SYSTEMATIC

~~ALGORITHMIC~~ TRADING. MTH9894

Lecture 1
Statistical Arbitrage Framework

Dmitry Rakhlin, PhD

Goldman Sachs Asset Management, Quantitative Investment Strategies

917-343-5355

Dmitry.Rakhlin@gs.com

COURSE STRUCTURE

SCOPE: OVERVIEW OF POPULAR THEMES IN QUANT TRADING

Lecture 1 (March 23th) Statistical Arbitrage

- Review of Avellaneda's and Meucci frameworks
- Course projects assignment

Lecture 2-3-4 (March 30th, April 6th, April 27th) Fundamental Quant Models

- Part 1: Common market anomalies,
- Part 2: More anomalies. Portfolio construction.
- Part 3: The role of factor decay. Cost of strategy implementation. Performance evaluation.

Lecture 5 (May 4th) Modern Agency Algo Strategies

- A bit of history. Modern algos. SORs
- Performance evaluation and all kinds of TCA

Lecture 6 (May 11th) High Frequency Trading Pt.1 (by Jarrod Yuster – CEO, Pico Trading)

- Raison d'être, infrastructure design and requirements
- Current state of research & market regulations.

Lecture 7 (May 18th) High Frequency Trading Pt.2 – Impact on Markets. Final project review

- Current market ecosystem. Who are liquidity takers and liquidity providers?
- Interaction of market participants. Another look at how trading impact investment performance.

Lecture 8 (May 25th) Course project presentation

Final Projects

BUILD A BETTER STRATEGY. DEFEND IT.

- Research, replicate and (hopefully) improve existing investment strategy
- Working in groups of 2 or 3
- You may use any framework (R, Matlab, Python, Java …). The code will be reviewed
- **April 6th:** a list of suggested strategies will be given. Examples:
  a. Avelllaneda/Meucci statarb framework
  b. Value/Growth/Quality quant models
  c. Market Sentiment / Sector Rotation model
- Feel free to suggest your own strategy.
  a. Pre-requisite: there should be a publication/white paper/research paper
  b. I will need to review it before approving
- **April 27th:** Working groups have to be finalized. Strategies are clearly stated.
- **May 18th:** Progress review. Groups are given 3-4 min to update/ask questions.
- **May 25th:** Final presentations. Each group will have 15 min to present the results.

Time Series vs. Cross-Sectional Trading Strategies

- **Time Series**
  - A. Statistical Arbitrage
  - B. HFT
  - C. Agency Algos

  Features:
  - Each stock (or small group of stocks) is evaluated and traded independently of the rest of tradable universe
  - ETFs/Futures are used primarily to hedge the exposures

  Pros:
  - Take advantage of asynchronous arrival of data

  Cons:
  - **Typically** doesn't evaluate an opportunity within full context of risk drivers. Hedging is basic
  - Almost by design is focused on mean-reverting high-turnover strategies. (Exceptions are less popular technical momentum indicators – breakouts, crossovers….)

- **Cross-Sectional Regressions (some short-term statarb use them)**
- **Hybrid solutions (these are used as well)**

Time Series vs. Cross-Sectional Trading Strategies

Fama-MacBeth 2 stage regression

1. Use time series data to regress each asset against the proposed risk factors to determine that asset's beta (β^j) for that risk factor.

$$R_t^j - R^f = \alpha + \beta^j (R_t^m - R_t^f) + \epsilon_t^j$$

2. Use cross-sectional data to regress all asset returns for a fixed time period against the estimated betas (β^js) to determine the risk premium for each factor.

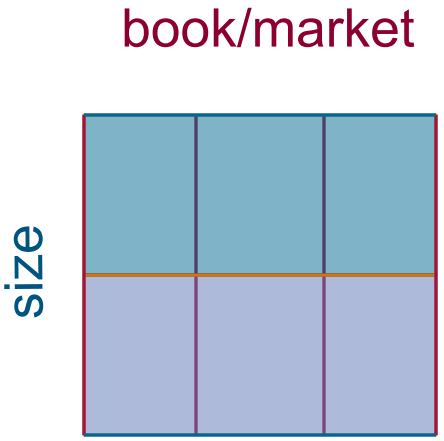
$$R^j_{\text{next month}} = a + b\hat{\beta}^j + e^j$$

b=market risk premium

Time Series vs. Cross-Sectional Trading Strategies

Fama French 3-Factor Model

- **Form 2x3 portfolios**
  - Size factor (SMB) – universe is split in 2 groups by size
    - Measure returns of small (portfolio) minus big (portfolio)
  - Book/Market factor (HML) – each group is split in 3 by B/M
    - Return of high (portfolio) minus low (portfolio)



- **For …** $R_t^j - R_t^f = \alpha^p + \beta^p(R_t^m - R_t^f)$

αs are big and βs do not vary much

- **For …** $R_t^p - R_t^f = \alpha^p + \beta^p(R_t^m - R_t^f) + \gamma^p \mathrm{SMB}_t^p + \delta^p \mathrm{HML}_t^p$

(for each portfolio p using time series data)
αs are zero, coefficients significant, high R^2. (Fama-French 3-factor model has been shown to explain 90% of the returns of diversified portfolio, while CAPM explains only 70%). For 4 & 5 factor model (+momentum & quality)

http://www.econ.yale.edu/~af227/pdf/Buffett's%20Alpha%20-%20Frazzini,%20Kabiller%20and%20Pedersen.pdf

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2287202

Statistical Arbitrage Framework

Marco Avellaneda, Jeong-Hyun Lee (2008): http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1153505
Attilio Meucchi (2009): http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1404905

1. Rule-based strategy aiming to profit from temporary relative dislocations of related securities
2. Trading book is market neutral (zero beta)
3. Mechanism for generating excess return is statistical (e.g. necessarily based on historical observations)
4. Holding period from from minutes to months. Faster strategies generate high T-costs => market making infrastructure becomes essential

Two Stocks described by price series P_t, Q_t (Pair Trading):

$$ln(P_t/P_{t_0}) = \alpha(t - t_0) + \beta ln(Q_t/Q_{t_0}) + X_t$$

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t$$

X_t - stationary, mean-reversing process: $\sigma(X_t) < \sigma_{max}$ for any t

Statistical Arbitrage Framework

N securities:

$$\frac{dP_t}{P_t} = \alpha dt + \sum_{j=1}^{n} \beta_j F_t^{(j)} + dX_t \quad \Leftrightarrow \quad R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \tilde{R}_i$$

$F_t^{(j)}$ factor returns – returns of "benchmark" portfolio representing systematic factors (simplest case: ETF returns & we have a "collection of pairs").

Market neutrality: $\overline{\beta}_j = \sum_{i=1}^{N} \beta_{ij} Q_i = 0, \quad j = 1, 2, ..., m$

$\{Q_i\}_{i=1}^{N}$ - dollars amounts invested in each of the stocks

$\overline{\beta}_j$ - portfolio betas, i.e. projection of portfolio returns on different factors

$$\sum_{i=1}^{N} Q_i R_i = \sum_{i=1}^{N} Q_i \left[\sum_{j=1}^{m} \beta_{ij} F_j \right] + \sum_{i=1}^{N} Q_i \tilde{R}_i = \sum_{j=1}^{m} \left[\sum_{i=1}^{N} \beta_{ij} Q_i \right] F_j + \sum_{i=1}^{N} Q_i \tilde{R}_i$$

PCA decomposition

How to Define/Find Factors? (so That Residual Process is Mean-Reverting)

1. Normalized Returns:

$$Y_{ik} = \frac{R_{ik} - \overline{R}_i}{\overline{\sigma}_i}, \quad \overline{R}_i = \frac{1}{M}\sum_{k=1}^{M} R_{ik}, \quad \overline{\sigma}_i^2 = \frac{1}{M-1}\sum_{k=1}^{M}(R_{ik} - \overline{R}_i)^2$$

2. Need to determine empirical correlation matrix:

$$\rho_{ij} = \frac{1}{M-1}\sum_{k=1}^{M} Y_{ik}Y_{jk}, \; \rho_{ii} = \frac{1}{M-1}\sum_{k=1}^{M}(Y_{ik})^2 = \frac{1}{M-1}\frac{\sum_{k=1}^{M}(R_{ik} - \overline{R}_i)^2}{\overline{\sigma}_i^2} = 1$$

3. Calculate its spectrum (eigenvalues and eigenvectors)

$$N \geq \lambda_1 > \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_N \geq 0.$$

$$v^{(j)} = \left(v_1^{(j)}, ..., v_N^{(j)}\right), \quad j = 1, ..., N.$$

4. And use the smallest eigenvectors to calculate minimum variance portfolio

BUT….

