Cointegration Strategy

Technical Report

Stratagem Technologies

February 15, 2018

1 Background

1.1 Cointegration

The basic intuition of the signal is that over a given look-back window there may exist a (temporary) *cointegrating* relationship between two time series (e.g. price processes, scoring rates, etc...) whose expected stability may be exploited. Formally, the cointegration property means that there exists a linear combination (portfolio) whose implied "spread series" is stationary and thus has a mean and variance that do not change with time (time-homogeneous).

1.2 Strategy

The fundamental building blocks of the strategy are very simple (see Fig **??**):

1. compute the cointegration signal;
2. apply a sequence of filters and transformations to the signal;
3. open a position based on the anticipated direction of the spread;
4. close the position to lock in gains or losses.

Each of the strategies considered herein — for which the signals generation processes are described in Chapter **??** — have this structure. The real complexity arises in the "Signal" and "Transformers | Filters" nodes which contain our main decision making logic. For example, a typical strategy might exploit the implied mean reversion in the spread by re-cast the spread series into a vector of standard scores[1] and using the current level of deviation from the mean as an indication of the future direction; i.e. where a large score would imply a negative trend due to reversion to the mean and vice-versa.

[1] Typically computed by evaluating $\frac{x-\mu_x}{\sigma_x}$.

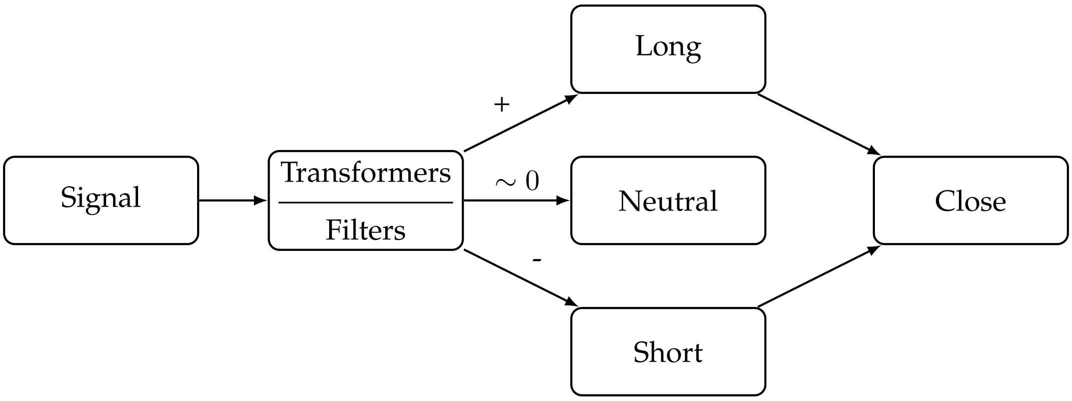


Figure 1.1: Structure of generic trading strategy which attempts to capture short-term trends in an asset's valuation.

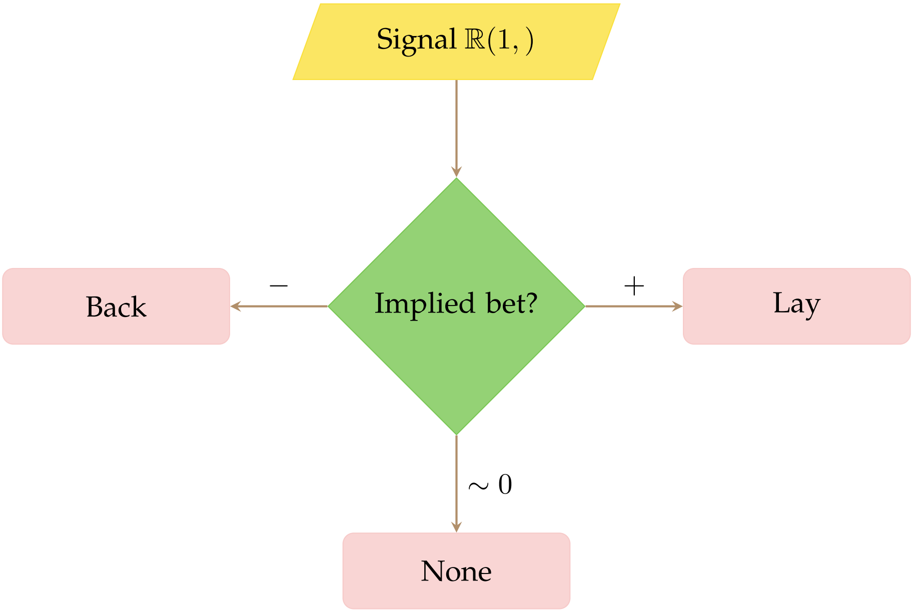


Figure 1.2: Illustration of generic signal to trade decision logic.

1.2.1 Interpreting signal values

The signal, $\boldsymbol{\beta} \in \mathbb{R}$, takes the form of a real valued vector of length m, where m is the number of markets in the portfolio. The sign and magnitude of each value dictates how to trade on each of the m markets (equivalently, stickers): negative values indicate short positions and, by symmetry, positive values indicate long positions (see Fig. **??**). The absolute value is then used to assign relative sizes to the trades in the portfolio; note that values close to zero should likely be ignored since the resulting size is often negligible.

1.2.2 Common strategy features

Naming The cointegration strategy is written to support many underlying mechanisms for generating a long/short signal based on an implied spread series. Further, it arbitrarily supports any market pair for $m \geq 2$. As such we need a clear naming convention that indicates exactly *what* we're trading on, and *what* method we're using to derive the signal. The following conventions are imposed:

For readability, it is suggested that the strategy name include "coint". For now the only concrete implementation of the strategy, based on market price processes, has the name: **coint_mpc**.

The strategy descriptor defines the method used inside the signal generator to extract a direction (prediction) from the spread series.

The strategy code should be used to add final details to the cointegration strategy such as which market to trade on.

This can be used to distinguish even further in case you wanted to do some specific tests, say with . Though the value of does not affect anything in the code, it does allow for manual debugging and filtering in the database.

Parameters Though each iteration of the cointegration strategy may have it's own set of parameters, some are common throughout:

the number of periods to include in the look-back window.

the sub-sampling rate for the look-back window; i.e. for a value 2, we sample every other value[2].

$\underline{t}$ the minimum holding time for a given portfolio; for practical reasons this is lower bounded by the in-play delay.

