Elements of Pairs Trading and Systematic Backtesting

by Dr Richard Diamond

CQF JAN 2022

1. Introduction to CQF Final Project

2. Time Series and Cointegration for Pairs Trading

3. Arbitrage: OU Process Signal Generation and Control

4. Trading Strategy Evaluation: Systematic Backtesting, Scorecards

Preparation

- Start downloading data and planning own implementation of the project.
- It is expected you apply own understanding. It is up to you to source and generate suitable data and features.

  Equities Data tutorial (*yfinance*). If you can't get hold: **make reasonable assumptions**, even generate the data.

  For example: you can make up reasonable CDS spreads and use correlation from equities.

  Refer to the relevant CQF Lectures and do extra reading on pricing methodologies and numerical techniques.

Code Adoption

A. You can adopt code for specific tasks, not model as a whole. Amend code it for your purpose, not copy/paste.

B. Where numerical techniques intense (and not part of the central model): use ready libraries with expertise. For example, portfolio optimisation best done using quadratic optimisation routine, not stochastic gradient Solver.

C. You are welcome to implement complex numerical methods vs. use of ready solution – if able to.

Numerical Techniques

Implement as necessary, numerical techniques
from the first principles.

What to code: pricing formulae, Black-Litterman calculation, SDE simulation, matrix form regression, Engle-Granger, interpolation, numerical integration, Cholesky, t-copula formula, CDS bootstrap, features computation...

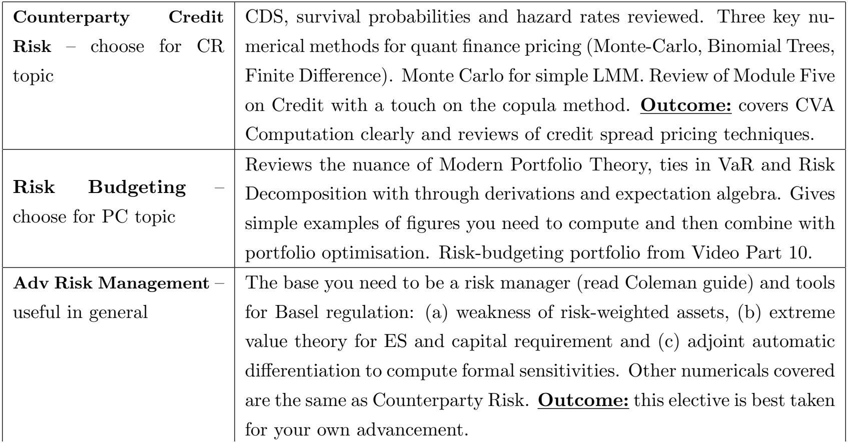
Use ready solutions for: covariance shrinkage, nearest correlation, ML numerical methods (eg, decision trees, neural nets), low latency RNs, kernel density (cdf estimation), QR-decomposition (PCA), EGARCH estimation, Johansen Procedure...

Project Report

- A full **mathematical description** of the models employed as well as numerical methods. Remember *accuracy and convergence*!
- Results presented using **a plenty of tables and figures**, which must be interpreted not just thrown at the reader.
- **Pros and cons** of a model and its implementation, together with possible improvements.
- **Demonstrate 'the specials'** of your implementation: own research, own coding of complex methods, use of the industrial-strength libraries of C++, Python.
- Instructions on how to use software if not obvious.
  The code must be thoroughly tested and well-documented.

CQF Electives

See Project Brief for the current relevant table.



Counterparty Credit Risk – choose for CR topic	CDS, survival probabilities and hazard rates reviewed. Three key numerical methods for quant finance pricing (Monte-Carlo, Binomial Trees, Finite Difference). Monte Carlo for simple LMM. Review of Module Five on Credit with a touch on the copula method. **Outcome:** covers CVA Computation clearly and reviews of credit spread pricing techniques.
Risk Budgeting – choose for PC topic	Reviews the nuance of Modern Portfolio Theory, ties in VaR and Risk Decomposition with through derivations and expectation algebra. Gives simple examples of figures you need to compute and then combine with portfolio optimisation. Risk-budgeting portfolio from Video Part 10.
Adv Risk Management – useful in general	The base you need to be a risk manager (read Coleman guide) and tools for Basel regulation: (a) weakness of risk-weighted assets, (b) extreme value theory for ES and capital requirement and (c) adjoint automatic differentiation to compute formal sensitivities. Other numericals covered are the same as Counterparty Risk. **Outcome:** this elective is best taken for your own advancement.

Final Day as advised

Don’t Extend Your Luck!

1. Introduction to CQF Final Project

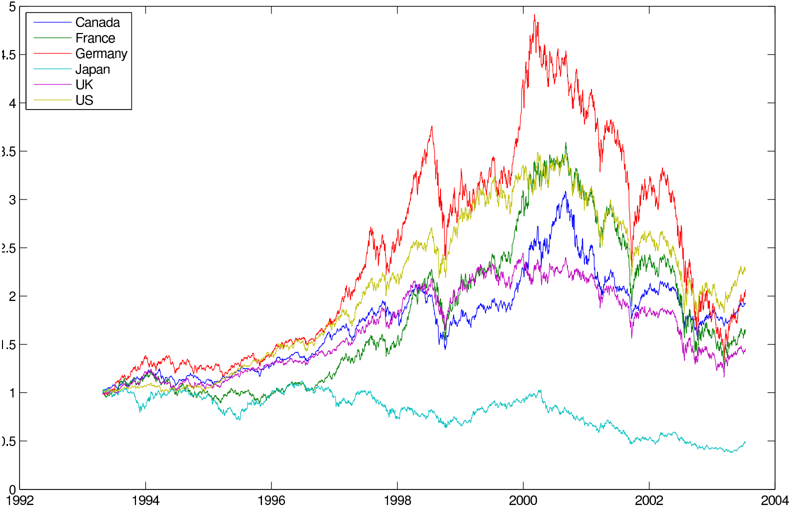
2. Time Series and Cointegration for Pairs Trading

3. Arbitrage: OU Process Signal Generation and Control

4. Trading Strategy Evaluation: Systematic Backtesting, Scorecards

Techniques from Time Series

Relative Equity Indices

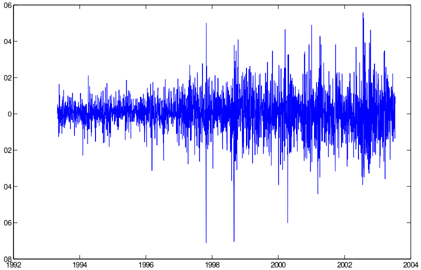


US Daily Index Returns

We will use index returns to demo *Vector Autoregression*.