Segway into Random Matrix Theory

L.LALOUX, P.CIZEAU, M.POTTERS, J.P.BOUCHAUD: math.nyu.edu/faculty/avellane/**Laloux**PCA.pdf

- ρ_{ij} has N(N-1)/2 entries that are determined from NT observations, so to large extent the structure of empirical correlation matrix is dominated by measurement noise
- Least risky portfolio has larger weights on eigenvectors with the smallest eigenvalues
- Smallest eigenvalues are the most sensitive to measurement noise. What to do?
- Compare spectrum of empirical matrix to the one composed of independent assets (assume that returns Y_{ik} are i.i.d. random variables)

$$\rho_C(\lambda) = \frac{1}{N}\frac{dn(\lambda)}{d\lambda} = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{max}-\lambda)(\lambda-\lambda_{min})}}{\lambda}$$

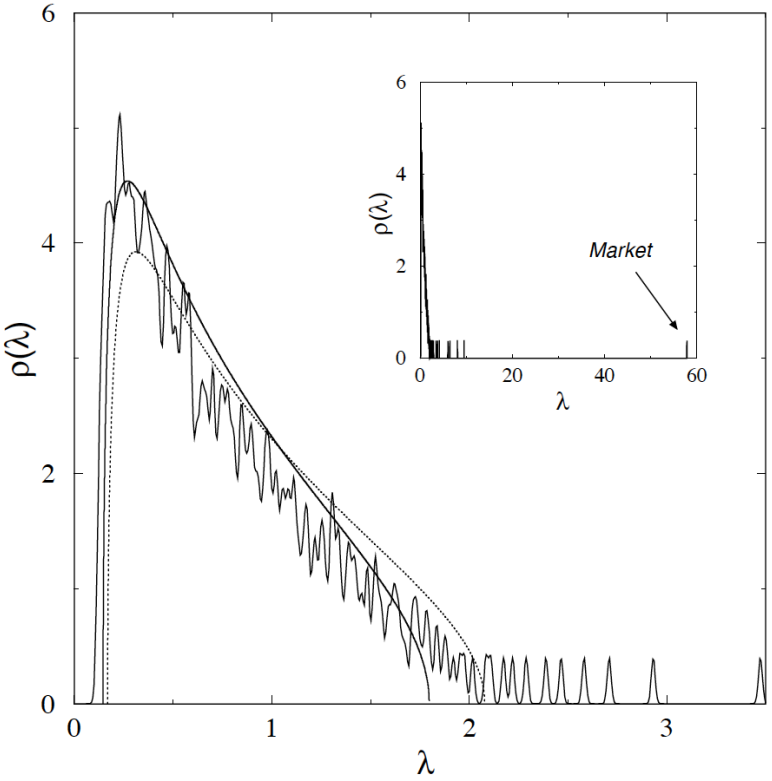
$$\lambda_{min}^{max} = \sigma^2(1+1/Q \pm 2\sqrt{1/Q}),\ \lambda \in [\lambda_{min}, \lambda_{max}], \sigma^2 = \text{var}(Y_{ik}) = 1$$

$$\text{in the limit } N \to \infty, T \to \infty \text{ and } Q = T/N \geq 1$$

Random Matrix Theory: Results

L.LALOUX, P.CIZEAU, M.POTTERS, J.P.BOUCHAUD: math.nyu.edu/faculty/avellane/**Laloux**PCA.pdf

1. No eigenvalues outside $[\lambda_{min}, \lambda_{max}]$, λ_{min} > 0 for *Q>1*
2. For finite *N* the "edges" are smoothed; the solution describing this blurriness is known



- SP500 daily returns 1991-96, *N*=406, *T*=1309
- "Market" eigenvector is positive for all stocks
- 94% of eigenvalues is in "noise" region
- Only 6% highest eigenvalues exceed λ_{max}, they explain 26% variance and carry information
- Empirical matrix can be "cleaned" by replacing noise band subspace by identity matrix (scaled to preserve matrix trace)
- "Clean" matrix still <u>underestimates</u> risk: optimizers "exploit" any amount of uncertainty in ρ_{ij} by biasing towards low-risk portfolios

PCA decomposition Cont'd

1. Need only keep *m*<*N* eigenvectors (above noise band)
2. "Clean" matrix with *trace(*$\overline{\rho}_{ii}$*)*=1 becomes

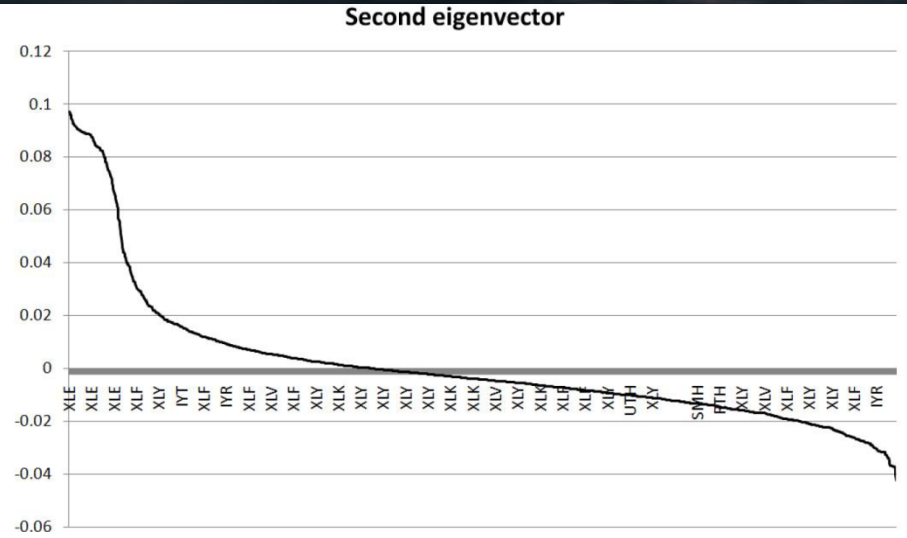
$$\overline{\rho}_{ij} = \sum_{k=0}^{m} \lambda_k \, v_i^{(k)} v_j^{(k)} + \epsilon_{ii}^2 \delta_{ij}, \quad \epsilon_{ii}^2 = 1 - \sum_{k=0}^{m} \lambda_k \, v_i^{(k)} v_i^{(k)}$$

3. Invest proportionally in each of *m* factors (consistent with market cap weighting) and then portfolio returns are:

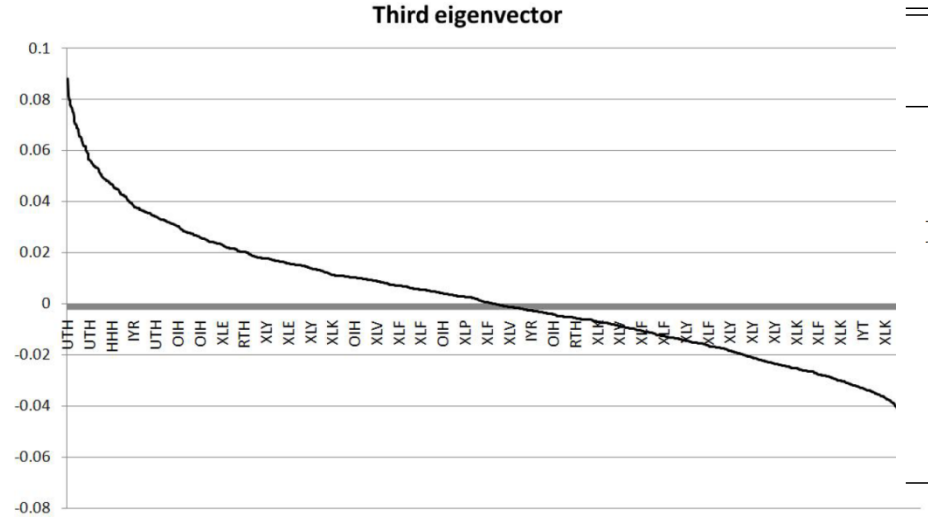
$$Q_i^{(j)} = \frac{v_i^{(j)}}{\overline{\sigma}_i} \qquad F_{jk} = \sum_{i=1}^{N} \frac{v_i^{(j)}}{\overline{\sigma}_i} R_{ik} \quad j = 1, 2, ..., m$$

Eigenvector Analysis

Top Ranking Eigenvectors *1<N<M* Represent Sector Rotations



Top 10 Stocks	Bottom 10 Stocks
Energy, oil and gas	Real estate, financials, airlines
Suncor Energy Inc.	American Airlines
Quicksilver Res.	United Airlines
XTO Energy	Marshall & Isley
Unit Corp.	Fifth Third Bancorp
Range Resources	BBT Corp.
Apache Corp.	Continental Airlines
Schlumberger	M & T Bank
Denbury Resources Inc.	Colgate-Palmolive Company
Marathon Oil Corp.	Target Corporation
Cabot Oil & Gas Corporation	Alaska Air Group, Inc.



Top 10 Stocks	Bottom 10 Stocks
Utility	Semiconductor
Energy Corp.	Arkansas Best Corp.
FPL Group, Inc.	National Semiconductor Corp.
DTE Energy Company	Lam Research Corp.
Pinnacle West Capital Corp.	Cymer, Inc.
The Southern Company	Intersil Corp.
Consolidated Edison, Inc.	KLA-Tencor Corp.
Allegheny Energy, Inc.	Fairchild Semiconductor International
Progress Energy, Inc.	Broadcom Corp.
PG&E Corporation	Cellcom Israel Ltd.
FirstEnergy Corp.	Leggett & Platt, Inc.

