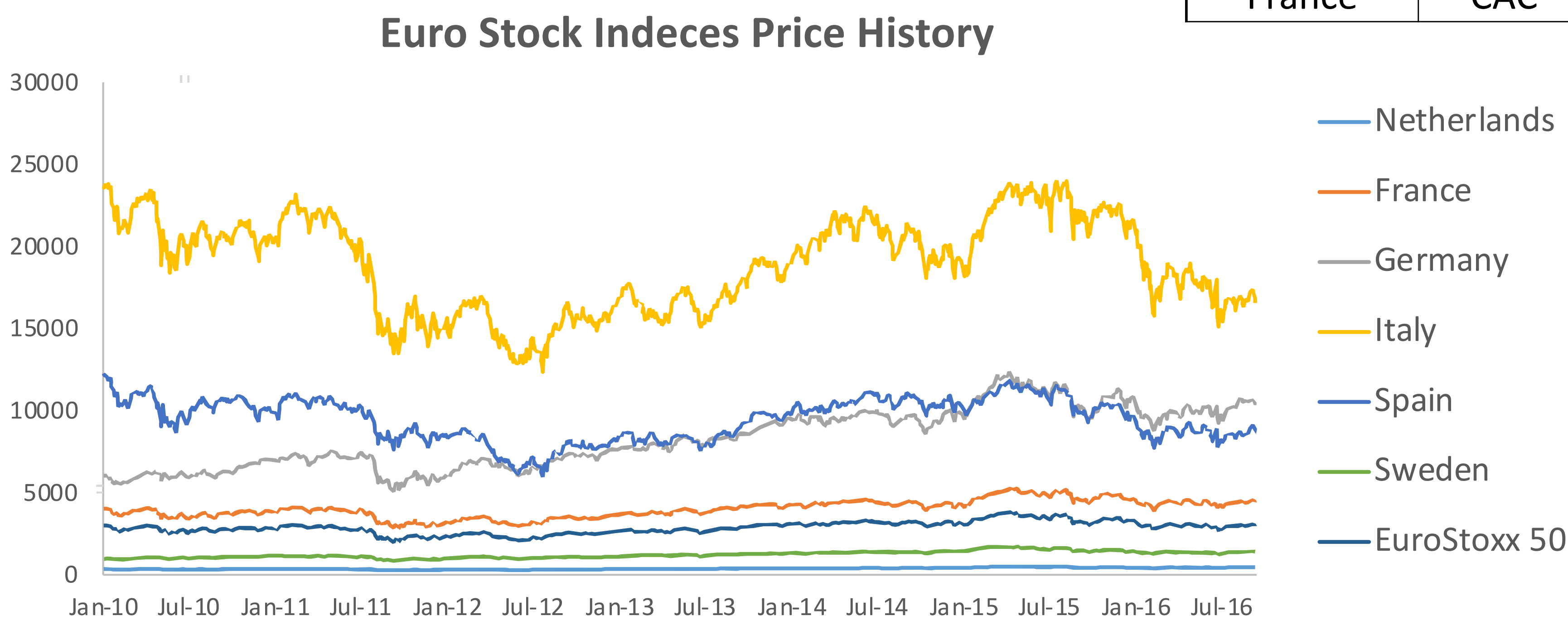
Potential project

- Strategy research on the Fed US treasury data presented here
  - PC neutral flies – diversify by constructing mean reversion strategies of many flies
  - Carry-rolldown – compute rolldown of a fly
  - Combine signals and flies to achieve a better performing portfolio
  - Compute ex-post pc return decomposition to measure effectiveness of pc-neutralization
  - Investigate the effect of changing the lookback window (2y in this example) on out of sample variance prediction: is there an optimal lookback scheme?
- Practice coding in python/pandas using accompanying jupyter notebook