- Canada, France, Germany, Japan, UK, US

Below is typical plot for the market daily returns (US). Observe the regimes of low, then high volatility.



Linear Model on Returns

For stationary returns, we set up a model-free endogenous system: variables depend on their past (lagged) values.

$$\begin{array}{rcl} R_t^{CA=1} & = & \beta_{1,0} + \beta_{11} R_{t-1}^{CA} + \beta_{12} R_{t-1}^{FR} + \ldots \beta_{1n} R_{t-1}^{US} + \ldots_{t-2} \ldots + \epsilon_{1,t} \\ R_t^{FR=2} & = & \beta_{2,0} + \beta_{21} R_{t-1}^{CA} + \beta_{22} R_{t-1}^{FR} + \ldots \beta_{2n} R_{t-1}^{US} + \ldots_{t-2} \ldots + \epsilon_{2,t} \\ \ldots & & \ldots \\ R_t^{US=n} & = & \beta_{n,0} + \beta_{n1} R_{t-1}^{CA} + \beta_{nn} R_{t-1}^{FR} + \ldots \beta_{nn} R_{t-1}^{US} + \ldots_{t-2} \ldots + \epsilon_{n,t} \end{array}$$

Consider forecasting powers of this model-free set up.

Empirical Forecasting

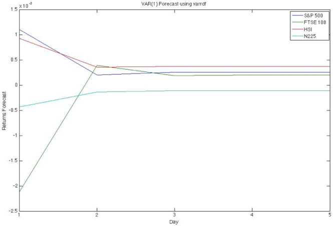
Vector Autoregression **FAILS** at forecasting daily returns (2011 data).

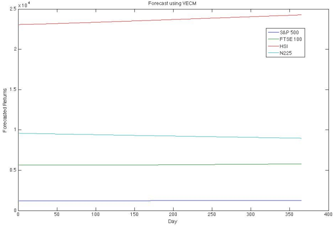
	S&P 500	FTSE 100	HSE	N225
MSE	0.0001	0.0001	0.0001	0.0001
MAPE	1.0175	1.3973	2.5325	1.0111

Table: Forecasting Accuracy: Market Index Returns (next day)

MAPE results suggest a deviation $O(100\%)$ to $O(200\%)$ per cent. Granted, daily returns for a broad market are a very small, close to negligible, quantity.

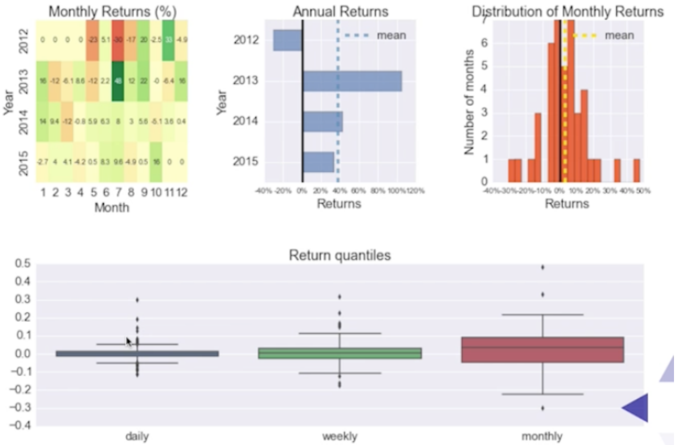
For properly stationary variables, Econometrics offers forecasting, impulse-response (IRF), and Granger causality analyses. **NONE** applicable to financial time series.





Without updating (recomputing) the regression, the forecast is a straight line.

Investment Performance



From: Quantopian Backtesting, T. Wiecki at QI2015

Vector Autoregression

VAR(p) is the structural equation model of *seemingly unrelated regressors*:

$$\begin{aligned}
y_{1,t} &= \beta_{1,0} + \beta_{11} y_{1,t-1} + \beta_{12} y_{2,t-1} + \ldots \beta_{1n} y_{n,t-1} + \ldots_{t-2} \ldots + \epsilon_{1,t} \\
y_{2,t} &= \beta_{2,0} + \beta_{21} y_{1,t-1} + \beta_{22} y_{2,t-1} + \ldots \beta_{2n} y_{n,t-1} + \ldots_{t-2} \ldots + \epsilon_{2,t} \\
\ldots & \quad \ldots \\
y_{n,t} &= \beta_{n,0} + \beta_{n1} y_{1,t-1} + \beta_{nn} y_{2,t-1} + \ldots \beta_{nn} y_{n,t-1} + \ldots_{t-2} \ldots + \epsilon_{n,t}
\end{aligned}$$

Instead of estimating by OLS line-by-line, all beta coefficients can be computed in concise form, **in one go**.

Dependent Matrix

1. *Dependent matrix* has *observations* for p lags removed. Time series in rows, $p+1$ to the most recent observation at T.

$$Y = [\mathbf{y}_{p+1}\;\mathbf{y}_{p+2}\cdots\mathbf{y}_T] = \begin{pmatrix} y_{1,p+1} & y_{1,p+2} & \cdots & y_{1,T} \\ y_{2,p+1} & y_{2,p+2} & \cdots & y_{2,T} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n,p+1} & y_{n,p+2} & \cdots & y_{n,T} \end{pmatrix}$$

$$\left[\cancel{y_{1,t=1}} \quad \cancel{y_{1,\ldots}} \quad \cancel{y_{1,p}} \quad y_{1,p+1} \quad y_{1,p+2}\cdots \quad y_{1,t=T}\right]$$

For **lag** $p=3$, we use the first three values to predict y_{p+1} (and so on).