Alternative to PCA: Sector Decomposition

Sector	ETF	Num of Stocks	Market Cap unit: 1M/usd		
			Average	Max	Min
Internet	HHH	22	10,350	104,500	1,047
Real Estate	IYR	87	4,789	47,030	1,059
Transportation	IYT	46	4,575	49,910	1,089
Oil Exploration	OIH	42	7,059	71,660	1,010
Regional Banks	RKH	69	23,080	271,500	1,037
Retail	RTH	60	13,290	198,200	1,022
Semiconductors	SMH	55	7,303	117,300	1,033
Utility	UTH	75	7,320	41,890	1,049
Energy	XLE	75	17,800	432,200	1,035
Financial	XLF	210	9,960	187,600	1,000
Industrial	XLI	141	10,770	391,400	1,034
Technology	XLK	158	12,750	293,500	1,008
Consumer Staples	XLP	61	17,730	204,500	1,016
Healthcare	XLV	109	14,390	192,500	1,025
Consumer discretionary	XLY	207	8,204	104,500	1,007
Total		**1417**	**11,291**	**432,200**	**1,000**

Pros:

- More intuitive
- ETFs are tradable

Cons:

- Requires 3rd-party market classification
- ETF composition keeps changing
- Bias towards large-cap stocks

Clear Sector Theme in 2nd and 3rd Eigenvectors

Cyclical
- Basic Materials
- Consumer Cyclical (luxury items)
- Financial Services
- Real Estate

Defensive
- Consumer Defensive
- Healthcare
- Utilities

Sensitive
- Communication Services
- Energy
- Industrials
- Technology

Cyclical (beta >1)
The cyclical super sector includes industries significantly impacted by economic shifts.

Defensive (beta <1)
The defensive super sector includes industries that are relatively immune to economic cycles. These industries provide services that consumers require in both good and bad times

Sensitive (beta ≈ 1)
The sensitive super sector includes industries which ebb and flow with the overall economy, but not severely so.

Modeling of Residual Returns

NOW WE KNOW SYSTEMATICS RETURNS (E.G. FACTOR/SECTOR). IT IS TIME TO MODEL RESIDUAL RETURNS

$$\frac{dP_t}{P_t} = \alpha dt + \sum_{j=1}^{n} \beta_j F_t^{(j)} + dX_t$$

1. **(Big) assumption: residuals follow mean-reverting Ornstein-Uhlembeck process**

$$dX_i(t) = \kappa_i \left(m_i - X_i(t)\right) dt + \sigma_i dW_i(t), \quad \kappa_i > 0.$$

2. **With parameters that vary slowly vs. Brownial motion increments**
3. **Fitting window is assumed to be 60 business days (quarterly reporting cycle)**
4. **Good news: we can test if O-U adequately describes the dynamics of residuals**

Calibrating O-U Process

$$dx_t = \Theta(\mu - x_t)dt + \sigma dB_t$$

$$\Rightarrow e^{\theta t}\left(dx_t + \theta X_t\, dt\right) = \left(\theta\mu\, dt + \sigma\, dB_t\right) e^{\theta t}$$

$$\Rightarrow \int_s^{s+\Delta s} d\left(e^{\theta t} X_t\right) = \int_s^{s+\Delta s} \theta\mu\, e^{\theta t}\, dt + \int_s^{s+\Delta s} \sigma\, e^{\theta t}\, dB_t$$

$$\Rightarrow e^{\theta(s+\Delta s)} X_{s+\Delta s} - e^{\theta s} X_s \cdot e^{-\theta(s+\Delta s)}$$

$$= \left[\mu\left(e^{\theta(s+\Delta s)} - e^{\theta s}\right) + \sigma\int_s^{s+\Delta s} e^{\theta t}\, dB_t\right] e^{-\theta(s+\Delta s)}$$

$$\Rightarrow X_{s+\Delta s} = e^{-\theta\Delta s} X_s + \left(1 - e^{-\theta\Delta s}\right)\mu + N\left(0, \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta\Delta s}\right)\right)$$

(Normal)

$$X = BF + a + u$$

Calibrating O-U Process Cont'd

$$dx_t = \Theta(\mu - x_t)dt + \sigma dB_t$$

$X = BF + a + U$

obtain the coefficients : B, a, U ← residual

from the regression.

$\therefore\ e^{-\theta} = B \Rightarrow \theta = -\ln B$

$\mu(1 - e^{-\theta}) = a \Rightarrow \mu = (1 - e^{-\theta})^{-1} a$

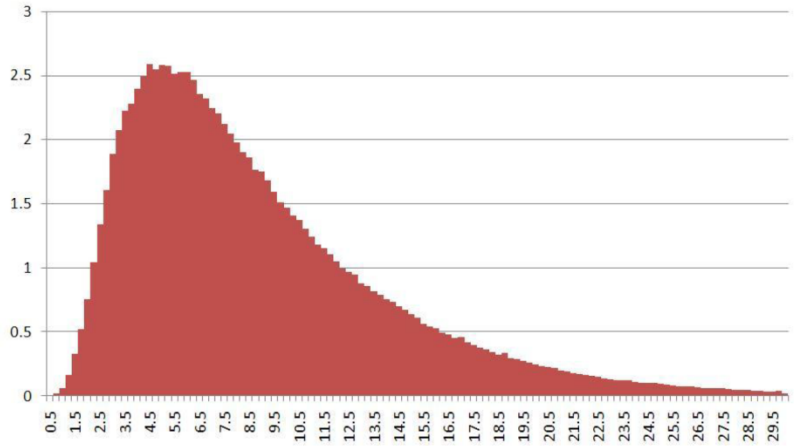
σ estimated from the residual

1. **1/θ describes the characteristic scale of mean reversion: 1/θ <<T=60 days**
2. **Check if the residuals are adequately described by N(0,ε) – likely not**
3. **StErr(μ) >> StErr(θ), so unless you believe stocks are grossly mispriced, you shouldn't let μ deviate too much from 0. Authors re-center the basket to zero:**

$$\overline{m}_i = m_i - \frac{1}{N}\sum_{j=1}^{N} m_j, i = 1, 2, ..., N$$

Calibrating O-U Process. Results:

$$dx_t = \Theta(\mu - x_t)dt + \sigma dB_t$$



	Days
Maximum	30
75%	11
Median	7.5
25%	4.9
Minimum	0.5

Averages for groups of stocks within a sector

ETF	Abs(Alpha)	Beta	Kappa	Reversion days	EquiVol	Abs(m)
HHH	0.20%	0.69	38	7	4%	3.3%
IYR	0.11%	0.90	39	6	2%	1.8%
IYT	0.18%	0.97	41	6	4%	3.0%
RKH	0.10%	0.98	39	6	2%	1.7%
RTH	0.17%	1.02	39	6	3%	2.7%
SMH	0.19%	1.01	40	6	4%	3.2%
UTH	0.09%	0.81	42	6	2%	1.4%
XLF	0.11%	0.83	42	6	2%	1.8%
XLI	0.15%	1.15	42	6	3%	2.4%
XLK	0.17%	1.03	42	6	3%	2.7%
XLP	0.12%	1.01	42	6	2%	2.0%
XLV	0.14%	1.05	38	7	3%	2.5%
XLY	0.16%	1.03	39	6	3%	2.5%
Total	0.15%	0.96	40	6	3%	2.4%

Define stock's score (deviation from the mean):

$$s_i(t) = \frac{X_i(t) - m_i}{\sigma_{eq,i}}$$

Trading Signals

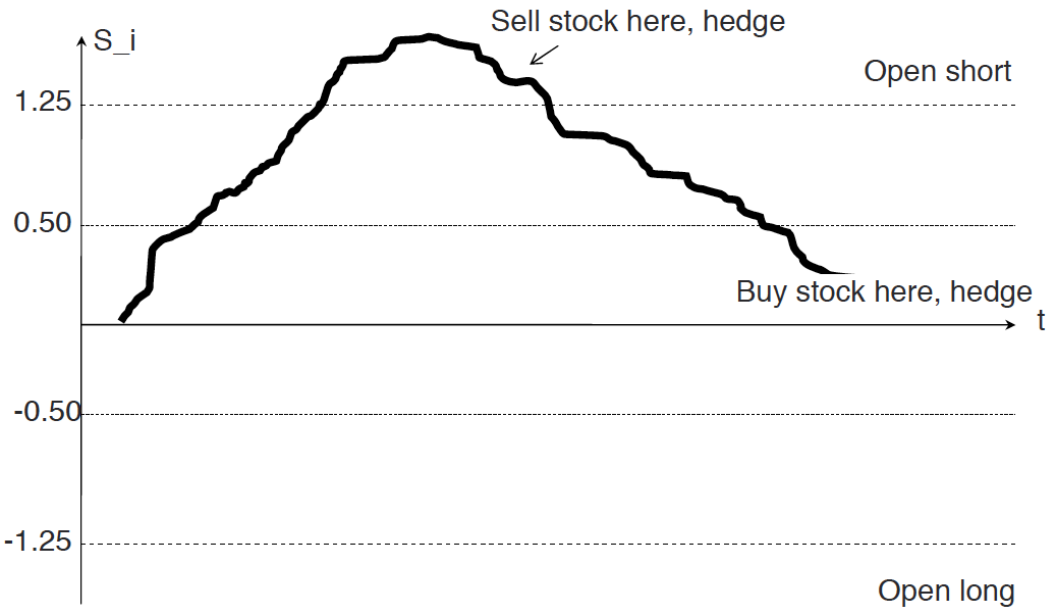
Time To Define Rules for our Rule-Based Trading Strategy:

Open long position if $s_i < -1.25$

Open short position if $s_i > +1.25$

Close long position if $s_i > -0.50$

Close short position if $s_i < +0.50$



Question of drift:

- Average drift 15bps
- Average volatility 300bps
- Average reversion 7 days

- Adjustment to score is 15*7/300=0.3, have minor effect on the outcome.