T the maximum holding time for a given portfolio; this is only comes into effect when no non-zero signal is observed.

r_{TP} The take-profit threshold on expected portfolio returns.

r_{SL} The stop-loss threshold on expected portfolio returns.

[2] The sub-sampling rate is especially important when using computationally intensive methods for predicting changes in the spread series. For our purposes this will not be the case, though it is still informative to investigate it's impact on returns.

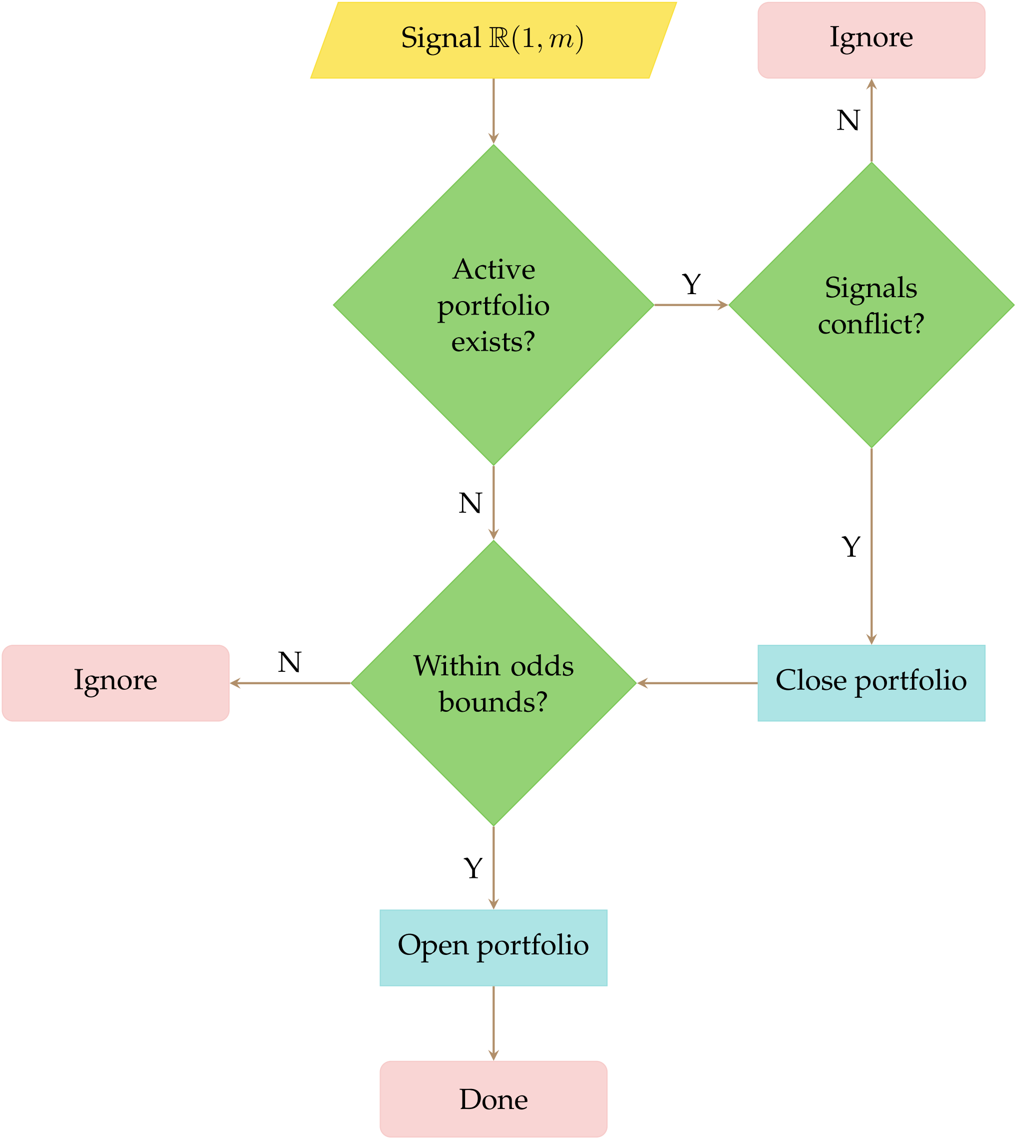


Figure 1.3:

Add description of this diagram and why it’s common to all strategies.