$$n = Nvar \qquad\qquad T = Nobs$$

3. *Explanatory data matrix (assume p=3)*

$$Z = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{y}_p & \mathbf{y}_{p+1} & \cdots & \mathbf{y}_{T-1} \\ \mathbf{y}_{p-1} & \mathbf{y}_p & \cdots & \mathbf{y}_{T-2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{y}_{n,1} & \mathbf{y}_{n,2} & \cdots & \mathbf{y}_{n,T-p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_{1,p} & y_{1,p+1} & \cdots & y_{1,T-1} \\ y_{2,p} & y_{2,p+1} & \cdots & y_{2,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n=Nvar,p} & y_{n,p+1} & \cdots & y_{n,T-1} \\ & & & \\ y_{1,p-1} & y_{1,p} & \cdots & y_{1,T-2} \\ y_{2,p-1} & y_{2,p} & \cdots & y_{2,T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n=Nvar,p-1} & y_{n,p} & \cdots & y_{n,T-2} \\ & & & \\ y_{1,1} & y_{1,2} & \cdots & y_{1,T-p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,T-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n=Nvar,1} & y_{n,2} & \cdots & y_{n,T-p} \end{bmatrix}$$

Coded in *Matlab*, the algorithm forms the matrix from the top,

```
ymat = y(nlag+1:end,:)'; % Forming dependent matrix Y

zmat = [ones(1,nobs-nlag)]; % Forming explanatory matrix Z
for i=1:nlag
    zmat = [zmat; y(nlag-i+1:end-i,:)'];
end;
```

$$n = Nvar \qquad T = Nobs$$

Residuals

4. Disturbance matrix (innovations, residuals)

$$\epsilon = \begin{bmatrix} \epsilon_{p+1} & \epsilon_{p+2} & \cdots & \epsilon_T \end{bmatrix} = \begin{bmatrix} e_{1,p+1} & e_{1,p+2} & \cdots & e_{1,T} \\ e_{2,p+1} & e_{2,p+2} & \cdots & e_{2,T} \\ \vdots & \ldots & \ddots & \vdots \\ e_{n,p+1} & e_{n,p+2} & \cdots & e_{n,T} \end{bmatrix}$$

Each row of residuals matches variables $y_1, y_2, \ldots, y_{n=Nvar}$ respectively. The most recent observation is at T.

Residuals are computed once we estimated $\hat{B}$

$$\hat{\epsilon} = Y - \hat{B}Z$$

Calculating VAR(p) Estimates

- Calculate the multivariate OLS estimator for *the coefficients*

$$\hat{B} = YZ'(ZZ')^{-1}$$

This estimator is consistent and asymptotically efficient.

- For the simple case of variables x and y, regression coefficients estimated with

$$\beta_1 = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

A bit of MLE

Consider the Log-likelihood function for multivariate Normal

$$\begin{aligned} L &= \prod_t^T N(y_t, x_t, \beta, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \\ \ln L &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \end{aligned}$$

To maximise the Log-Likelihood *by varying* β

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} &= \frac{1}{\sigma^2}(Y - X\beta)'X = 0 \\ \hat{\beta} &= YX'(XX')^{-1}. \end{aligned}$$

This is how $\hat{B} = YZ'(ZZ')^{-1}$ result was obtained.

Inference

1. Estimator of the *residual covariance matrix* with $T \equiv N_{obs}$

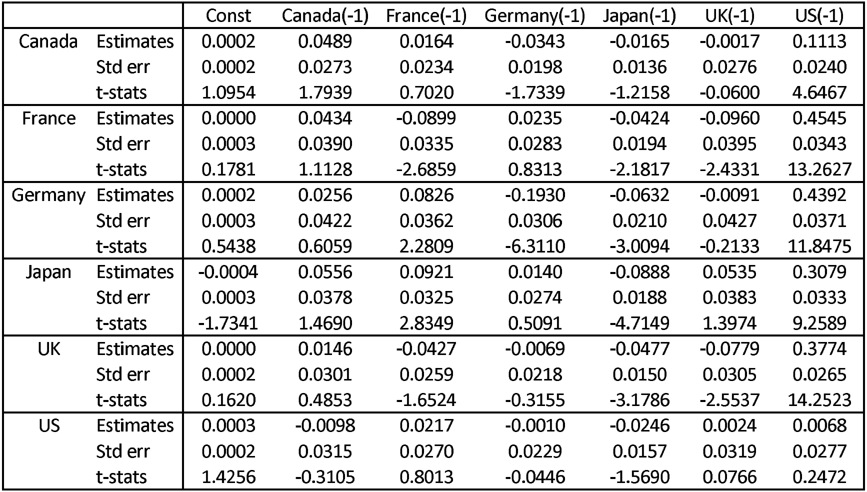
$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} \hat{\epsilon}_t \hat{\epsilon}_t'$$

2. Standard errors of beta coefficients will be inside the inverse of Information Matrix (on the diagnoal)

$$\mathbb{C}ov \left[Vec(\hat{B}) \right] = (ZZ')^{-1} \otimes \hat{\Sigma} \quad = \quad I^{-1}$$

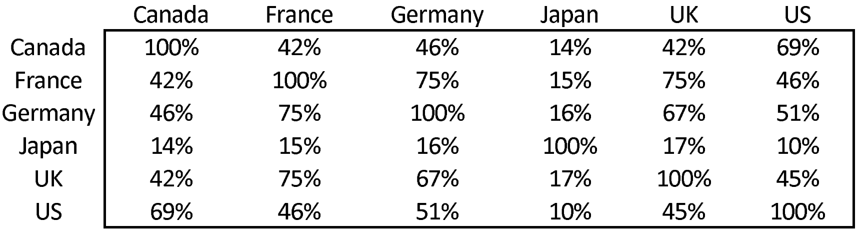
$\otimes$ is the *Kronecker product*.

VAR(1) Estimation



		Const	Canada(-1)	France(-1)	Germany(-1)	Japan(-1)	UK(-1)	US(-1)
Canada	Estimates	0.0002	0.0489	0.0164	-0.0343	-0.0165	-0.0017	0.1113
	Std err	0.0002	0.0273	0.0234	0.0198	0.0136	0.0276	0.0240
	t-stats	1.0954	1.7939	0.7020	-1.7339	-1.2158	-0.0600	4.6467
France	Estimates	0.0000	0.0434	-0.0899	0.0235	-0.0424	-0.0960	0.4545
	Std err	0.0003	0.0390	0.0335	0.0283	0.0194	0.0395	0.0343
	t-stats	0.1781	1.1128	-2.6859	0.8313	-2.1817	-2.4331	13.2627
Germany	Estimates	0.0002	0.0256	0.0826	-0.1930	-0.0632	-0.0091	0.4392
	Std err	0.0003	0.0422	0.0362	0.0306	0.0210	0.0427	0.0371
	t-stats	0.5438	0.6059	2.2809	-6.3110	-3.0094	-0.2133	11.8475
Japan	Estimates	-0.0004	0.0556	0.0921	0.0140	-0.0888	0.0535	0.3079
	Std err	0.0003	0.0378	0.0325	0.0274	0.0188	0.0383	0.0333
	t-stats	-1.7341	1.4690	2.8349	0.5091	-4.7149	1.3974	9.2589
UK	Estimates	0.0000	0.0146	-0.0427	-0.0069	-0.0477	-0.0779	0.3774
	Std err	0.0002	0.0301	0.0259	0.0218	0.0150	0.0305	0.0265
	t-stats	0.1620	0.4853	-1.6524	-0.3155	-3.1786	-2.5537	14.2523
US	Estimates	0.0003	-0.0098	0.0217	-0.0010	-0.0246	0.0024	0.0068
	Std err	0.0002	0.0315	0.0270	0.0229	0.0157	0.0319	0.0277
	t-stats	1.4256	-0.3105	0.8013	-0.0446	-1.5690	0.0766	0.2472