Back-Testing: P&L Equation

$$E_{n+1} = E_n + r\Delta t E_n + \sum_{i=1}^{N} Q_{i,n} R_{i,n} - r\Delta t \left(\sum_{i=1}^{N} Q_{i,n} \right) - \delta r \Delta t \sum_{i=1}^{N} \left| Q_{i,n} \right| - \varepsilon \sum_{i=1}^{N} \left| Q_{i,n+1} - Q_{i,n} \right|$$

$Q_{i,n}$ = investment in stock i at the start of period n

$R_{i,n}$ = dividend-adjusted return of stock over period n

r = Fed Funds rate or reference rate for cash

$r + \delta r$ = interest paid for cash on long stock

$r - \delta r$ = interest received for cash on short stock

ε = market impact + clearing & commissions

E_n = equity in the account at start of period n

Typically, we will assume $\varepsilon = 5$ bps $= 0.0005$,

and $\delta r = 0$, for simplicity

Back-Testing: Performance Metrics

$$\mu = \frac{1}{\Delta t N_{\text{periods}}} \sum_{n=1}^{N_{\text{periods}}} \frac{E_n - E_{n-1}}{E_{n-1}}$$

Expected return over simulation period

$$\sigma^2 = \frac{1}{\Delta t N_{\text{periods}}} \sum_{n=1}^{N_{\text{periods}}} \left(\frac{E_n - E_{n-1}}{E_{n-1}} - \mu \Delta t \right)^2$$

Variance over simulation period

$$S = \frac{\mu - r}{\sigma}$$

Sharpe Ratio

The Sharpe ratio measures returns above the risk-free rate.

It is independent of the leverage of the strategy (dimensionless).

Back-Testing: Procedure

-- Large, diversified trading universe of U.S. equities (~ 1000 names)

-- Select within the trading universe those stocks that have a trading signal (large magnitude of s-score) and open trades

-- Monitor for closing trades through s-score as well

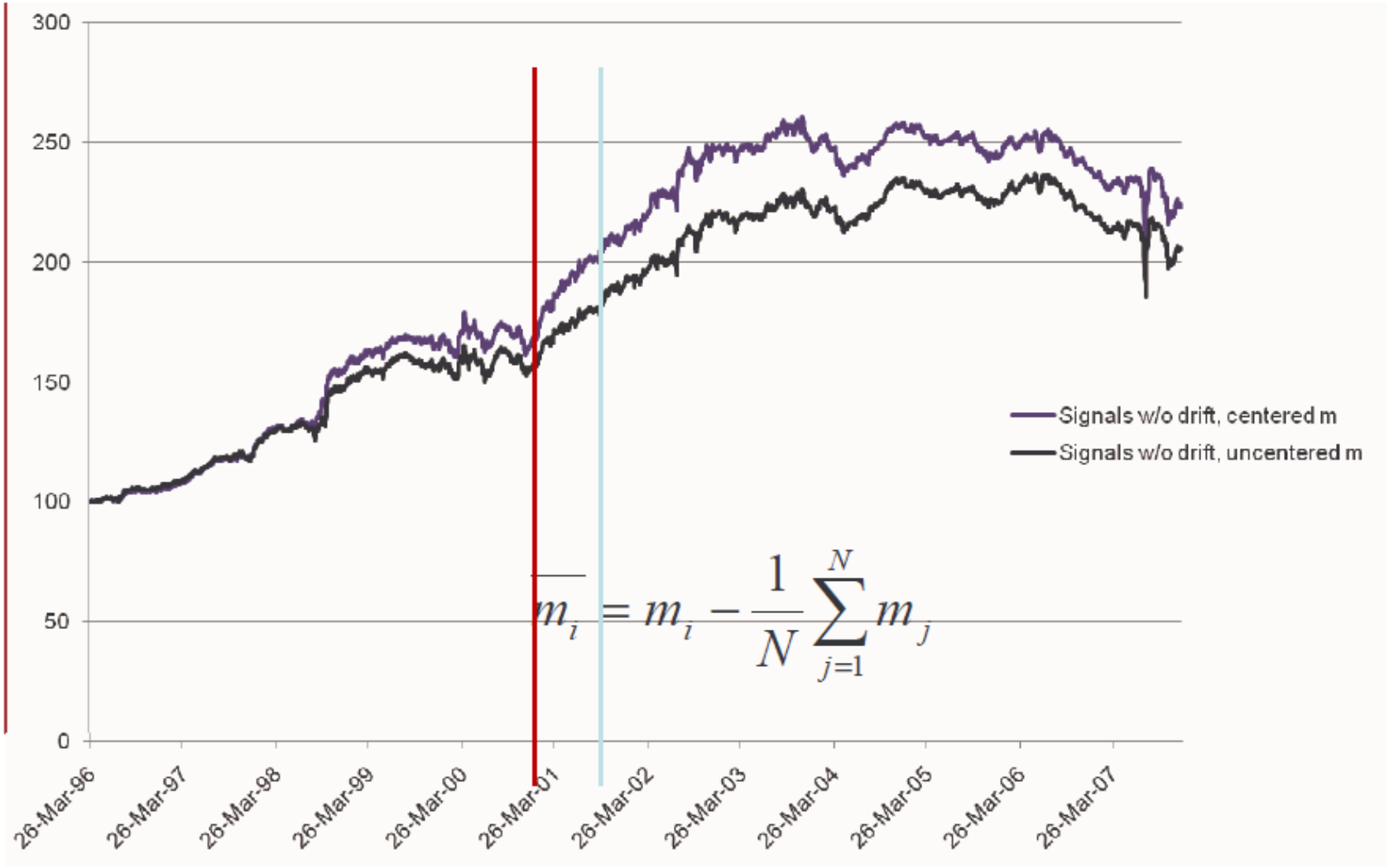
-- Keep all sectors beta-neutral by using ETFs to balance the portfolio and maintain sector-neutrality

-- Leverage = 2+2 (i.e. $2 long, $2 short for $1 of capital)

-- Expected Volatility for this Leverage < 7% annualized (< 50 bps/day)

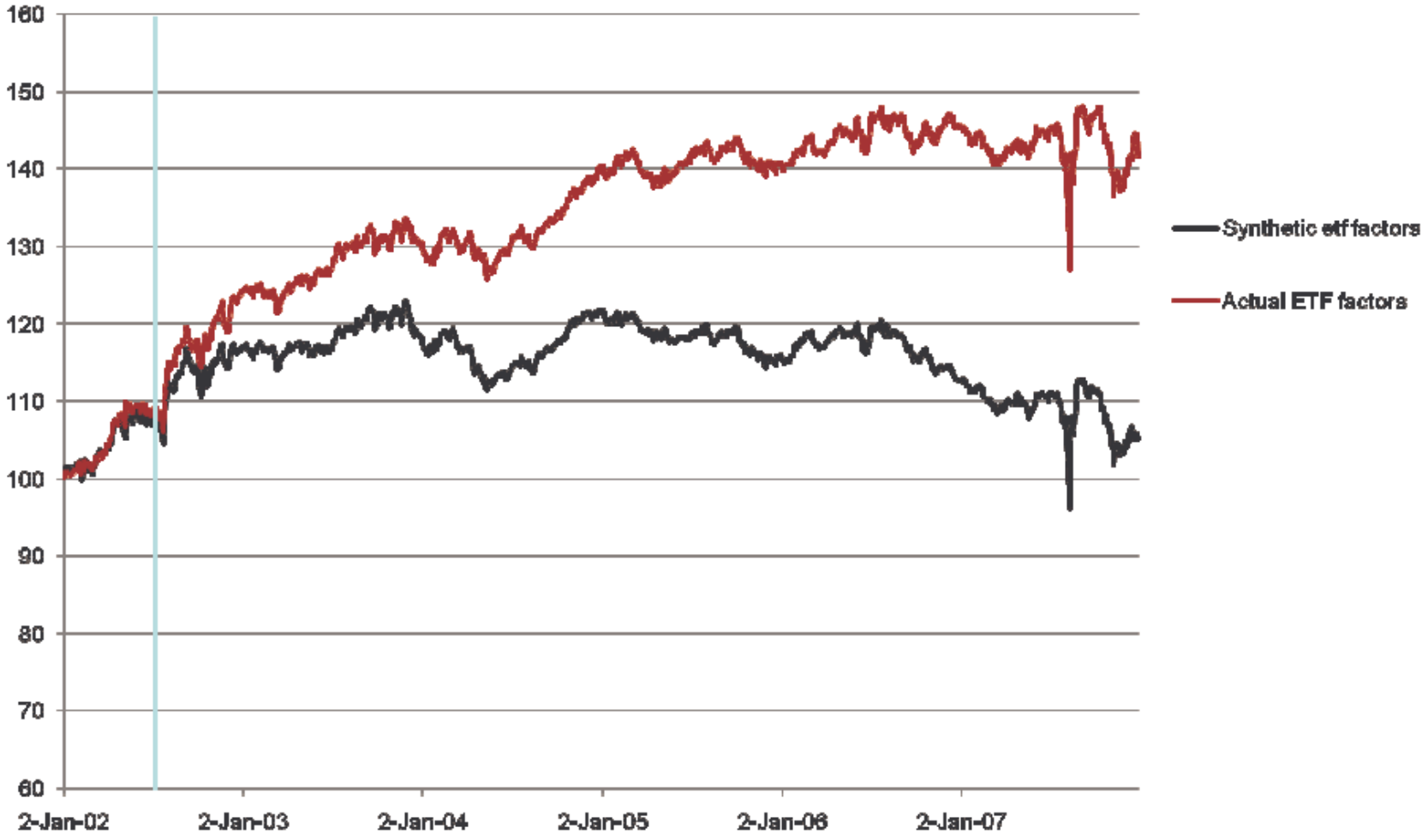
$$\Lambda = \frac{\sum_{i=1}^{N} |Q_{i,n}|}{E_n} = \text{leverage ratio} \quad \Lambda = \frac{\text{Long Market Value} + |\text{Short Market Value}|}{\text{Equity}}$$

Using Synthetic ETFs as factors



Signals Based on Actual ETFs

ACTUAL ETFS ARE TRADABLE (WHICH MIGHT REINFORCE MEAN-REVERSION). THEIR COMPOSITION IS DYNAMIC.