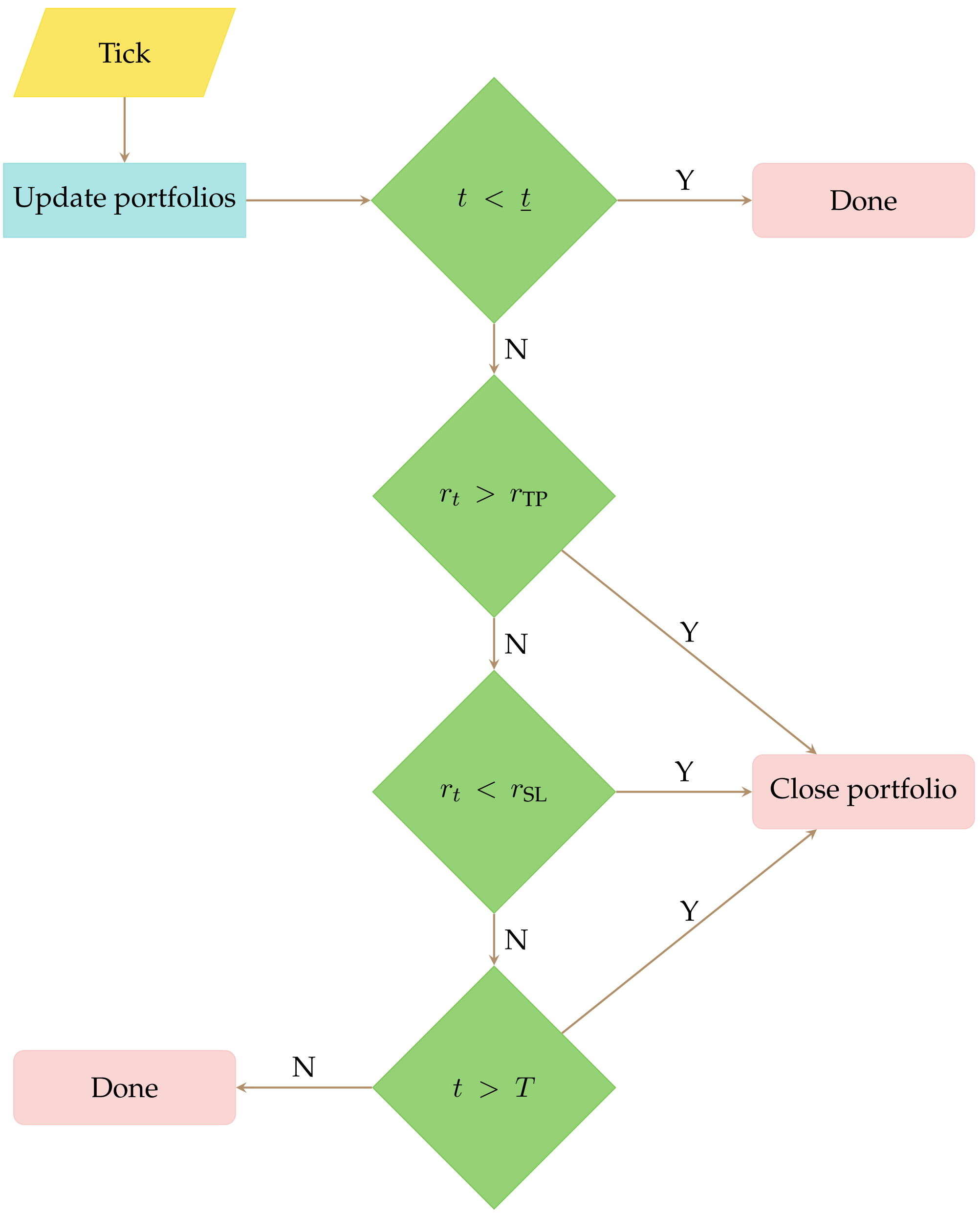


Figure 1.4:

Add description of this diagram and why it's common to all strategies.

2 Basis Estimation

A robust estimate of the basis vector is crucial for developing a trading strategy with any confidence. While the Johansen procedure is very efficient and produces good estimates in theory, the time series used in practice are often very noisy and have periods of intermittent cointegration (structural breaks) which can damage our overall estimate of the relationship. To improve our estimate we use a system of "weak" experts and combine them to form robust basis vector with, empirically speaking, much better convergence properties.

The weak estimators are formed by applying the Johansen procedure over varying data slices, such as a moving window for using the entire observed dataset up to time t. One may then apply smoothing, such as (exponentially-weighted) moving averages or Bayesian parameter estimation procedures, on top of this to further improve robustness of each individual expert.

2.1 Normal-Normal model

We consider a univariate Gaussian model for the observed values of the basis vector, $\mathcal{N}(\mu, \sigma^2)$, in which the variance is assumed known but the mean is estimated. We let $D = (x_1, \ldots, x_n)$ be the data such that the likelihood takes the form,

$$\begin{aligned} p(D \mid \mu, \sigma^2) &= \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2), && (2.1) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]. && (2.2) \end{aligned}$$

Assuming a constant σ^2, we have the following relationship,

$$\begin{aligned} p(D \mid \mu, \sigma^2) &\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right], && (2.3) \\ &\propto \mathcal{N}(\bar{x} \mid \mu, \frac{\sigma^2}{n}). && (2.4) \end{aligned}$$

For simplicity, we use the natural conjugate prior for $p(\mu) \propto \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$ to obtain nice closed-form solutions for the estimate updates. After some simple manipulation one obtains the following update rules:

$$\begin{aligned} \sigma_n^2 &= \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}, && (2.5) \\ \mu_n &= \sigma_n^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right). && (2.6) \end{aligned}$$

2.1.1 Good priors

It's important that a good prior, $p(\mu)$, is used in order to get an practicable estimation of the basis early on. To do this, we look at the historical distributions of basis vectors (Tab. **??**) and assume

	μ_0	σ_0	m	c
Whole match	0.4920	0.0634	0.0024	0.4986
First quarter	0.4954	0.1188	0.0033	0.5067
Second quarter	0.4950	0.1138	0.0038	0.4893
Third quarter	0.4890	0.1132	0.0035	0.4865
Fourth quarter	0.4915	0.1121	0.0016	0.5000
First half	0.4925	0.0938	0.0031	0.4959
Second half	0.4922	0.0834	0.0028	0.5025

Table 2.1: Emprirical prior distributions over the cointegration basis vector of basketball team scores between 06/2015 to 11/2017.

that they are indicative of future relationships. For completeness, we include the distributions for different periods of each match, where μ_0 and σ_0 are the mean and standard deviation on the prior, respectively.

To further improve our estimate we can consider the relationship between the natural handicap on the match and the means of the given basis distributions. The gradient, m, and intercept, c, may then be used to form an adjusted estimate of μ_0, given by the following,

$$\mu_0 \sim m \cdot \mathrm{NH} + c, \tag{2.7}$$

where NH is the natural handicap.

2.2 Combining estimators

Based on Gareth's voting scheme technical report, we used a minimum variance estimator (MVE) to combine the weak experts. This assumes that the individual estimates are unbiased, which is likely untrue, but forms a decent first method for combining in a single unbiased strong estimator. In practice one may know a priori that some estimators should be weighted higher but we leave this is for future work.

The MVE estimator produces the following weights for a system of K estimators,

$$w_i = \frac{1/\sigma_i^2}{\sum_{k=1}^{K}(1/\sigma_k^2)}, \tag{2.8}$$

which yields a weighted average for the mean, with variance given by,

$$\sigma^2 = \left(\sum_{k=1}^{K} \frac{1}{\sigma_k^2} \right)^{-1}. \tag{2.9}$$

2.3 Case Study

To illustrate the effectiveness of the MVE estimator, we present an example[1] using a basketball match between the Dallas Mavericks and the New York Knicks.

[1] Note that the results do carry across matches, this is not just a one off.