Residual Covariance Matrix



	Canada	France	Germany	Japan	UK	US
Canada	100%	42%	46%	14%	42%	69%
France	42%	100%	75%	15%	75%	46%
Germany	46%	75%	100%	16%	67%	51%
Japan	14%	15%	16%	100%	17%	10%
UK	42%	75%	67%	17%	100%	45%
US	69%	46%	51%	10%	45%	100%

- since our residuals $\sim N(0, \sigma^2)$ this is also correlation.
- notice the correlation for US/Canada and UK/France, UK/Germany pairs. That hints at **collinearity**, a difficulty to separate.

Optimal Lag Selection

Optimal Lag p is determined by the lowest values of AIC, BIC statistics, constructed using the penalised likelihood principle.

- *Akaike Information Criterion*

$$AIC = \log|\widehat{\Sigma}| + \frac{2k'}{T}$$

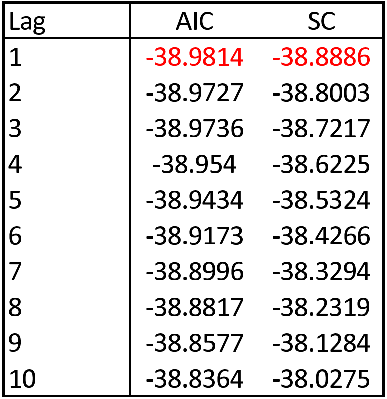
- *Bayesian Information Criterion* (also Schwarz Criterion)

$$SC = \log|\widehat{\Sigma}| + \frac{k'}{T}\log(T)$$

$k' = n \times (n \times p + 1)$ is the total number of coefficients in VAR(p)

$|\widehat{\Sigma}|$ is the determinant of the residual covariance matrix

Example: Optimal Lag Selection



Lag	AIC	SC
1	-38.9814	-38.8886
2	-38.9727	-38.8003
3	-38.9736	-38.7217
4	-38.954	-38.6225
5	-38.9434	-38.5324
6	-38.9173	-38.4266
7	-38.8996	-38.3294
8	-38.8817	-38.2319
9	-38.8577	-38.1284
10	-38.8364	-38.0275

Stability Condition

It requires for the eigenvalues of each relationship matrix B_p to be inside the unit circle (< 1).

Eigenvalue	Modulus < 1
-0.22	0.22
-0.17	0.17
-0.01-0.11i	0.11
-0.01+0.11i	0.11
0.04	0.04
-0.01	0.01

This VAR system satisfies stability condition $|\lambda \mathbf{I} - \mathbf{B}| = 0$.

If $p > 1$, coefficient matrix for each lag B_p to be checked separately.

Cointegration Analysis

Investigates the long-run relationship between **Prices**, also known as error correction. Consider cases:

- Global market indicies, such as FTSE vs DAX: cointegration transpires over the 15-20 year period – daily Prices.

- Sections of the yield curve, such as r_{10Y} and r_{25Y}, GBP ‘LIBOR’ 2013-15.

- Segments of the commodities market, eg, heating oil vs. gas, agricultural commodities.

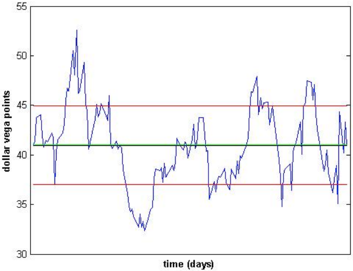
Moving onto **cointegration in equities** (daily Prices)

- Marriott vs IHG was **an M&A situation** Jan 2014 - Jan 2017, when Marrott was looking for an acquisition. More in *Cointegration Lecture*.
- FP Tutorial (forthcoming) Ford vs. GM for 2011-2015 (full years). Cointegrated but spread ‘smoothed’, trades taking up to 6 months duration.
- Learning Cointegration article (Appendix B) has APPL vs GOOG and AMZN vs EBAY for Feb 2012 - Feb 2013. AMZN vs EBAY had a reverting spread, up to 8-10 trades.

Cointegrated System

Prices move together in long term = stationary spread.

$$e_t = \text{Price}_t^A \pm \beta_C \text{Price}_t^B \pm ...$$



Linear combination of stochastic Prices (alike GBM) reduced to one common factor: spread e_t!

Estimating Cointegration - Pairwise

Pairwise Estimation: select two **Prices** (Asset A, Asset B) likely to have a stationary spread: gas vs. heating oil futures, two pharmas in a merger.

- **Step 1.** Regress Asset A price P_t^A on P_t^B, and test the fitted residual by ADF with lag=1. If stationary, proceed.
- **Step 2.** Confirm correction equations for $\Delta P_t^A =$ ..., $\Delta P_t^B =$ ... .
- **Step 3.** Use OU SDE to evaluate mean-reversion: μ_e, σ_{eq}.

$$\mu_e, \pm Z^* \sigma_{eq}$$

Estimating Cointegration – Engle-Granger in detail

Step 1. Obtain the fitted residual and ADF-test for stationarity.

$$P_t^A = \widehat{\mu_e} + \widehat{\beta}_C P_t^B + \widehat{e}_t \qquad \Longrightarrow \qquad \widehat{e}_t = P_t^A - \widehat{\beta}_C P_t^B - \widehat{\mu_e}$$

- Cointegrating vector $\boldsymbol{\beta}'_{Coint} = [1, -\widehat{\beta}_C]$ and equilibrium level is $\mathbb{E}[\widehat{e}_t] = \mu_e$
- **If the residual non-stationary** then no long-run relationship exists and regression is spurious.