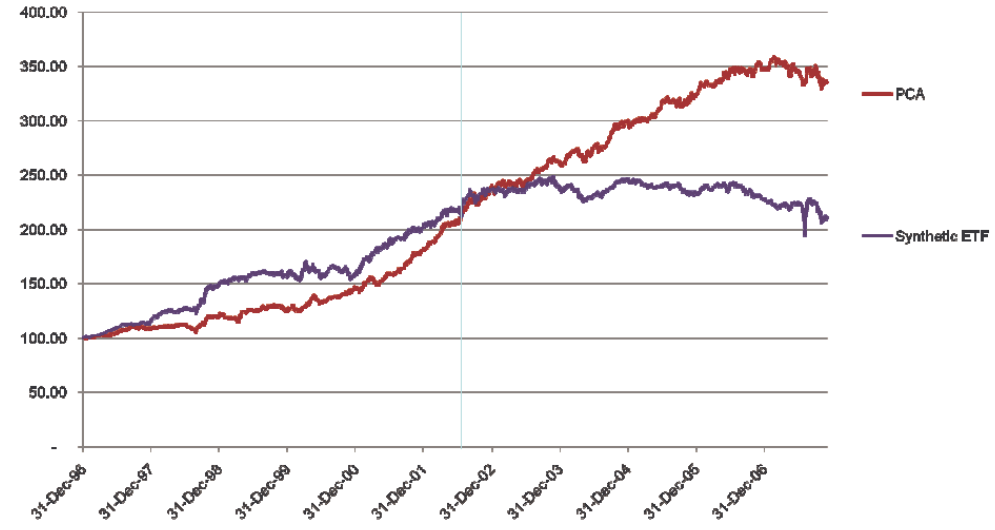


Sharpe Ratios: Synthetic vs. Actual ETFs

	HHH	IYR	IYT	OIH	RKH	RTH	SMH	UTH	XLE	XLF	XLI	XLK	XLP	XLV	XLY	Porfolio
1996	1.7	1.7	(1.2)	1.0	0.8	(0.6)	0.6	1.4	0.6	2.3	0.5	1.5	(0.5)	1.1	0.4	1.7
1997	0.1	1.5	(0.0)	2.5	1.2	1.1	2.2	1.1	(1.0)	2.3	0.6	1.1	0.4	1.5	2.9	3.6
1998	0.9	(0.5)	(0.5)	0.8	2.5	1.8	2.4	2.0	1.1	2.1	0.8	3.0	0.1	(0.1)	2.8	3.4
1999	(1.0)	(1.3)	1.5	(1.3)	(0.7)	0.3	1.2	(1.2)	1.4	1.9	1.1	1.9	(1.1)	0.1	0.6	0.8
2000	(0.4)	1.0	1.2	(0.6)	2.1	0.1	(0.7)	0.7	1.0	0.2	(0.8)	0.9	0.1	(0.5)	(1.1)	0.3
2001	(0.9)	2.8	0.7	0.6	2.7	1.5	(0.9)	0.6	1.6	0.1	1.9	1.9	0.6	1.4	3.3	2.9
2002	1.9	1.5	(0.1)	1.0	2.1	0.7	(0.5)	(1.1)	(1.3)	1.6	0.8	2.0	1.3	0.0	1.8	2.0
2003	0.5	0.0	(0.4)	(0.4)	2.6	1.3	(1.3)	(0.9)	0.1	(0.4)	(0.8)	2.5	(0.6)	(1.0)	(1.1)	0.1
2004	0.7	0.1	1.2	0.3	1.3	(0.4)	0.1	(1.1)	0.6	0.1	1.1	1.2	(0.0)	(0.8)	(0.0)	0.8
2005	0.1	(2.1)	(0.3)	(0.8)	(0.1)	0.2	0.5	(2.1)	0.0	(0.8)	(0.1)	1.0	(1.1)	(0.6)	(0.5)	(1.3)
2006	(0.7)	(1.8)	(0.1)	(0.3)	1.6	(0.4)	(0.2)	0.3	(0.7)	(1.1)	0.9	0.7	(0.9)	(1.0)	1.1	(0.5)
2007	2.1	(2.1)	0.6	(1.4)	(1.1)	(0.9)	0.1	(1.1)	(0.8)	(1.0)	1.0	(0.0)	0.0	(0.6)	1.1	(0.5)
	0.4	0.1	0.2	0.1	1.2	0.4	0.3	(0.1)	0.2	0.6	0.6	1.5	(0.2)	(0.0)	0.9	1.1

	HHH	IYR	IYT	OIH	RKH	RTH	SMH	UTH	XLE	XLF	XLI	XLK	XLP	XLV	XLY	Porfolio
2002	1.9	2.1	1.4	0.6	2.4	1.5	(0.7)	(0.2)	(0.2)	1.8	0.7	1.5	1.8	(0.1)	2.4	2.7
2003	(0.2)	0.8	(0.3)	(0.5)	1.4	1.1	(1.0)	(0.1)	0.5	0.6	(0.6)	2.6	0.3	(0.4)	(0.4)	0.8
2004	0.9	1.6	(0.7)	0.4	0.5	0.1	0.2	(0.4)	0.6	0.6	1.4	1.9	0.5	(0.6)	0.3	1.6
2005	0.3	(1.5)	0.8	(0.6)	0.3	0.5	0.5	(1.1)	(0.1)	0.9	0.6	1.3	(0.7)	0.2	0.0	0.1
2006	(0.2)	(1.3)	0.0	(0.2)	0.9	(0.1)	0.5	1.7	(0.5)	(0.6)	1.7	1.7	(0.0)	(0.4)	2.0	0.7
2007	(0.4)	(0.3)	0.0	(1.3)	(1.2)	(0.7)	0.9	(0.7)	(1.0)	(0.6)	1.1	0.6	0.4	(0.5)	1.3	(0.2)
	0.4	0.2	0.2	(0.3)	0.7	0.4	0.1	(0.1)	(0.1)	0.5	0.8	1.6	0.4	(0.3)	0.9	0.9

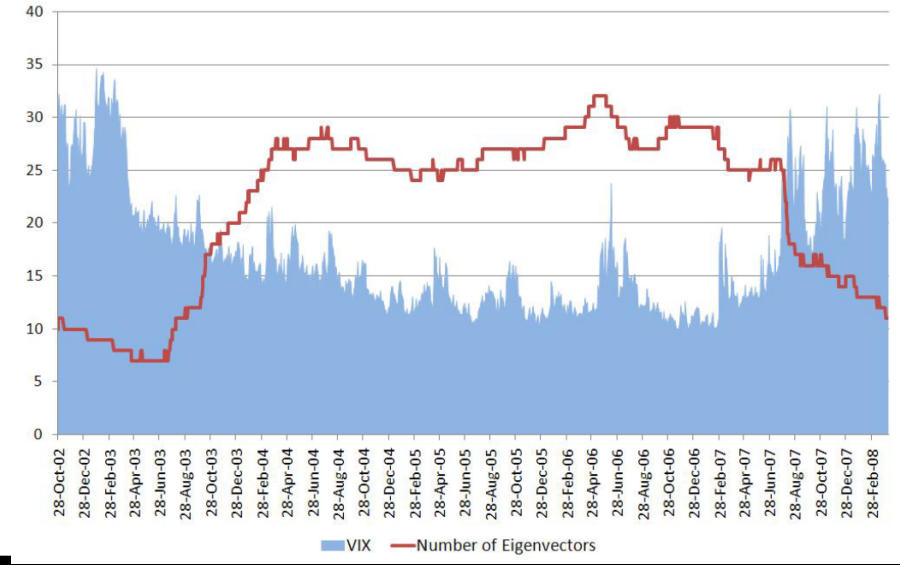
Signals Based on 15 PCA Factors Outperform