Example: European Stock Market Indeces

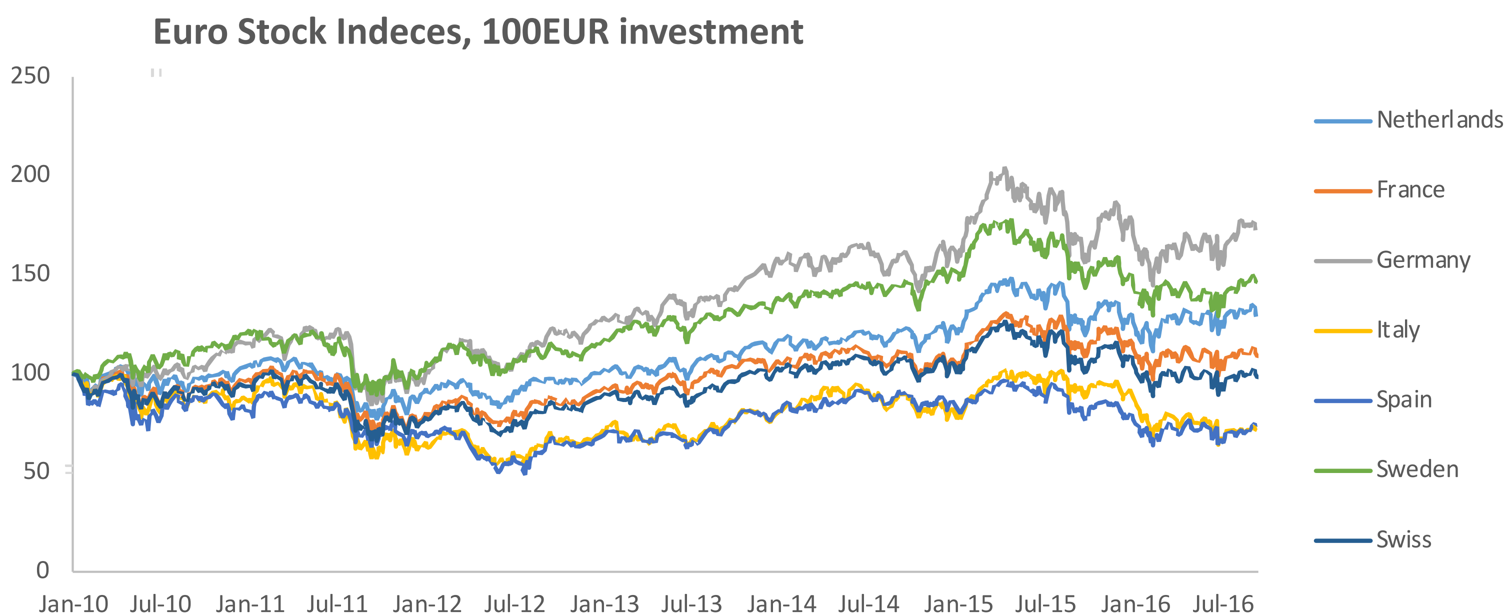
- Look at universe of 7 European stock market indeces
- Period: 1/4/2010 to 9/16/2016
- Part of the largest single market in the world
- Similar time zone – comparing daily closing marks makes sense
- Same currency: EUR

Country	**Index**
Netherlands	AEX
Germany	DAX
Italy	FTSEMIB
Spain	IBEX
Sweden	OMX
EuroStoxx 50	SX5E
France	CAC



Pre-Analysis data preparation

- Normalize history to a 100 EUR investment in 1/4/2016
  - Works because indeces are asset prices, hence represent theoretical investment vehicles
- Changes look correlated and mostly driven by one major factor: Good PCA candidate!
- Asset returns distributions are more lognormal than normal
  - Means percentage returns, or log(returns) are somewhat normally distributed
  - Hence Black-Scholes options pricing works well
  - PCA input would be a time series of log returns: diffs of log of price
  - $r_t = \ln(p_t) - \ln(p_{t-1}) = ln\left(\frac{p_t}{p_{t-1}}\right)$, where r_t is the log return for t, and p_t is the price at t



Covariance / Correlation matrices

- Demean daily log returns so that mean is zero
- Optionally rescale each index log returns to unit variance
- Calculate covariance matrix
  - Variance on diagonal
  - Covariance off-diagonal