Step 2. Plug the residual from Step 1 into **error correction** equation

$$\begin{aligned} \Delta P_t^A &= \phi \Delta P_t^B - (1-\alpha)\widehat{e}_{t-1} \\ \Delta P_t^A &= \phi \Delta P_t^B - (1-\alpha)\underbrace{\left(P_{t-1}^A - \beta_C P_{t-1}^B - \mu_e\right)}_{} \end{aligned}$$

- It is required **to confirm the significance** for $(1-\alpha)$ coefficient.

Statistical Arbitrage

- Signal Generation from Cointegrated Spread
- Fitting to OU Process (how good the mean-reversion is)
- Trade Bounds Optimisation and Backtesting

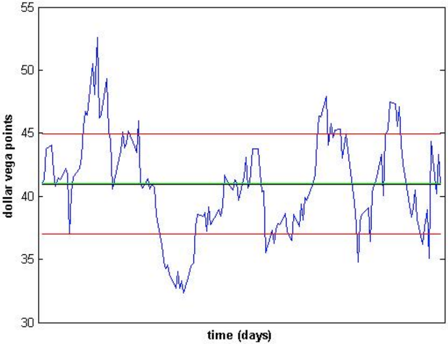
Statistical Arbitrage

Cointegrated prices have a mean-reverting spread $e_t = \boldsymbol{\beta}'_{\boldsymbol{Coint}} \boldsymbol{P_t}$, when it goes *significantly* above/below μ_e, it gives a signal.

1. **How to generate P&L?** Trade design and algorithmic considerations.

2. **How to evaluate P&L?** Drawdown control and backtesting.

Signals from Mean-Reverting Spread



Signal generation and positions for assets A and B.

$$e_t \gg \mu_e \quad \text{enter with} \quad [-100\%\, P^A, +\beta_C\%\, P^B]$$

$$e_t \ll \mu_e \quad \text{enter with} \quad [100\%\, P^A, -\beta_C\%\, P^B]$$

To make the trading systematic and controlled, you will need:

- **Loadings** $\boldsymbol{\beta}_{Coint}$ give positions, the spread is coint residual

$$e_t = P_t^A + \beta_B P_t^B + \cdots + \beta_G P_t^G$$

- **Bounds** $\mu_e \pm Z\,\sigma_{eq}$ give **entry** signal, while **exit** at $e_t \approx \mu_e$

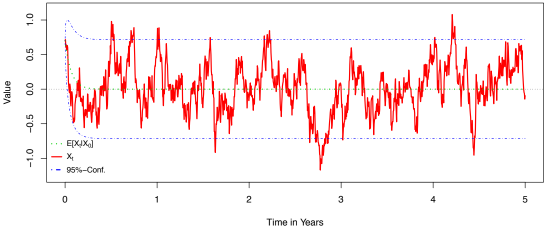
  Instead of assuming $Z = 1$ you can vary in the range $[0.7, 1.3]$ or as fitting to your spread.

- **Half-life** between the crossings $e_t = \mu_e$.

$$\widetilde{\tau} \propto \ln 2/\theta$$

  Average time between exists that fix positive P&L.

OU Process simulated



We consider the process because it generates **mean-reversion**. Your empirical spread e_t might/might not be as good as this.

$$de_t = -\theta(e_t - \mu_e)\,dt + \sigma_{OU}\,dX_t \qquad (1)$$

- $\theta \ll 0$ is the speed of reversion to the equilibrium μ_e
- σ is the scatter of BM diffusion (not of reversion σ_{eq}).

Fitting to OU Process

$$e_{t+\tau} = \left(1 - e^{-\theta\tau}\right)\mu_e + e^{-\theta\tau} e_t + \epsilon_{t,\tau}$$

Two terms of SDE solution: reversion and autoregression

$$e_t = C + Be_{t-1} + \epsilon_{t,\tau} \qquad \text{run a regression}$$

$$e^{-\theta\tau} = B \quad \Rightarrow \quad \boxed{\theta = -\frac{\ln B}{\tau}} \tag{2}$$

$$\left(1 - e^{-\theta\tau}\right)\mu_e = C \quad \Rightarrow \quad \boxed{\mu_e = \frac{C}{1 - B}} \tag{3}$$

Signal-generating Bounds

$$\sigma_{eq} = \sqrt{\frac{\text{SSE} \times \tau}{1 - e^{-2\theta\tau}}} \tag{4}$$

SSE is sum of squared residuals of your regression for e_t, for this AR(1). It represents covariance, SSE $\times \tau = \Sigma_\tau$.

σ_{OU} is parameter of the SDE, Brownian Motion diffusion *over each small dt*. Not needed for trading *per se*.

$$\begin{aligned} \sigma_{OU} &= \sigma_{eq}\sqrt{2\theta} \\ &= \sqrt{\frac{2\theta\,\text{SSE}}{1 - e^{-2\theta\tau}}}. \end{aligned}$$

OU Fit – Model Risk

IN PRACTICE we want to trade with tight bounds$Z < 1$ of the higher frequency spread.

$$\mu_e \pm Z\,\sigma_{eq}$$

For the largest profit per trade, typically $Z > 1.5$, the strategy is prone to the breakouts (partitioning of the coint relationship).

Ex ante testing for regime-change is of little help. Adaptive estimation with Kalman or other filtration means unwanted rebalancing, however.

You are constructing the model (cointegration) as much as you are testing for it. There are a number of ways where model not suitable, typically, (a) spread too tight, below bid/ask spread, and (b) OU process might not fit well.

Before we conclude, the words of wisdom from Fischer Black:

1. “In the real world of research, conventional tests of [statistical] significance seem almost worthless.”

2. “It is better to estimate a model than to test it. Best of all, though, is to explore a model.”

On model risk in time series from American Statistical Society:

1. Running multiple tests on the same dataset at the same stage of an analysis increases the chance of obtaining at least one invalid result.’

2. Selecting one ‘significant’ result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion.

EXTRA. Brief use introduction for multivariate cointegration

Vector Error Correction

Returns are modelled with Vector Autoregression but forecasting is poor.