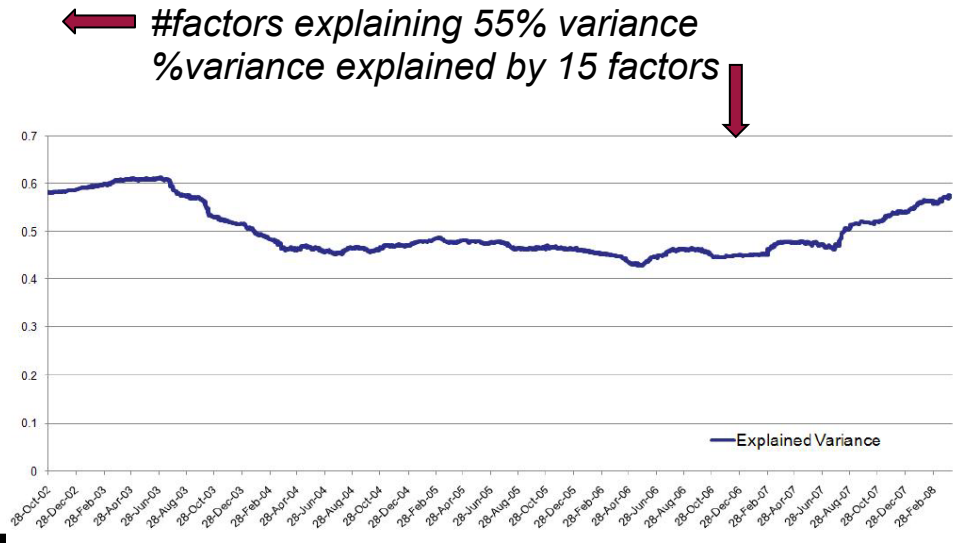


year	AnnRetPL	AnnVolPL	SharpePL
1997	9%	5%	1.73
1998	11%	6%	1.71
1999	3%	7%	0.49
2000	16%	7%	2.44
2001	22%	8%	2.86
2002	28%	7%	3.73
2003	8%	7%	1.23
2004	14%	5%	2.56
2005	8%	5%	1.53
2006	7%	5%	1.42
2007	-6%	16%	(0.36)

Sensitivity to Number of PCA Factors

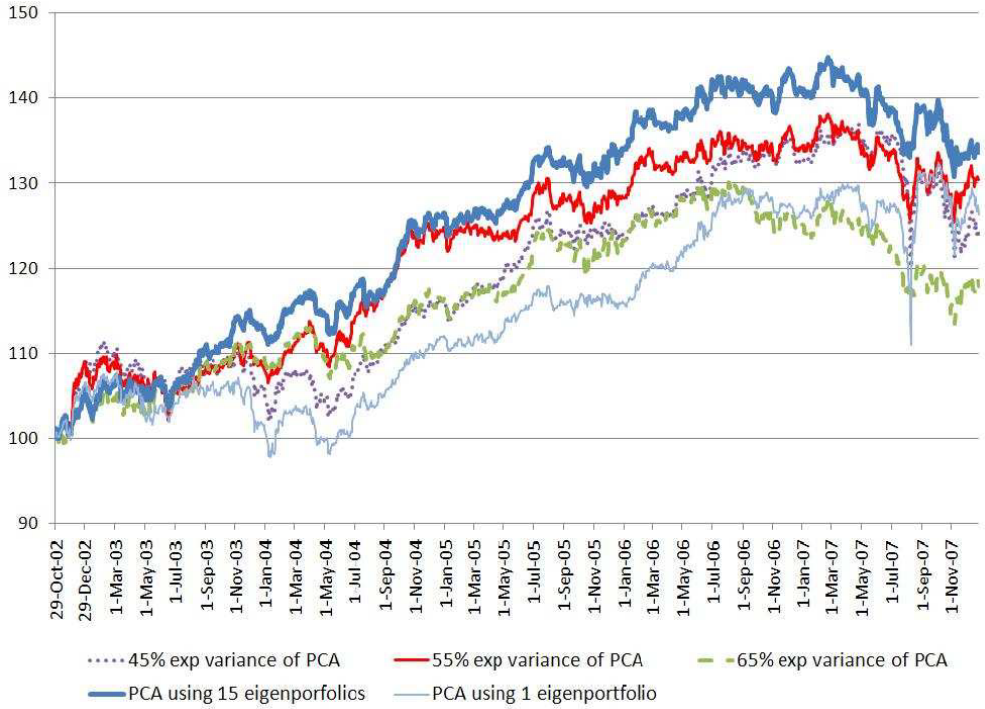
1. The separation between systematic and idiosyncratic components is a theoretical construct. **It is not observable**!
2. **Too few factors** (think CAPM) lead to large variance of residuals and low sensitivity to size, industry, etc. Residuals are less likely to mean-reverts over short time periods
3. **Too many factors** lead to noise trading (the magnitude of deviations is smaller, so potential profits will get eaten by transaction costs). Also eigenvectors may belong to noise band, so out-of-sample performance may be (very) poor.
4. Fixed number of factors explains different % of total variance in various market conditions





Sensitivity to Number of PCA Factors Cont'd

P&L and Sharpe Ratios for Different Truncation Levels:



	1 Eigenportfolios	15 Eigenportfolios	45 % Exp Variance	55% Exp Variance	65% Exp Variance
2003	-0.7	0.9	-0.5	-0.1	0.4
2004	1.7	2.2	1.7	2.6	1.3
2005	0.8	1.2	1.3	0.6	1.0
2006	1.8	1.0	1.3	0.9	0.3
2007	0.0	-0.7	-0.7	-0.4	-0.9
Since Inception	0.7	0.9	0.6	0.7	0.4

Trading Time vs. Actual Time

1. **Statistics on equity returns can be done**
   A. -- in actual time (% change/day)
   B. -- in trading time (% change per share/day) - Incorporates volume information
2. **In trading-time framework, mean-reverting signals (S-Scores) are**
   A. -- weaker when volume is heavy
   B. -- stronger when volume is light

$$\varepsilon = \frac{\Delta S}{S} - \beta \frac{\Delta I}{I} \qquad \text{(usual residual)}$$

$$\bar{\varepsilon} = \frac{< \Delta V >}{\Delta V} \varepsilon, \qquad \Delta V = \text{daily volume} \qquad < \Delta V > = \text{average } \Delta V$$

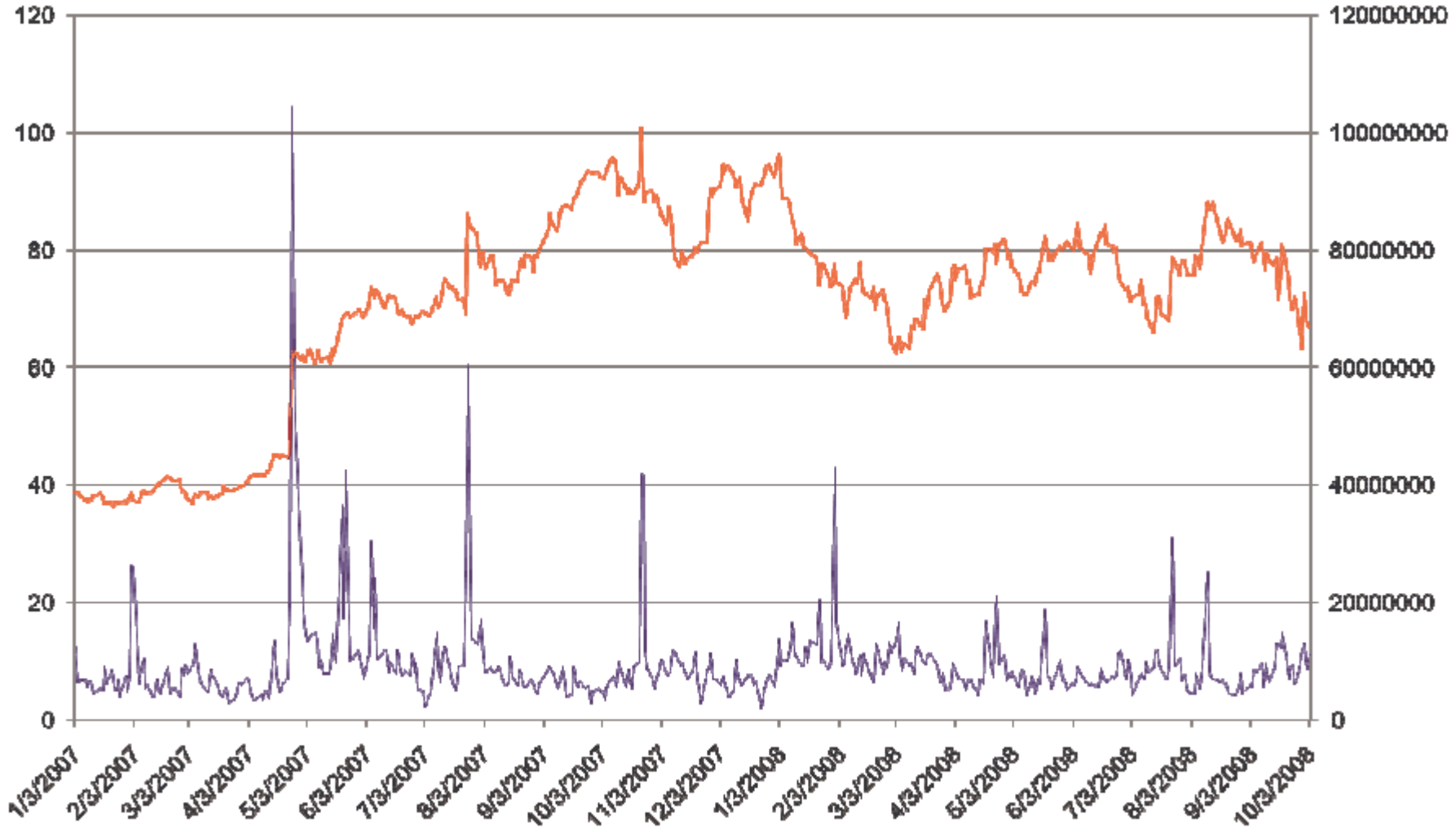
$$Y_t = \sum_{i=1}^{t} \bar{\varepsilon}_i$$