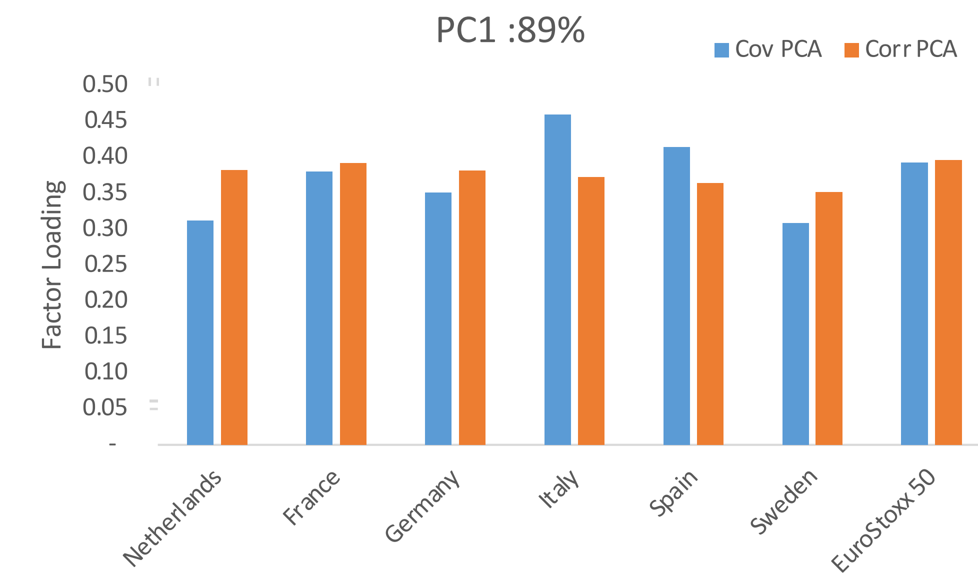
	AEX	DAX	FTSEMIB	IBEX	OMX	SX5E	CAC
AEX	1.4E-04	1.5E-04	1.4E-04	1.7E-04	1.5E-04	1.2E-04	1.5E-04
DAX	1.5E-04	1.9E-04	1.7E-04	2.1E-04	1.9E-04	1.5E-04	1.9E-04
FTSEMIB	1.4E-04	1.7E-04	1.7E-04	1.9E-04	1.7E-04	1.4E-04	1.8E-04
IBEX	1.7E-04	2.1E-04	1.9E-04	2.9E-04	2.4E-04	1.6E-04	2.2E-04
OMX	1.5E-04	1.9E-04	1.7E-04	2.4E-04	2.5E-04	1.4E-04	2.0E-04
SX5E	1.2E-04	1.5E-04	1.4E-04	1.6E-04	1.4E-04	1.6E-04	1.5E-04
CAC	1.5E-04	1.9E-04	1.8E-04	2.2E-04	2.0E-04	1.5E-04	2.0E-04

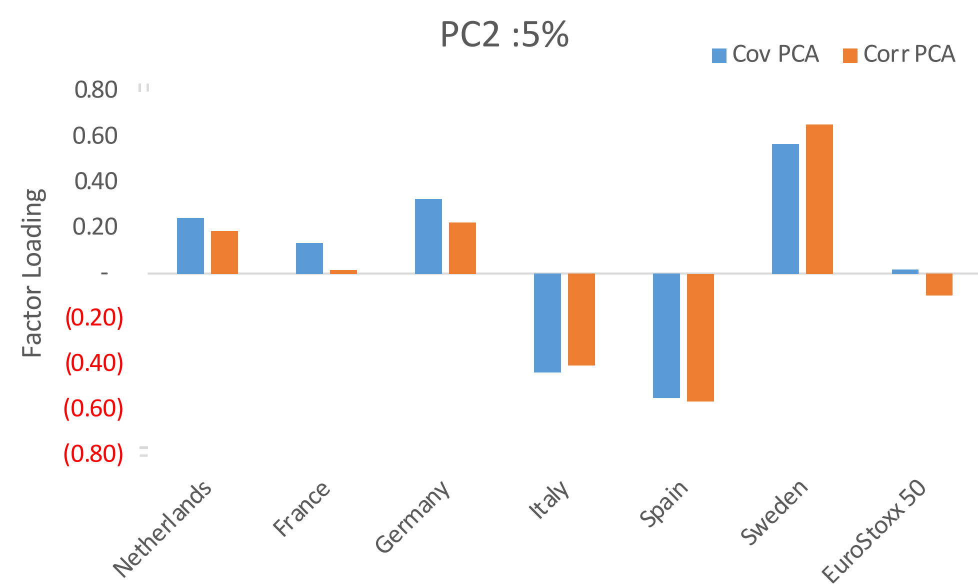
- Correlation matrix: $Cov(x, y) = Corr(x, y)\sqrt{Var(x)Var(y)}$