Prices can be tied up with special error correction equations:

$$\boldsymbol{\Delta P_t} \;=\; \boldsymbol{\Pi}\, \boldsymbol{P_{t-1}} + \boldsymbol{\Gamma_1 \Delta P_{t-1}} + \epsilon_t$$

$\Longrightarrow$ $\boldsymbol{\Pi}$ must have reduced rank, otherwise *rhs* will not balance *lhs*.

Now to make this look alike Engle-Granger, we decompose coefficients $\boldsymbol{\Pi} = \alpha\, \beta'_{Coint}$

$$\boldsymbol{\Delta P_t} \;=\; \alpha\, \underbrace{(\beta'_C\, \boldsymbol{P_{t-1}} + \mu_e)}\, + \boldsymbol{\Gamma_1 \Delta P_{t-1}} + \epsilon_t$$

Cointegrating Vector Estimators β'_{Coint}

	1	2	3	4	5	6	7
Canada	6.78395	-1.96320	-9.07554	7.03629	2.56142	6.25519	-2.08045
France	4.86921	4.86043	-2.08623	-7.28739	2.28808	-1.59825	-1.60875
Germany	-15.76001	-5.94947	0.12170	3.34469	-0.01972	-4.04040	4.24522
Japan	-1.22250	5.52024	-0.70856	1.03285	-0.17938	-0.08242	1.76463
UK	27.19903	-13.06796	-0.55980	-0.36245	-1.03954	-1.76308	0.23821
US	-10.25644	13.17254	7.00734	-0.56186	-5.15207	2.16214	-2.37646
Const	-117.01015	-5.47002	59.45116	-32.77753	5.05186	-8.11528	-7.19582

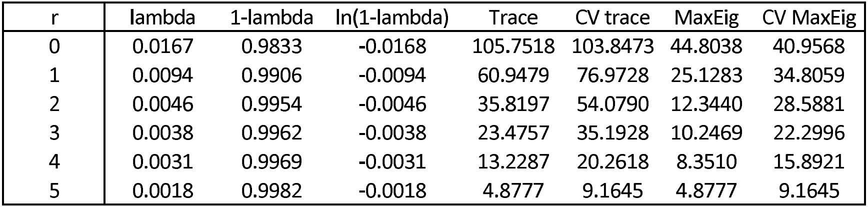
- $n - 1$ columns are linearly dependent on the 1st column.
- $r = 1$ columns of β are cointegrating vectors, take the first column and standardise it.

$$\begin{bmatrix} 1 & 0.7178 & -2.3231 & -0.1802 & 4.0093 & -1.5119 & -17.2481 \end{bmatrix}$$

The allocations $\hat{\beta}'_{Coint}$ provide a mean-reverting spread.

Sequential Testing for Cointegration Rank

Trace Statistic and Maximum Eigenvalue tests rely on eigenvalues of **Π**.



r	lambda	1-lambda	ln(1-lambda)	Trace	CV trace	MaxEig	CV MaxEig
0	0.0167	0.9833	-0.0168	105.7518	103.8473	44.8038	40.9568
1	0.0094	0.9906	-0.0094	60.9479	76.9728	25.1283	34.8059
2	0.0046	0.9954	-0.0046	35.8197	54.0790	12.3440	28.5881
3	0.0038	0.9962	-0.0038	23.4757	35.1928	10.2469	22.2996
4	0.0031	0.9969	-0.0031	13.2287	20.2618	8.3510	15.8921
5	0.0018	0.9982	-0.0018	4.8777	9.1645	4.8777	9.1645

- Trace statistic $H_0 : r = r^*$, and $H_1 : r > r^*$. **Table above $r^* = 1$**

$$LR_{r^*} = -T \sum_{i=r^*+1}^{n} \ln(1 - \lambda_i)$$

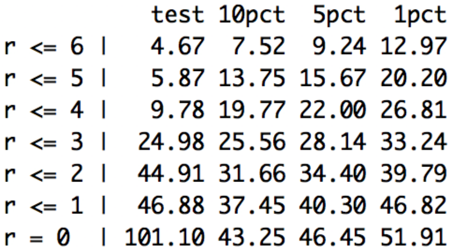
- Maximum eigenvalue statistic $H_0 : r = r^*$, and $H_1 : r = r^* + 1$

$$LR_{r^*} = -T \ln(1 - \lambda_{r^*+1})$$

Implementation Notes - R

Johansen Procedure is a useful **screening tool**. It implements ready multivariate cointegration.

The workhorse is *ca.jo()* function from the **R package** *urca*.



```
           test 10pct  5pct  1pct
r <= 6 |   4.67  7.52  9.24 12.97
r <= 5 |   5.87 13.75 15.67 20.20
r <= 4 |   9.78 19.77 22.00 26.81
r <= 3 |  24.98 25.56 28.14 33.24
r <= 2 |  44.91 31.66 34.40 39.79
r <= 1 |  46.88 37.45 40.30 46.82
r = 0  | 101.10 43.25 46.45 51.91
```

cajorls() presents the output as a set of familiar OLS equations with EC term, separate line for each price.

Implementation Notes - Python

from statsmodels.tsa.stattools import coint

coint(PriceA, PriceB)

Critical values were fixed to MacKinnon(2010) after been wrong for about 2011-2017!

import statsmodels.tsa.stattools as ts

ts.adfuller() gives a rudimentary output for DF Test for stationarity.

Requirement (TS Topic): implement Engle-Granger procedure from the first principles. Enclosed R code gives a complete example.

Implementation Notes - Python (Multivariate)

```
import statsmodels.tsa.vector_ar.vecm as cajo
johansen_test = cajo.coint_johansen(PricesData, 0, 2)
```

Python routines output will be very similar to VECM output from R routines (package *urca*).

To install dev version, use *git()* instead of *pip*. Refer to the source code comments to understand inputs and outputs.

`https://www.statsmodels.org/dev/generated/statsmodels.tsa.vector_ar.vecm.VECM.html`

Relevant Econometric Advances

1. Estimation of regression adaptively, via a state-space model known as Kalman filter, removes the need for rolling parameters `www.thealgoengineer.com/2014/online_linear_regression_kalman_filter/`

   (a) Recursive re-estimation of coint residual $\widehat{e}_t = P_t^A - \widehat{\beta}_C P_t^B - \widehat{\mu_e}$ 'contradicts' the idea of long-term error correction: $\widehat{\beta}_C$ stable.
   (b) But you can apply Kalman filter for the fine-tuning of OU process.