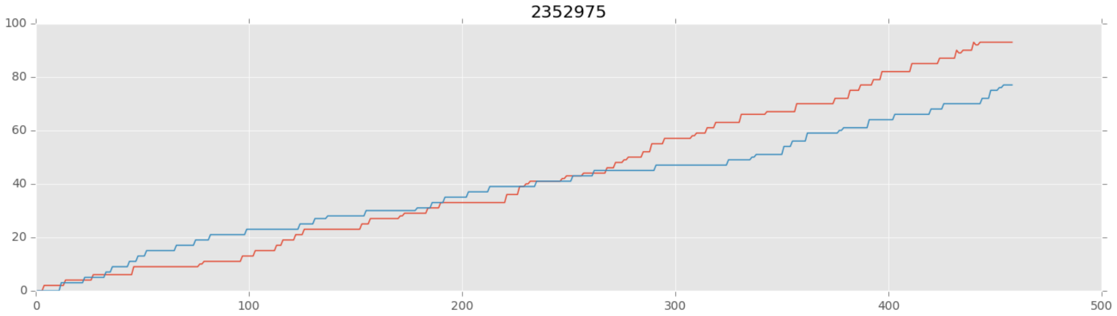
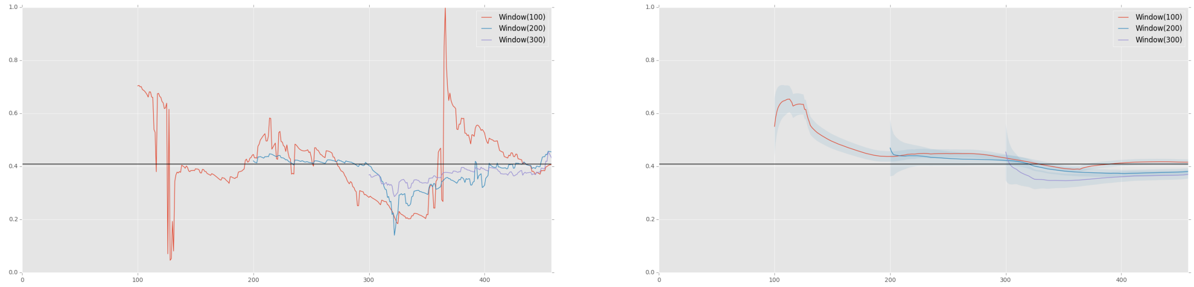


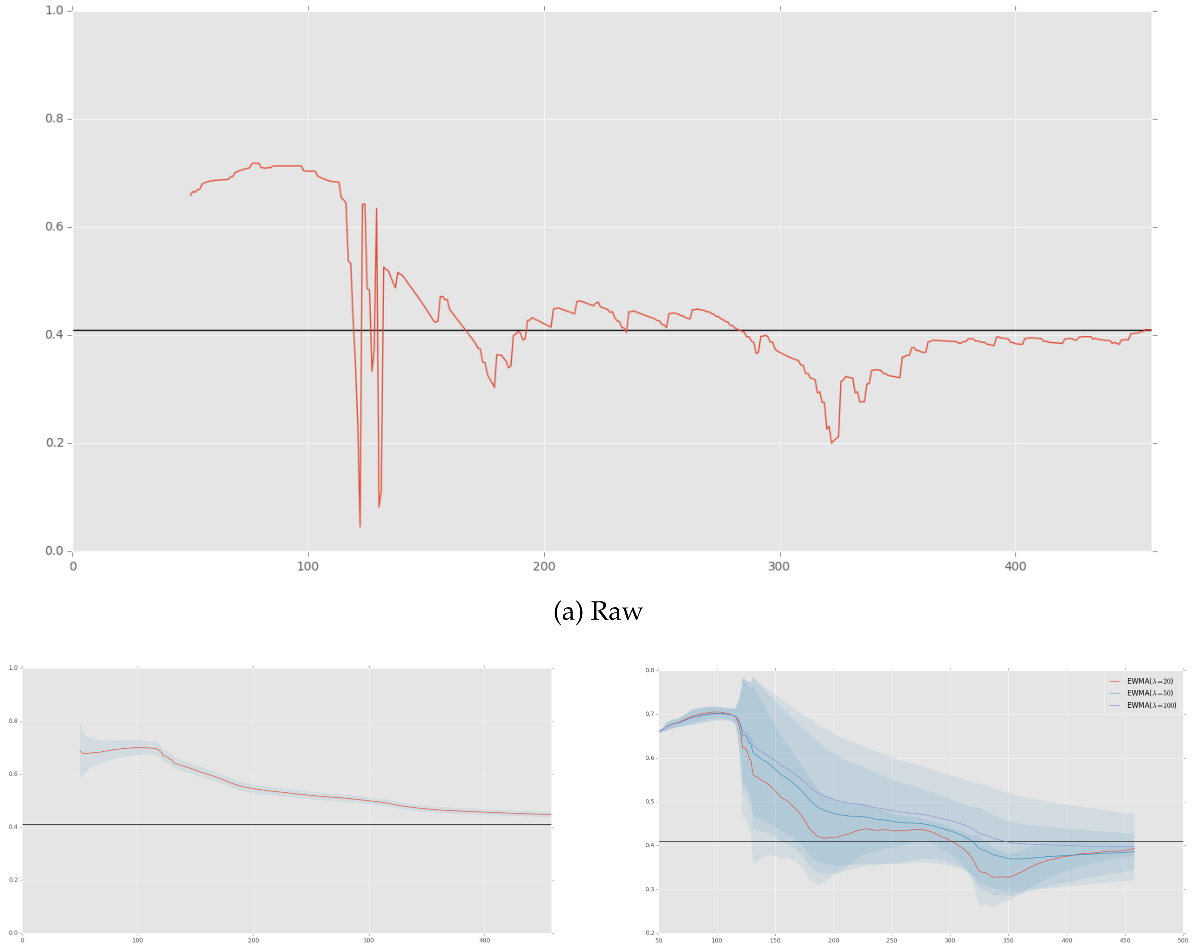
Figure 2.1: Scores of teams A (red) and B (blue) for the basketball match with event id 2352975.



(a) Raw

(b) Bayes

Figure 2.2: Estimations of the cointegration basis (first component) given using a moving window.



(a) Raw

(b) Bayes

(c) Exponentially-weighted moving average

Figure 2.3: Estimations of the cointegration basis (first component) given using an accumulated window of all observed estimations.