	AEX	DAX	FTSEMIB	IBEX	OMX	SX5E	CAC
AEX	100%	94%	92%	85%	82%	84%	94%
DAX	94%	100%	94%	89%	87%	84%	98%
FTSEMIB	92%	94%	100%	84%	80%	84%	95%
IBEX	85%	89%	84%	100%	89%	76%	92%
OMX	82%	87%	80%	89%	100%	72%	91%
SX5E	84%	84%	84%	76%	72%	100%	83%
CAC	94%	98%	95%	92%	91%	83%	100%

Principal Components

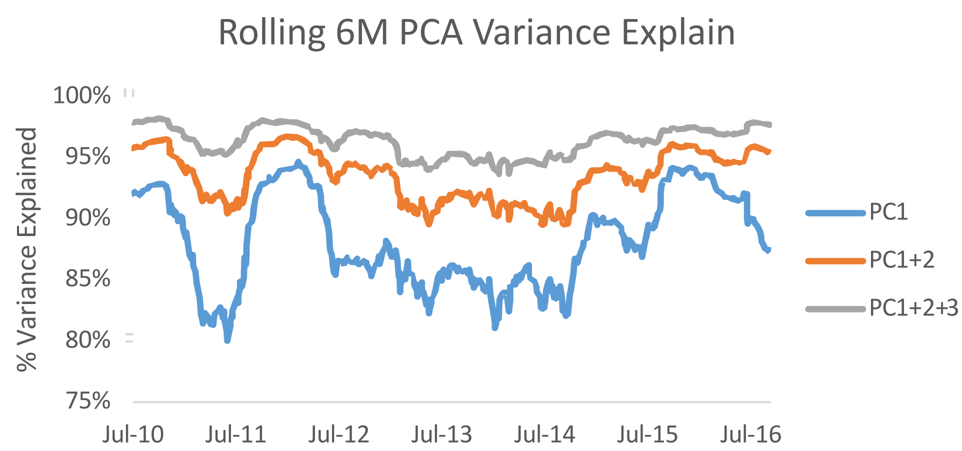
- First principal component
  - Explains 89% of daily return variance!
  - PC1 loadings approximately equal among indeces
  - To first order, all 7 euro indeces move together
  - First factor can be interpreted as “Euro Stocks” factor
    - Concept often seen in financial news
- Second principal component
  - Explains about half of remaining variance
  - Shows country differentiation
    - Northern vs Southern Europe divide!
    - France is neither!
- Covariance vs. Correlation matters little
  - Mostly because variances similar among indeces
- Sign does not matter!