$$dY = \kappa(m - Y)dt + \sigma dW$$

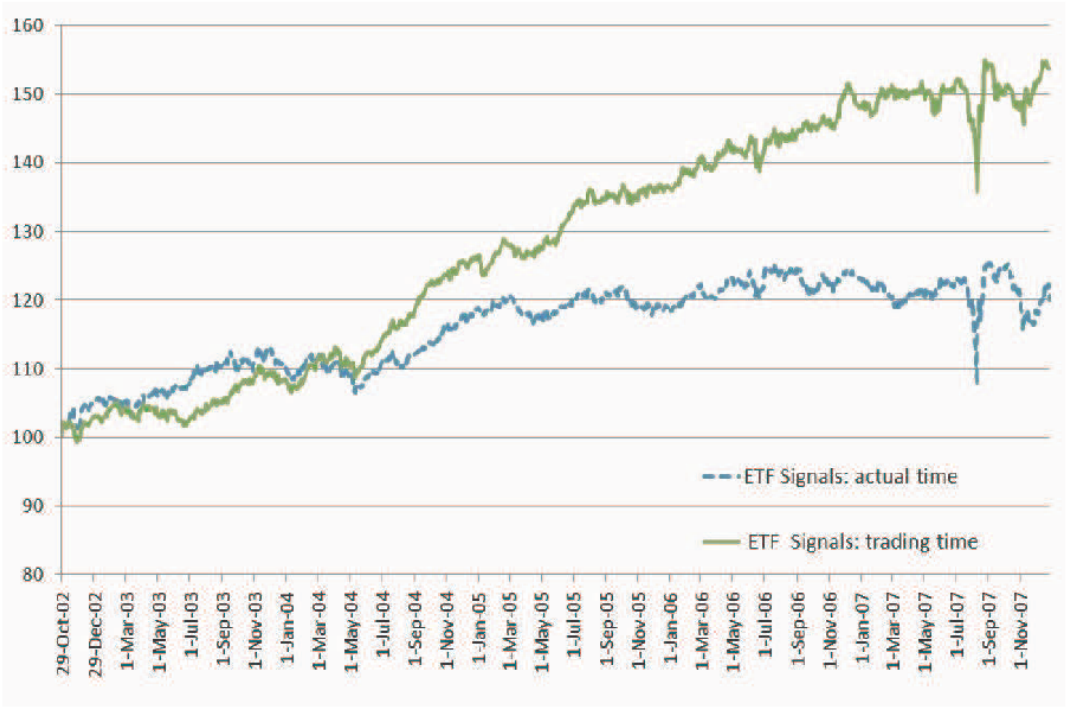
Estimate AR-1 / OU process for the new process Y(t)

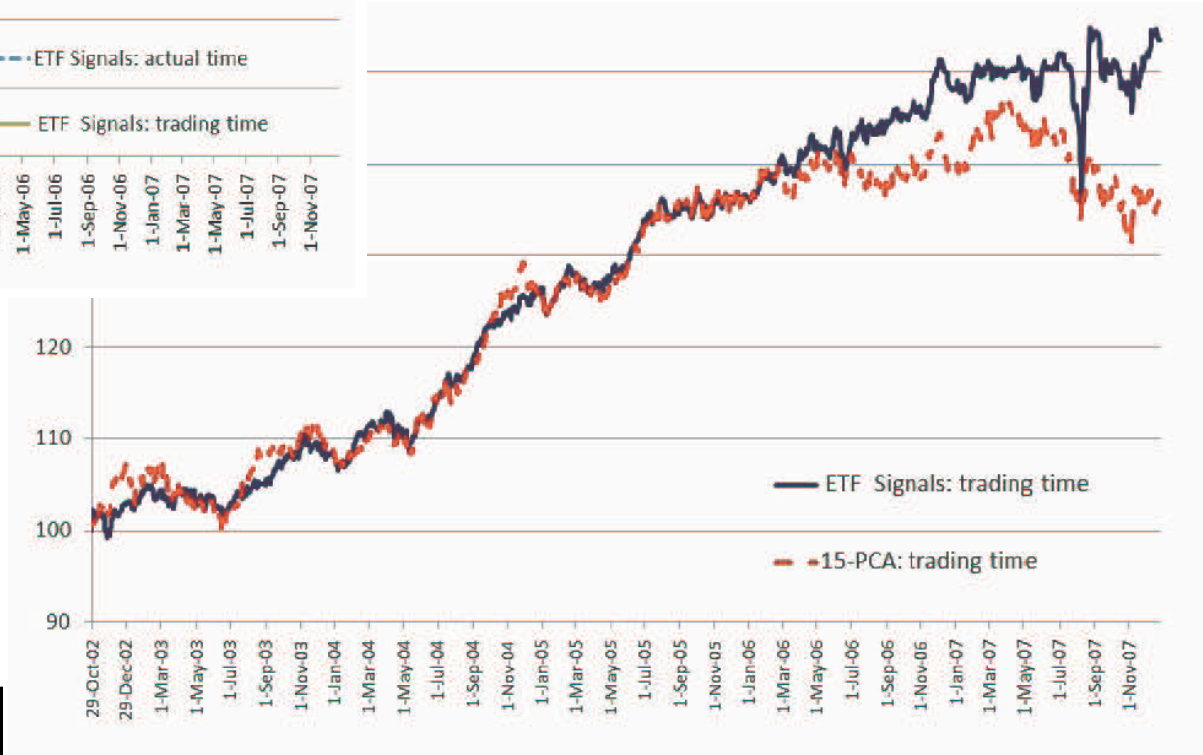
Events That Trading Time Suppresses:

AMZN January 2007- October 2008:



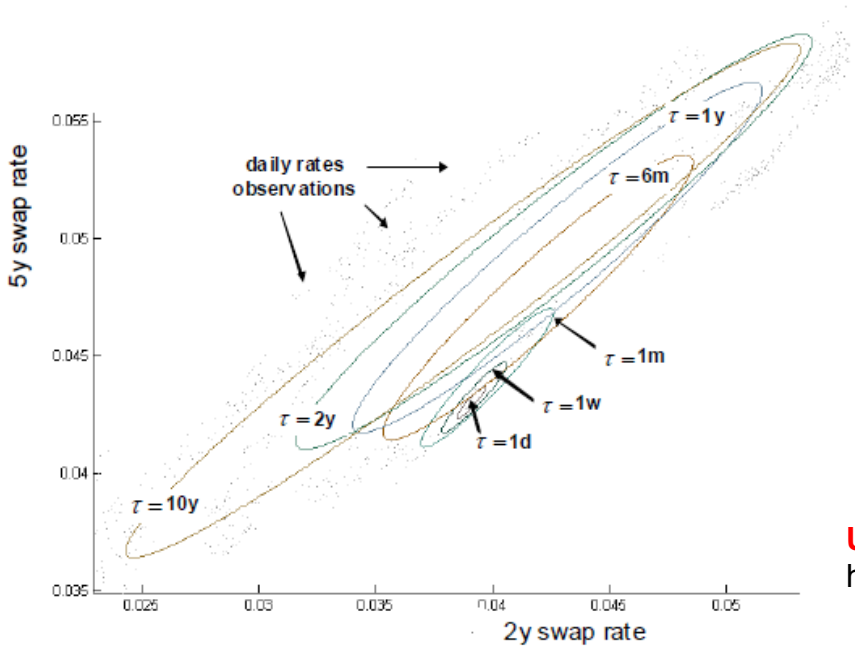
Trading Time vs. Actual Time: Results:





Attilio Meucci – Review of Statistical Arbitrage

HTTP://PAPERS.SSRN.COM/SOL3/PAPERS.CFM?ABSTRACT_ID=1404905



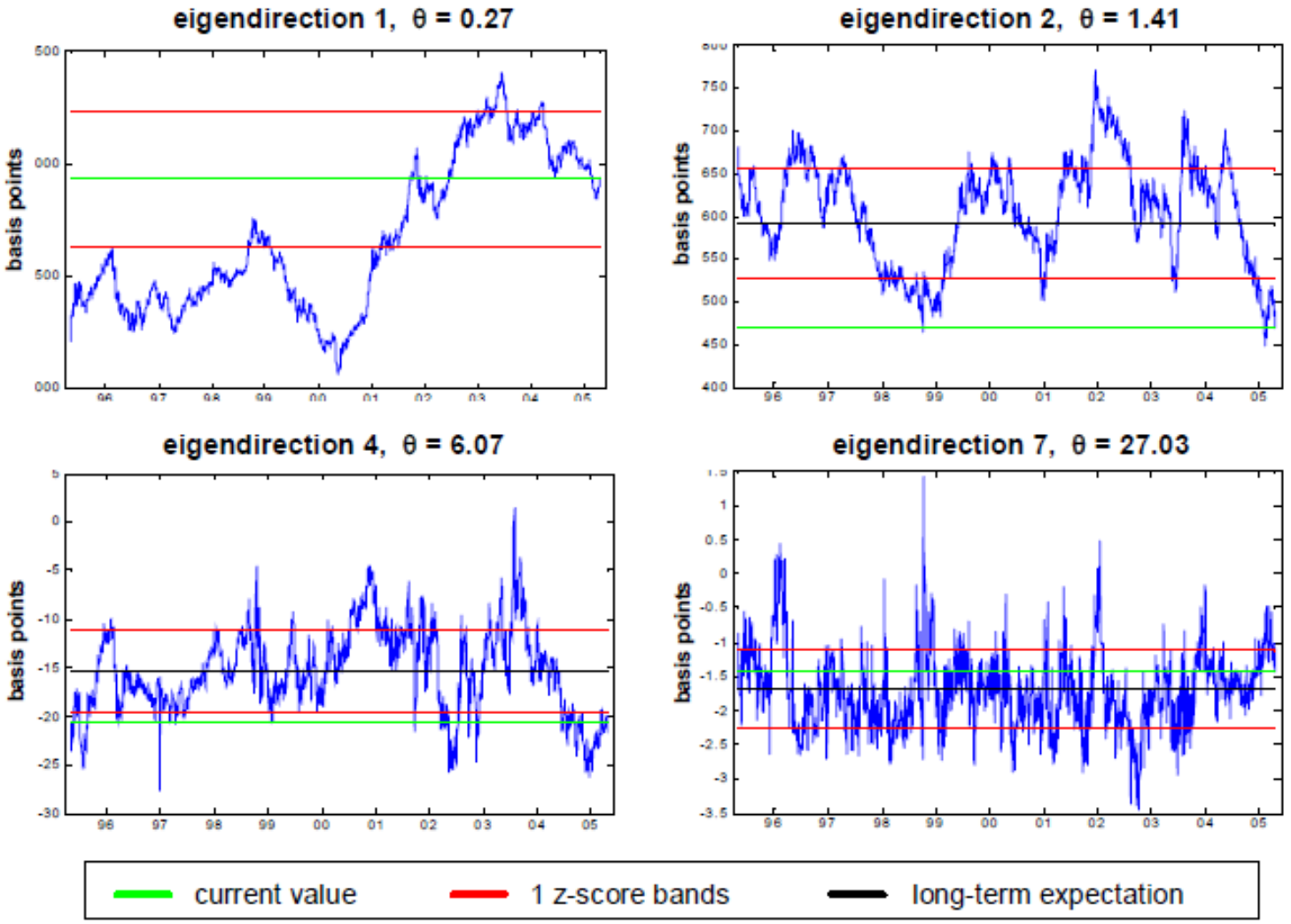
Useful Matlab Code:
http://www.mathworks.com/matlabcentral/fileexchange/24120

1. **The combination that minimizes conditional variance is the best candidate for cointegration:** $\tilde{\mathbf{w}} \equiv \underset{\|\mathbf{w}\|=1}{\operatorname{argmin}} \left[\operatorname{Var}\left\{Y_{\infty}^{\mathbf{w}} | \mathbf{x}_0\right\}\right].$

2. **Solution is** $\tilde{\mathbf{w}} \equiv \mathbf{e}^{(N)}$**, the eigenvector relative to the smallest eigenvalue** $\lambda^{(N)}$

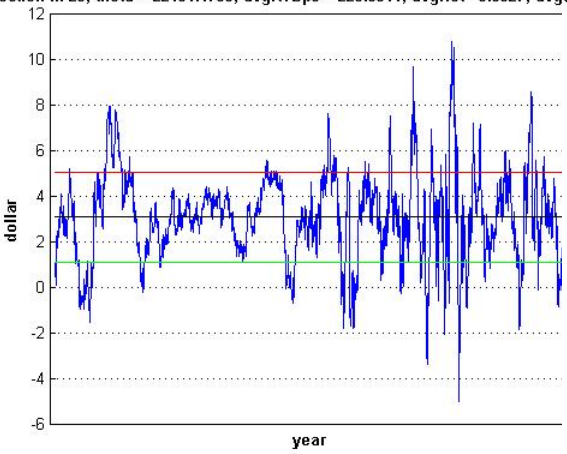
3. **This approach assumes knowledge of the true covariance !!!**

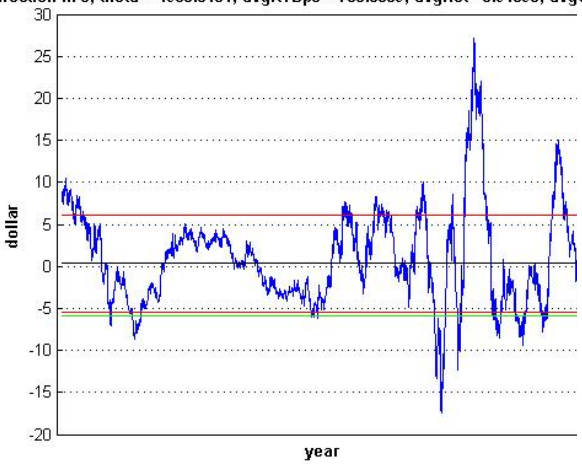
In-Sample Results

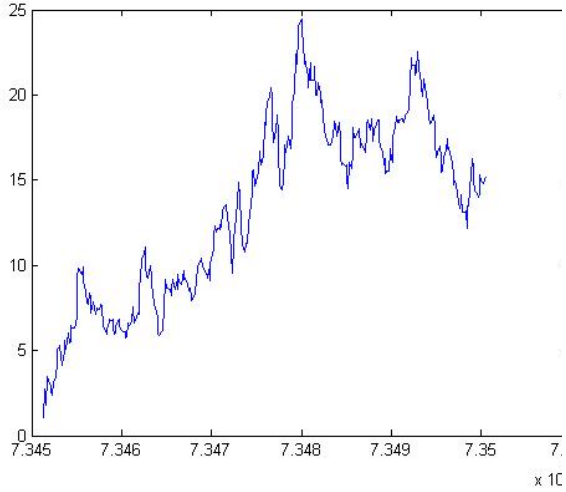


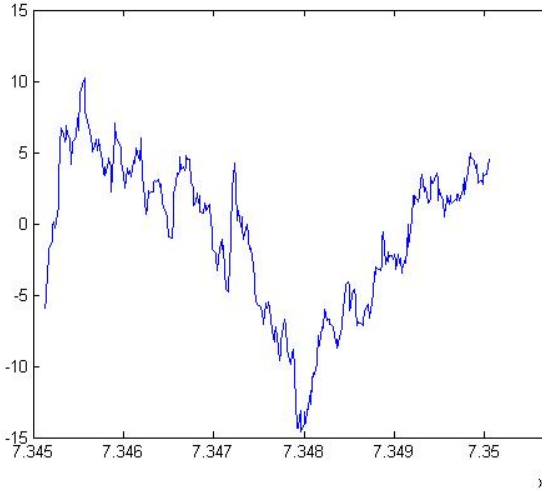
Out-of-Sample Results

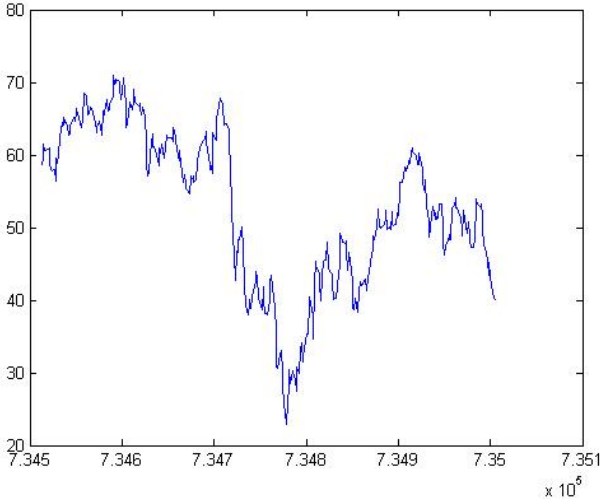
endirection n. 20, theta = 22184.1708, avgRTBps = 225.8344, avgNet= 3.0527, avgendirection n. 8, theta = 4966.0151, avgRTBps = 730.6509, avgNet= 0.94693, avgendirection n. 4, theta = 504.2117, avgRTBps = 3047.1447, avgNet= 18.8145, avgGross= 15











Improvements

1. **Choice of investable universe & proper risk model**

- Stocks in your investable universe should be related (fundamentally, economically, geographically, etc.). You will get more eigenvalues of empirical correlation matrix outside of the noise zone which explain higher proportion of total variance.
- Reduce the dimensionality of covariance matrix by building a multifactor model.
- Compute cumulative residual returns. Always trade well diversified basket (optimize). You will come up with a decent factor that still works, but pretty crowded (alpha decays fast, capacity is not great)

2. **Use "volume" time**
3. **Trade on temporary (flow-induced) dislocations rather than on permanent (news-induced) dislocations = avoid trading on company-specific news. Avoid certain sectors (biotech)**
4. **Explore less efficient markets outside of the U.S.**
5. **Relative value strategies usually work better for mid- to large-cap names; "sweet spot" is probably bottom 500 stocks in R1000…**
6. **Don't overfit!!! :"The probability of backest overfitting" by D.Bailey, J.Borwein, M. Lopez de Prado, Q Zhu** (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2326253)