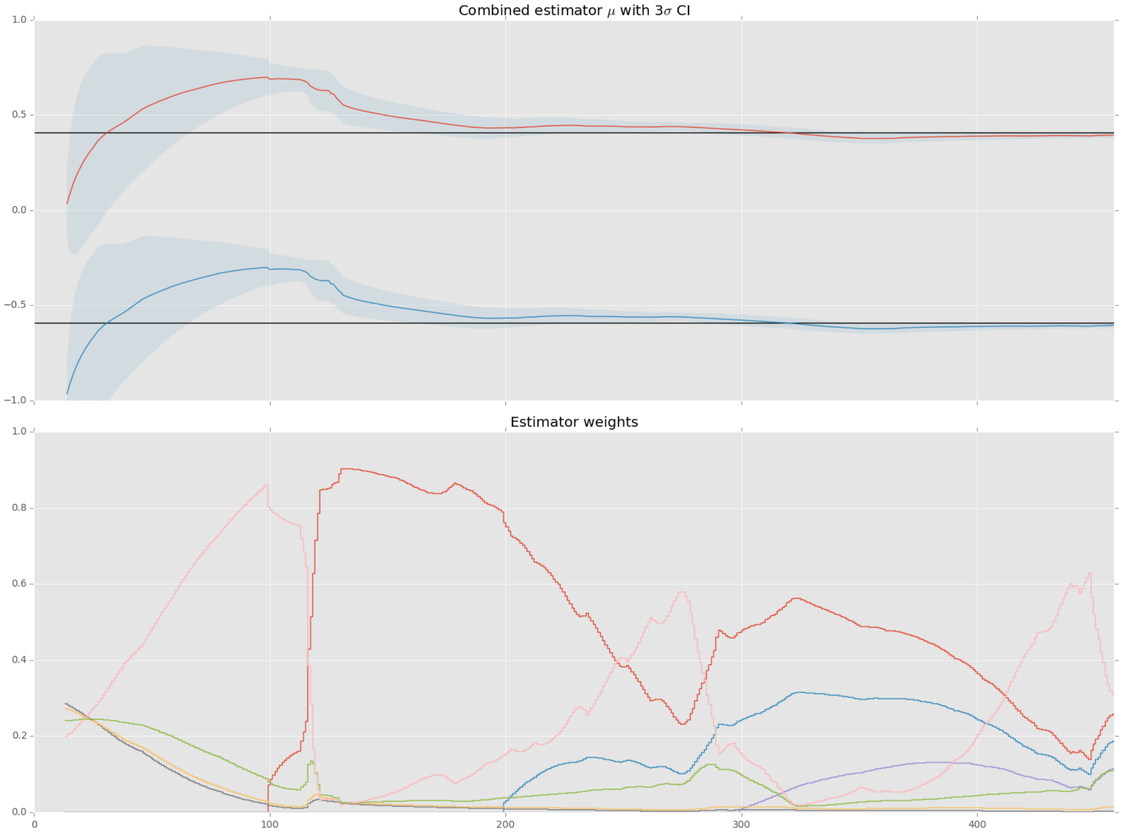


Figure 2.4: Minimum variance estimation of the cointegration basis using a combination of Bayesian windowed estimators and exponentially-weighted moving average cumulative estimators.

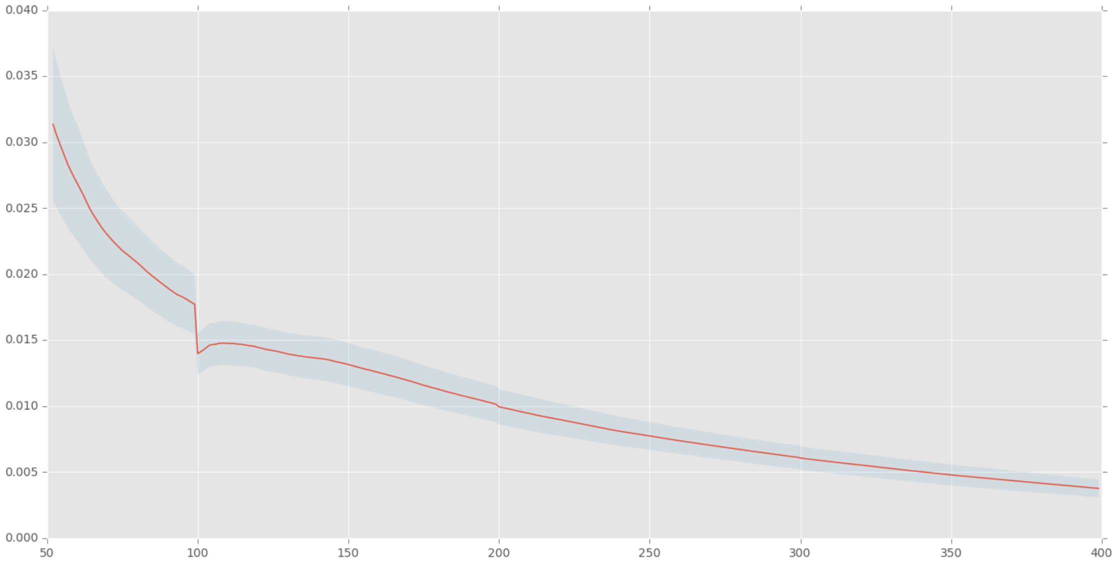


Figure 2.5: Distribution of the mean squared error of the estimation against the best estimate across all recorded matches in the period 06/2015 to 11/2017. The mean (red line) is quoted along with the 3σ confidence interval.

3 Strategies

3.1 Common

Thus far, the strategies have been constrained to multi-sport variants, none have been specific to a single sport; though the next step would be to exploit any causal relationship between the cointegration spread of basketball team scores and market prices. These are all located in the repository. Some of the common utilities that you should probably make use are also in the repo. The main ones to use and extend are:

1. — This utility (see sgmsystematic/trading/providers/model) takes tick data as input (not necessarily prices) and return the basis vectors for each of the stickers in the group. It makes use of the static methods in the class (see sgmresearchbase/analysis/cointegration).

2. — This is the main component used to generate trading signals for the MPC trading strategy. It makes use of subclasses of the class and converts tick data into bases and finally into a directional trading signal to be used by the trader.

3. — This class keeps track of a position opened on multiple markets and computes the expected return given any partial/full executions and the current market prices. It also has some utility plotting methods which are especially useful for debugging and evaluating performance online.

I've included as many useful utilties for generating signals, tracking ticks and smoothing estimations, etc, in the research base repo. All are fully documented and tested. I would advise going over the MPC strategy first and getting an idea about the pipeline to convert tick data into an actual trading strategy. The next task will likely want to be to incorporate the improved estimator (for which there are many possible parameterisations) and to build and investigate a strategy deriving from the team scores and exploiting causul relationships with the market.