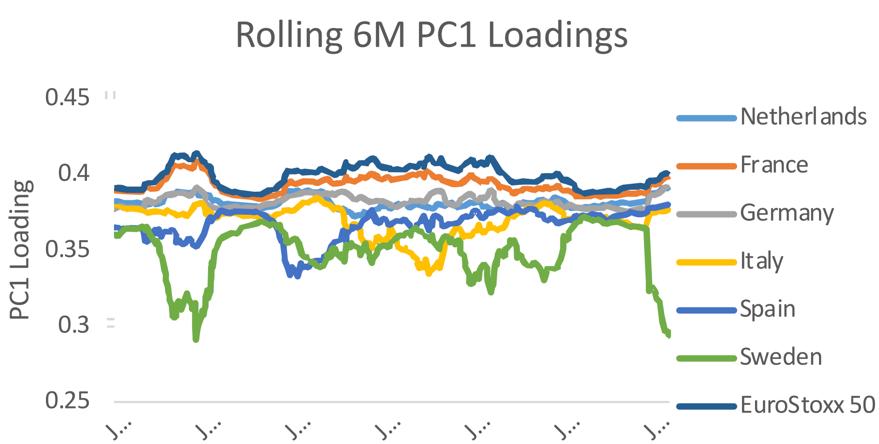


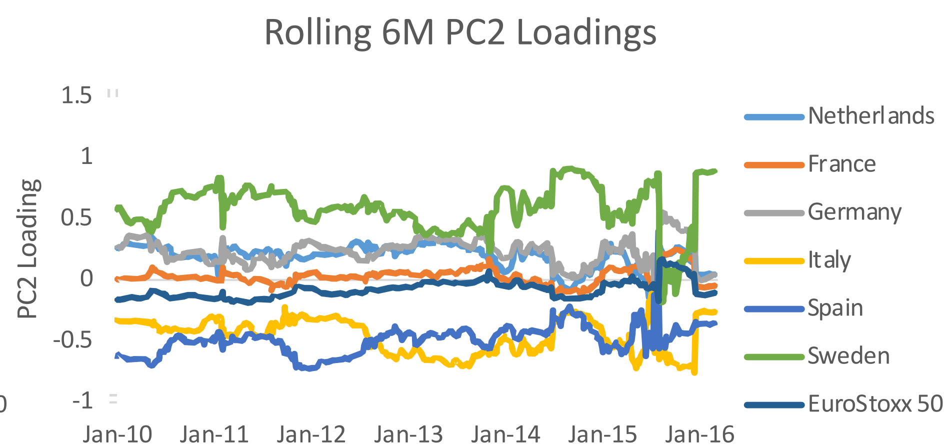


Stability/Robustness

- Rolling 6M PCA can uncover instabilities in the cumulative result
- Indications are that the PCA is robust
  - Variance explain in top 3 factors is consistently high and stable
  - PC1 loadings are quite stable
  - PC2 loadings also relatively stable
- PC analysis is robust and is likely to be robust and predictive