2. `cran.r-project.org/web/views/Robust.html`

3. `cran.r-project.org/web/views/Econometrics.html`

1. Introduction to CQF Final Project

2. Time Series and Cointegration for Pairs Trading

3. Arbitrage: OU Process Signal Generation and Control

4. Trading Strategy Evaluation: Systematic Backtesting, Scorecards

EXTRA. Strategy Backtesting and Evaluation

- Systematic Backtesting. Alpha and Beta
- Trading Efficiency. Ratios and Scorecards

Systematic Backtesting

1. We will look at **how to relate P&L** to the market and factors, to understand what drives P&L, what you make money on.
2. Then, we will talk about **evaluating P&L** with drawdown control and VaR.
3. You can look for suitable models for algorithmic **order flow** and liquidity impact. [Optional]

Alpha and Beta

Beta is the strategy's market exposure, for which you should not pay much as it is easy to gain by buying an ETF or index futures contract.

Alpha is the excess return after subtracting return due to market movements.

$$R_t^S = \alpha + \beta R_t^M + \epsilon_t$$

$$\mathbb{E}[R_t^S - \beta R_t^M] = \alpha$$

$R_t^M = R_t - r_f$ is the time series of returns representing **the market factor**.

Risk-Reward Ratios

Information Ratio (IR) focuses on risk-adjusted *abnormal* return, the risk-adjusted alpha!

$$\frac{\alpha}{\sigma(\epsilon)}$$

(That doesn't tell us how much dollar alpha is there. It can be eaten by transaction costs.)

Sharpe Ratio measures return per unit of risk. Familiar form:

$$\frac{\mathbb{E}(R_t - r_f)}{\sigma(R_t - r_f)}$$

Factors

Evaluating performance **against factors** is the central part of the backtesting.

We saw the separation of alpha and beta in regression *wrt* one market factor

$$R_t^S = \alpha + \beta R_t^M + \epsilon_t$$

We see that a factor is a time series of changes, similar to the series of asset returns.

Named Factors

- **Up Minus Down** (UMD) or **momentum** factor would leverage on stocks that are going up. The recent month's returns are excluded from calculation to avoid a spurious signal.
- **Small Minus Big** (SMB) factor shorts large cap stocks, so β^{SMB} measures the tilt towards small stocks.
- Long-short **High Minus Low** (HML) or **value** factor: buy top 30% of companies with the high book-to-market value and sell the bottom 30% (expensive stocks).

1) Except for HML, the impact/presence of other factors questionable.
2) Since 2015, Fama-French moved to 5-factor model that include profitability RMW and investment CMA but ignore the proper 3) Momentum factor and 4) Low Volatility (Betting Against Beta) factors.

Factors Backtesting

So how do we check against those factors?

Set up a regression!

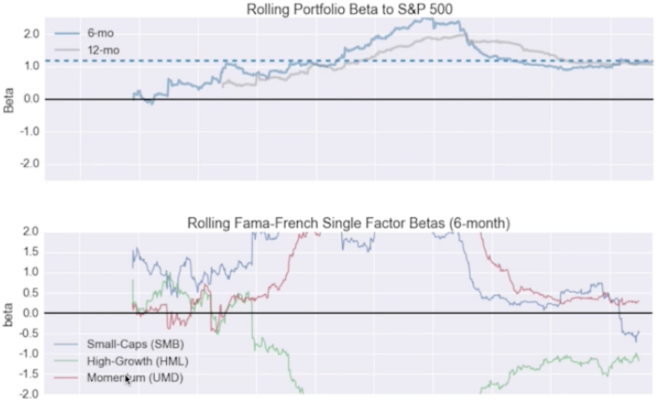
$$R_t^S = \alpha + \beta^M R_t^M + \beta^{HML} R_t^{HML} + \epsilon_t$$

where R_t^{HML} is return series from the long-short HML factor.

- We can **add factors** to this regression.
- We can have **rolling estimates** of these betas for each day/week.

Factors Backtesting (Advanced)

- Scale returns to have the same volatility as the benchmark – put on the same plot for correct comparison.
- Rolling Sharpe Ratio – changes **not** desirable).
- Rolling market factor beta – $\beta > 1$ **not** desirable.
- Rolling betas *wrt* to UMD (momentum), SMB, and industry sectors.



From: *Portfolio and Risk Analytics with PyFolio*, T. Wiecki, QI 2015

Drawdowns

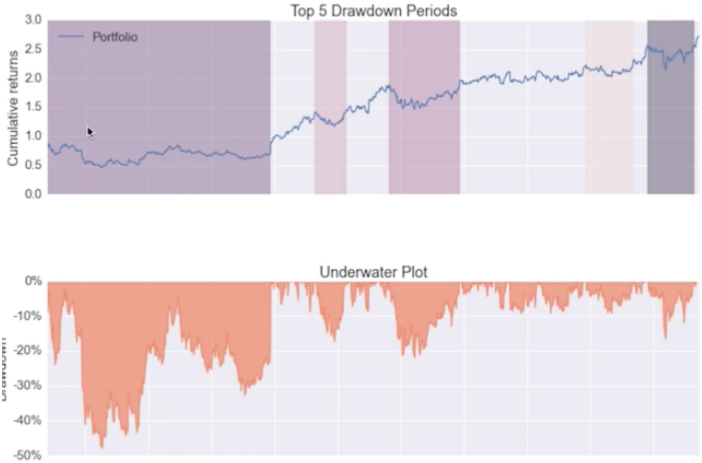
The drawdown is the cumulative percentage loss, given the loss in the initial timestep.

Let's define the highest past peak performance as High Water Mark

$$\mathrm{DD}_t = \frac{\mathrm{HWM}_t - P_t}{\mathrm{HWM}_t}$$

where P_t is the cumulative return (or portfolio value Π_t).

It makes sense to evaluate a maximum drawdown over past period $\max_{t \leq T} \mathrm{DD}_t$.



From: Quant Insights, Oct 2015, *Portfolio and Risk Analytics with PyFolio*, Thomas Wiecki (Quantopian)

Drawdown Control

The strategy must be able to survive without running into a close-out.