3.1.1 Notes from discussions with Gareth

Some of the main points are:

1. Improve the mean reversion signal by using robustified Bollinger bands and/or control charts rather than a simple estimation of the mean.

2. Investigate the best ways to sample the team scores and thus the market prices.

3. Investigate causality between scores and prices using Granger causality and/or an explicit model of the lead-lag relationship.

3.2 MPC

The main strategy developed was the market price cointegration (MPC) strategy.

Parameters

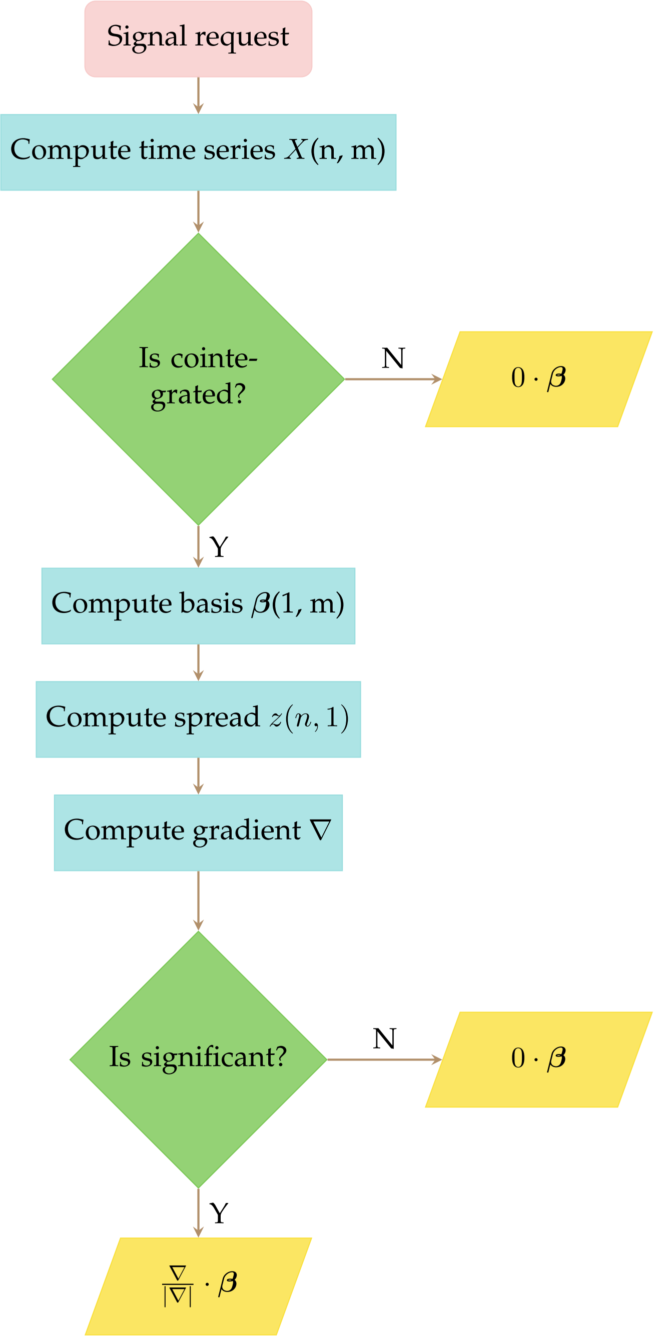
The strategy descriptor tells the framework which pair to trade and provides arguments to that type. For example, one pair that has been implemented is , which corresponds to the full-time moneyline and total games over/under markets. We may also choose to provide parameters to this pair, the exact specification of which are dictated by the implementation. Consider, for example, trading FT12, A versus B, on exclusively ATP matches. Such a code would take the following form: **ft12_tennis.ATP**. Note that the pair and it's arguments come before and after the period, respectively. Some further examples are given below:

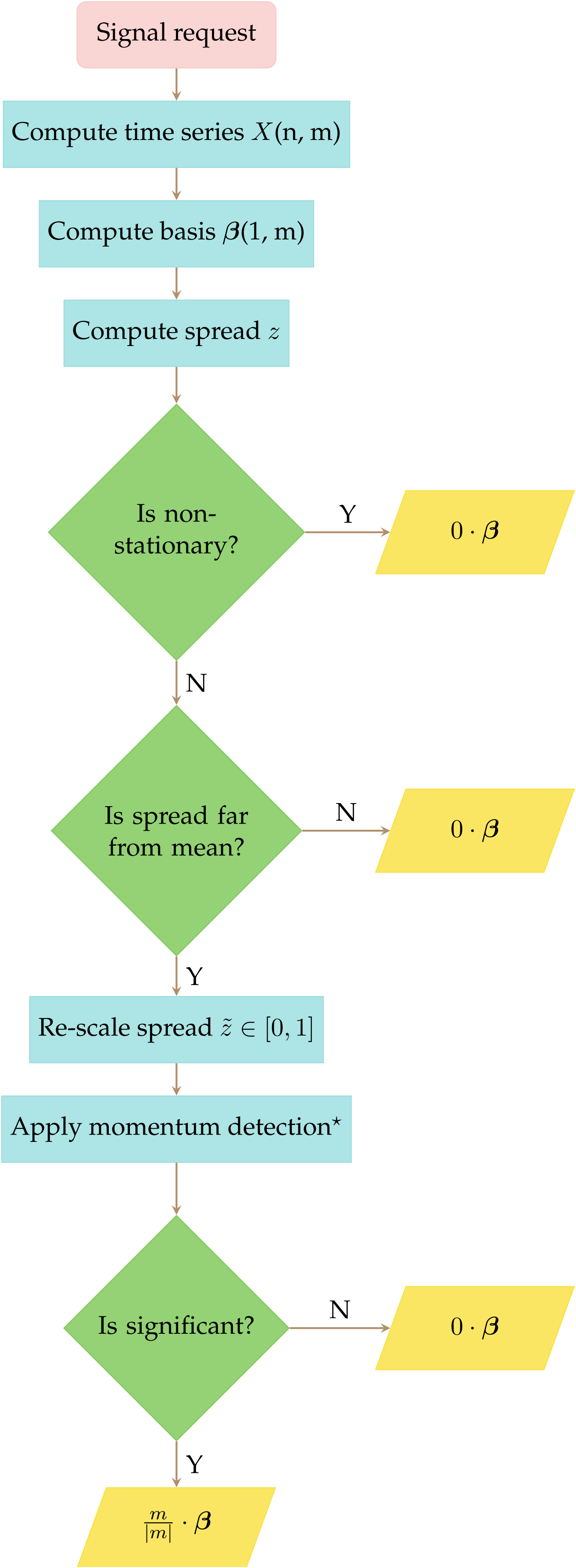
- **ft12_tennis.ATP** — full-time moneyline, A versus B, on ATP matches.
- **ft12_tgou.ATP** — full-time moneyline versus total games over/under, on ATP matches.