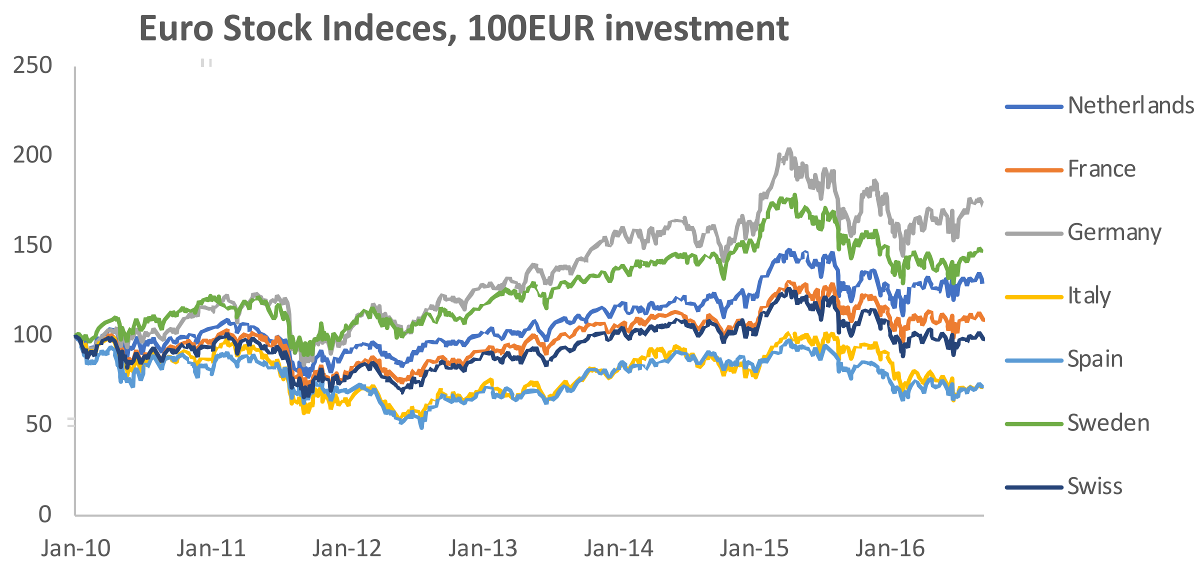


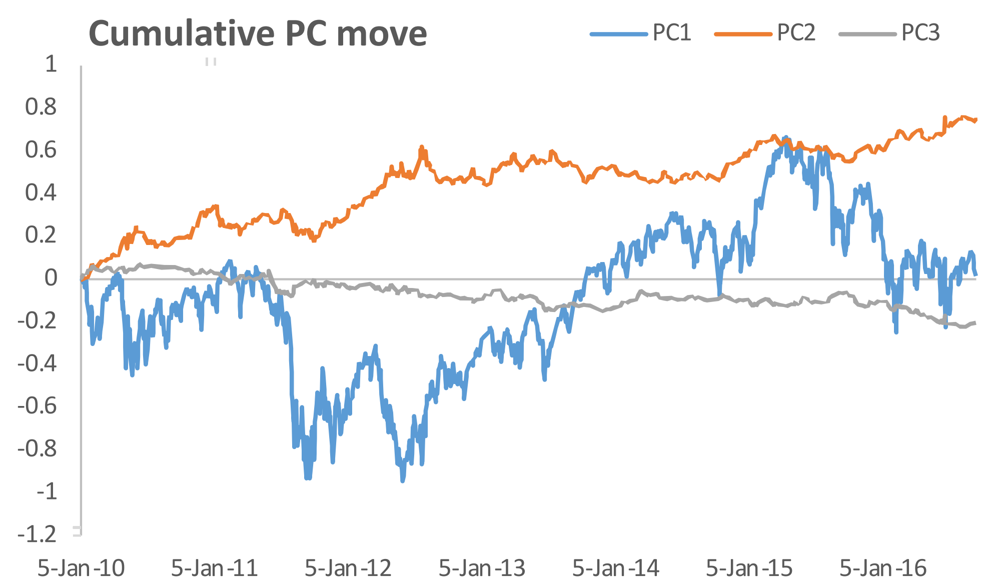




PCA portfolios

- A principal component portfolio is a basket of tradeable products where each one is weighted by the PC weight
  - PC1 portfolio is expected to have most of the variance
  - Sometimes PC2 or PC3 portfolios can have interesting properties like mean reversion that can be exploited
- Daily realizations for PC portfolios can be computed as
  - $\Delta PC_t^i = \Delta X_t \cdot PC_i$
- In our example
  - PC1 realizations reflect the general trend across European markets
  - PC2 reflects the divergence between Northern and Southern European stock markets