It makes sense to pre-define Maximum Acceptable Drawdown (MADD) and trace

$$\mathrm{VaR}_t \leq \mathrm{MADD} - \mathrm{DD}_t$$

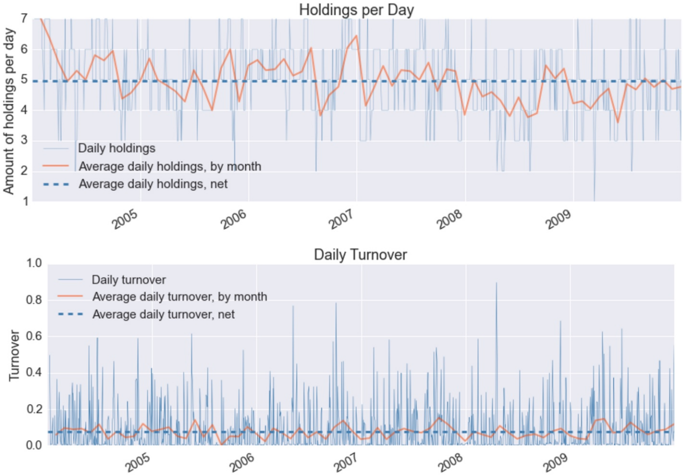
where VaR_t is today's VaR and DD_t is current drawdown.

Backtesting Risk and Liquidity - Summary

1. Does cumulative P&L behave as expected (eg, for a cointegration trade)? Behaviour of risk measures (volatility/VaR/Drawdown)?

2. Is P&L coming from a few large trades or many smaller trades? Does all profit come from a particular period. Concentration in assets and its attribution – as intended?

3. Turnover (good or bad?), impact of transaction costs (slippage). Plot the P&L value (or its alpha) vs. number of transactions.



From: Quant Insights, Oct 2015, *Portfolio and Risk Analytics with PyFolio*, Thomas Wiecki (Quantopian)

Algorithmic Flow

3. Optionally, introduce liquidity and algorithmic flow considerations (a model of order flow). How would you be entering and accumulating the position? What impact *your transactions* will make on the market order book?

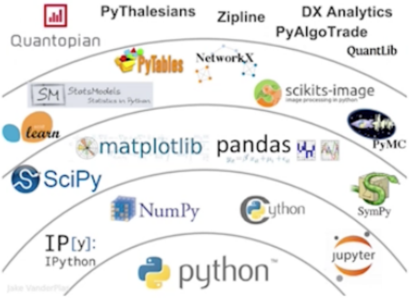
4. Related issue is the possible leverage for the strategy. While the maximum leverage is $1/\text{Margin}$, the more adequate solution is a maximally leveraged market-neutral gain or alpha-to-margin ratio

$$\text{AM} = \frac{\alpha}{\text{Margin}}.$$

Python Ecosystem

The Quant Finance PyData Stack

Source: [Jake VanderPlas: State of the Tools](https://www.youtube.com/watch?v=5GINDD7qbP4)



`https://www.cqf.com/about-cqf/program-structure/cqf-qualification/advanced-electives`

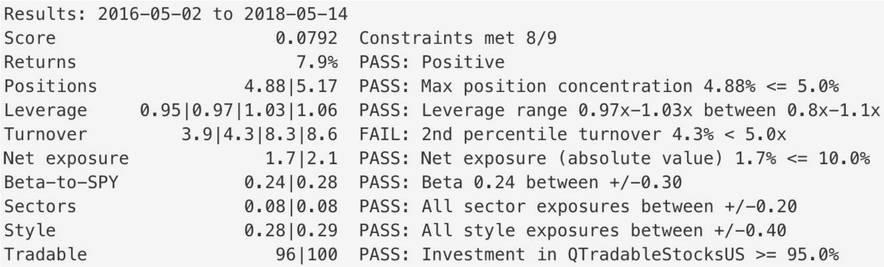
1. https://github.com/quantopian/PYFOLIO
   https://quantopian.github.io/pyfolio/notebooks/single_stock_example/

2. https://github.com/quantopian/ALPHALENS
   https://github.com/quantopian/alphalens/blob/master/alphalens/examples/alphalens_tutorial_on_quantopian.ipynb

Github examples above to showcase the useful aspects of backtesting: **a.** rolling beta *wrt* S&P500 plot, **b.** rolling Sharpe Ratio plot, and **c.** various Ratios in scorecards.

Instead of using those ready packages you will code **own** analytics. *PYFOLIO* package is deprecated, no longer maintained.

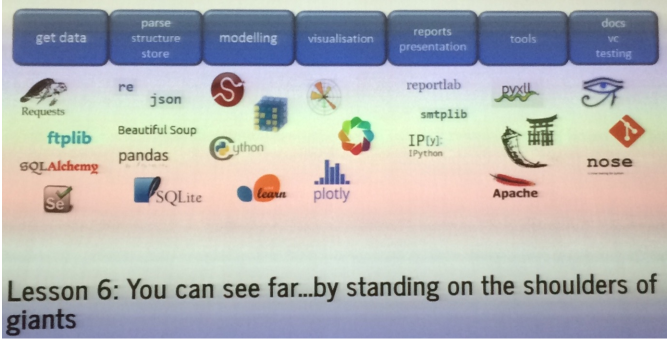
Quick Algo Checker



```
Results: 2016-05-02 to 2018-05-14
Score                        0.0792  Constraints met 8/9
Returns                        7.9%  PASS: Positive
Positions                 4.88|5.17  PASS: Max position concentration 4.88% <= 5.0%
Leverage        0.95|0.97|1.03|1.06  PASS: Leverage range 0.97x-1.03x between 0.8x-1.1x
Turnover            3.9|4.3|8.3|8.6  FAIL: 2nd percentile turnover 4.3% < 5.0x
Net exposure                1.7|2.1  PASS: Net exposure (absolute value) 1.7% <= 10.0%
Beta-to-SPY               0.24|0.28  PASS: Beta 0.24 between +/-0.30
Sectors                   0.08|0.08  PASS: All sector exposures between +/-0.20
Style                     0.28|0.29  PASS: All style exposures between +/-0.40
Tradable                     96|100  PASS: Investment in QTradableStocksUS >= 95.0%
```

From inaccessible: `https://www.quantopian.com/posts/contest-constraint-check-notebook-with-compact-output`

Developing a Trading Business



Lesson 6: You can see far...by standing on the shoulders of giants

There are libraries for anything: data download, regression and ML, backtesting and tear sheets/trading analytics.

Building an Energy Trading Business from Scratch

Teodora Baeva (BTG Pactual), Quant Insights 2015 (CQF Institute event).

END OF WORKSHOP