3.2.1 Mark 0 (COR)

3.2.2 Mark 1 (MR)

3.2.3 Mark 2 (Hybrid)


Signal request
Compute time series X(n, m)
Is cointegrated?
N
$0 \cdot \boldsymbol{\beta}$
Y
Compute basis $\boldsymbol{\beta}$(1, m)
Compute spread $z(n, 1)$
Compute gradient ∇
Is significant?
N
$0 \cdot \boldsymbol{\beta}$
Y
$\frac{\nabla}{|\nabla|} \cdot \boldsymbol{\beta}$


Signal request
Compute time series X(n, m)
Compute basis $\boldsymbol{\beta}$(1, m)
Compute spread z
Is non-stationary?
Y
$0 \cdot \boldsymbol{\beta}$
N
Is spread far from mean?
N
$0 \cdot \boldsymbol{\beta}$
Y
Re-scale spread $\tilde{z} \in [0, 1]$
Apply momentum detection*
Is significant?
N
$0 \cdot \boldsymbol{\beta}$
Y
$\frac{m}{|m|} \cdot \boldsymbol{\beta}$

A Appendix

A.1 Preliminary Evaluation

For simplicity, we begin with an evaluation of the signal over A/B outright winner market pairs for tennis ATP matches between 2017/08/01 and 2017/09/30 (amounts to a total of 347 events where 220 are actually traded). Though the existence of cointegration within this pairing is unclear, there is a much stronger guarantee of liquidity which makes for less sparse backtest results.

The following assumptions are made about the environment

1. Zero in-play delay.
2. Immediate execution at either the mid-price or top of the book — this also implies that we do not consider prevalent effects such as adverse selection.
3. Perfect connection such that we experience zero latency and no loss of messages in transit to the exchange.
4. We do not consider path dependency (perhaps due to randomness) or any of the complexities associated with cashing out, amounting to an assumption of infinite liquidity.

A.1.1 Mid-price execution

Here we assume that the agent is executed favourably at the mid-prices of the two legs. Thus, the strategy effectively tracks the spread series, taking a long, short or neutral position based on the cointegration signal.

A few key observations:

1. For the short holding period we consider here (until the next time step) short lookback windows are favourable (Fig. **??**).
2. The sub-sampling rate appears to have a considerable effect on the performance, though this relationship is not the same of all lookback windows. Crucially, for the 200 period window, larger values for the sampling-rate (sampling less) have lower associated variance.
3. The proportion of wins to losses behaves in the opposite way, where larger values of the sub-sampling rate have increased variance and a greater proportion of outlier cases (Fig. **??**).
4. The consistency ratio appears to be best for larger lookback windows and low sampling rates.