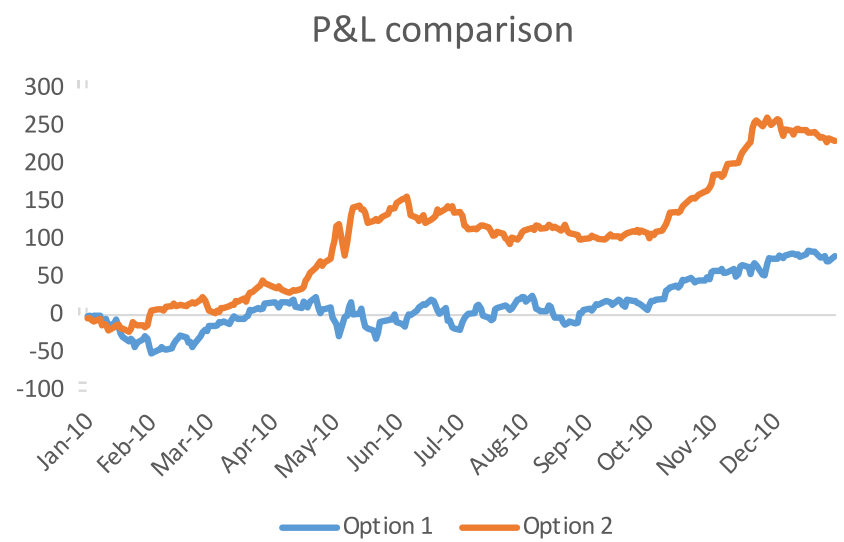


PCA for hedging

- The investment process (especially relative value strategies), usually blends a variety of strategies in an overall portfolio, seeking diversification
  - The resulting total portfolio may have unintended overall exposure
- PCA allows investor to hedge volatile components that do not contribute alpha
- For portfolio with weights V across components, the exposure to n-th PC factor is $V \cdot w_n$
- To eliminate exposure to first PC factor, we can construct $V' = V - (V \cdot w_1)w_1$
  - The resulting portfolio will be PC1-neutral, i.e. $V' \cdot w_1 = 0$
  - The stdev of resulting portfolio $= {V'}^T \Sigma V'$
  - If the alpha of the strategy is largely orthogonal to PC1, this should preserve the alpha while reducing volatility
    - Results in higher Sharpe ratio – can invest and make more money for the same amount of risk

PCA Hedging example

- Investment parameters
  - Thesis: German companies will outperform their peers in Europe
  - Size: Target 100mm euro annualized daily volatility
  - Period: Jan 1, 2010 – Jan 1, 2011
- Option 1: Buy German stock index (DAX)
  - DAX annualized vol = 20%
  - Size: 500mm EUR = 100mm / 0.2
  - Return: 75mm EUR, Sharpe Ratio = 0.75
- Option 2: Construct PC1 neutral portfolio
  - PC1 neutral annualized vol = 6.7%
  - Size: 1,500mm EUR = 100mm / 0.067
  - Return: 230mm EUR, Sharpe Ratio = 2.3



	PC1 weights	Target	PC1 neutr	Position (mm)	Return (mm)	Return (%
Netherlan	0.383	0	(0.14)	(215)	(11)	5%
France	0.390	0	(0.15)	(219)	6	-3%
Germany	0.377	1	0.86	1,280	199	16%
Italy	0.379	0	(0.14)	(213)	28	-13%
Spain	0.365	0	(0.14)	(205)	38	-19%
Sweden	0.361	0	(0.14)	(203)	(45)	22%
EuroStoxx	0.391	0	(0.15)	(220)	13	-6%

PCA Conclusions

- Strong evidence of a single underlying factor driving most of the returns across European stock markets
  - Very high explanatory power for a single factor
  - Consistent with the common notion of Euro Stocks as a single factor
- In PC2, strong evidence of North vs. South European differentiation
  - Approximately half of ex-PC1 variance explained
  - Consistent with the notion that there are cases Northern and Southern European countries have divergent interests
- Possible uses (a very non-exclusive list)
  - Produce PC-neutral portfolio
  - Use it for a North vs. South trading strategy
    - PC1 weights give you the proper way to hedge out PC1 exposure
  - Frankfurt exchange opens earlier than Spain exchange – predict and trade by proxy Spain market just before it opens
  - PC1 provides a good set of weights to construct a basket of stocks that captures the “Euro Stocks” factor

PCA Limitations

- PCA would always give you a result, but robustness can be an issue
  - Important to study robustness
    - Robustness to rolling time windows in time series PCA
    - Robustness to varying the number of factors going into the PCA
- While not necessary, transforming input variables into something as close as possible to a multivariable normal distribution helps
  - If the data is not described well by variance along a cardinal direction at the data center of mass, results can be misleading
- PCA does not imply a causal relationship (as opposed to regression)
  - This can be good and bad, depending on the problem
  - For trying to predict a dependent variable based on independent variables, regression can be more appropriate

Ideas for Final Project

- Perform PCA on the top 15 stock indeces by market value from 2003 to today
  - Use Google Sheets to gather data
- Compare robustness of factors for daily, 3 day, and weekly returns
  - How do time zones affect your analysis
- How has the variance explained evolved? Can you discern any long term effects of the financial crisis?
- Are the loadings of the first couple of factors meaningful?
- Welcome to discuss ideas for PCA on other multivariable datasets where you think the correlations are interesting and meaningful

Legal Notice

- Nothing in this presentation constitutes investment, legal, accounting, or tax advice.

18.642 Topics in Mathematics with Applications in Finance
Fall 2024