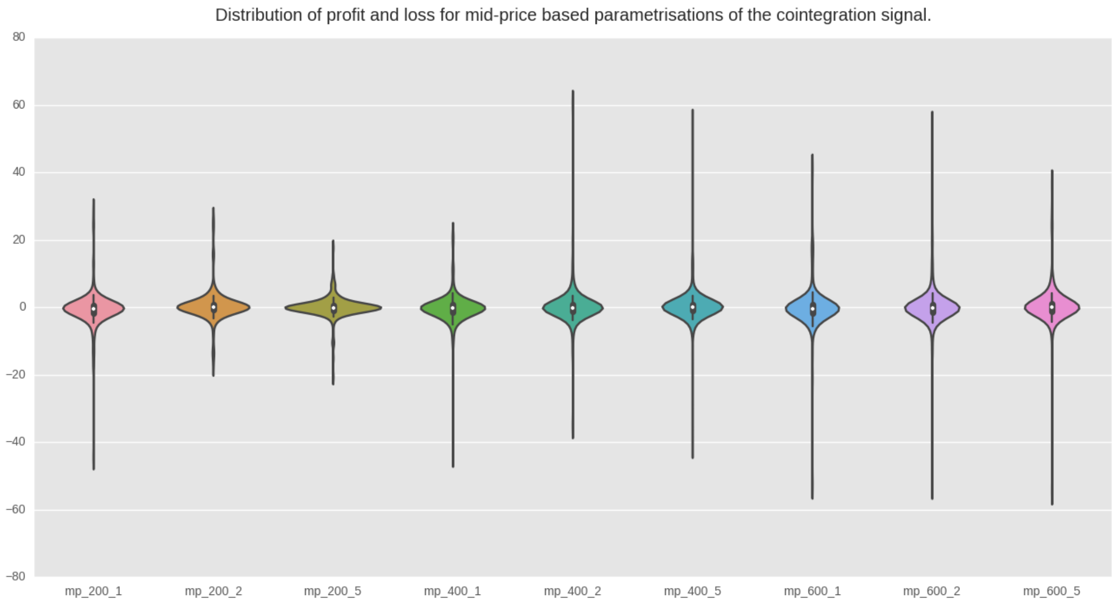


Figure A.1

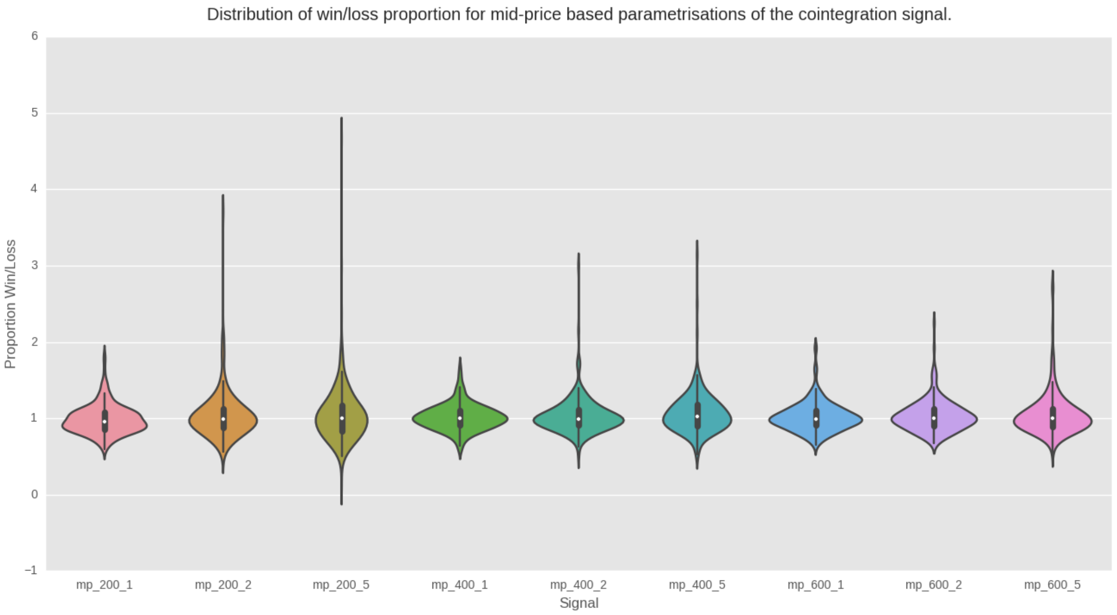


Figure A.2

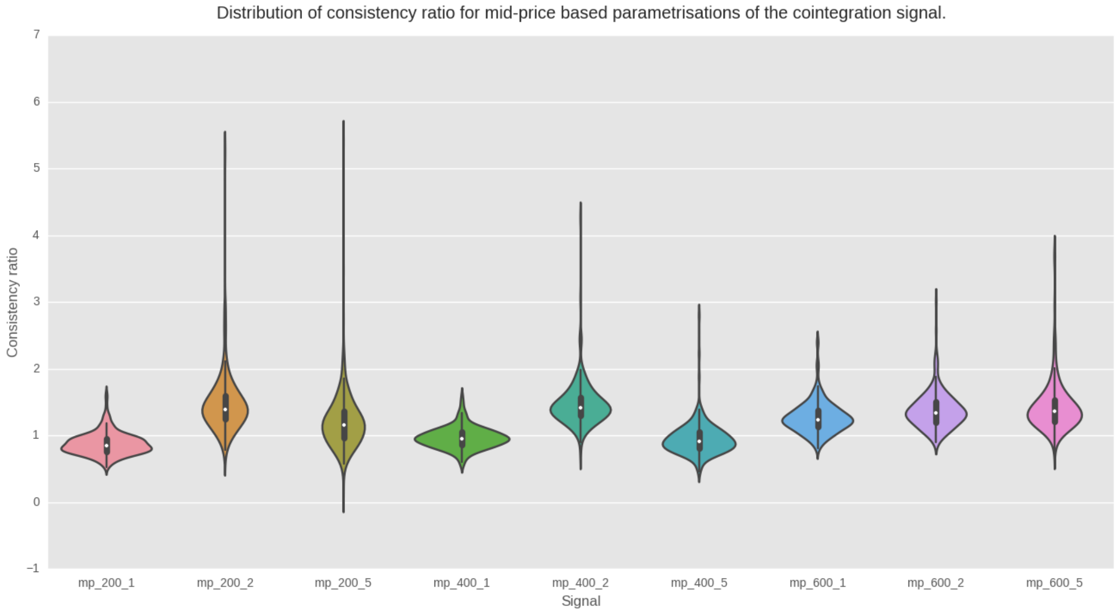


Figure A.3

A.1.2 Spread crossing execution

Here we assume that the agent always crosses the spread, walking the book in each of the two legs.

A few observations:

1. The strategy performs very poorly when crossing the spread on every trade opportunity. This is likely due to the fact that the holding time is very short, so the cost of entry/exit are applied very frequently.

2. Win/loss proportion is also very poor. Further the consistency ratio shows that the strategy was never profitable on a step-by-step basis. Again, we need longer and/or more informed holding times.

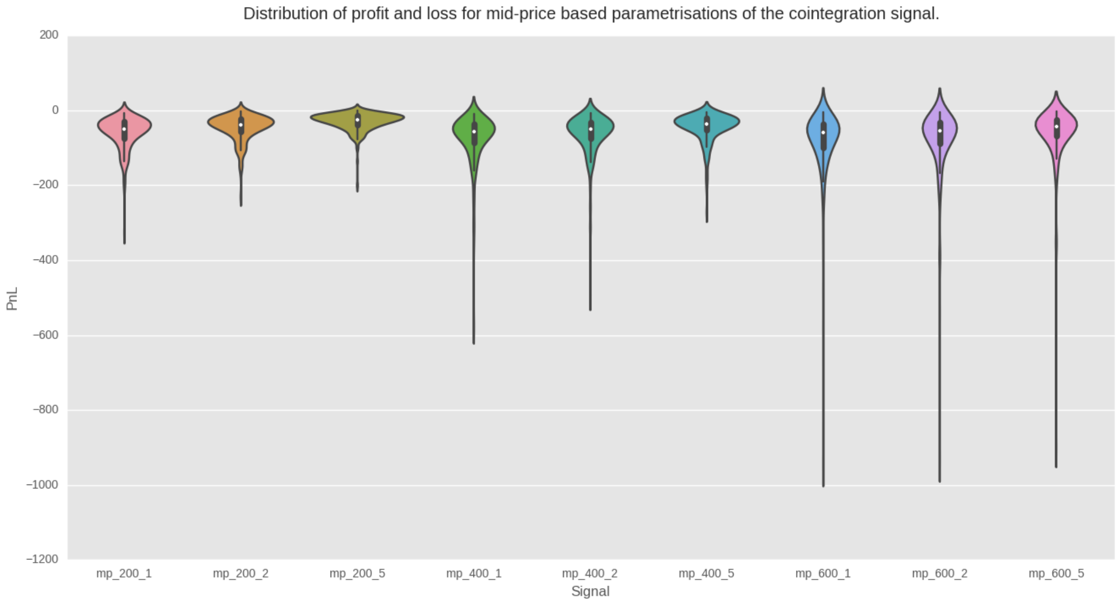


Figure A.4

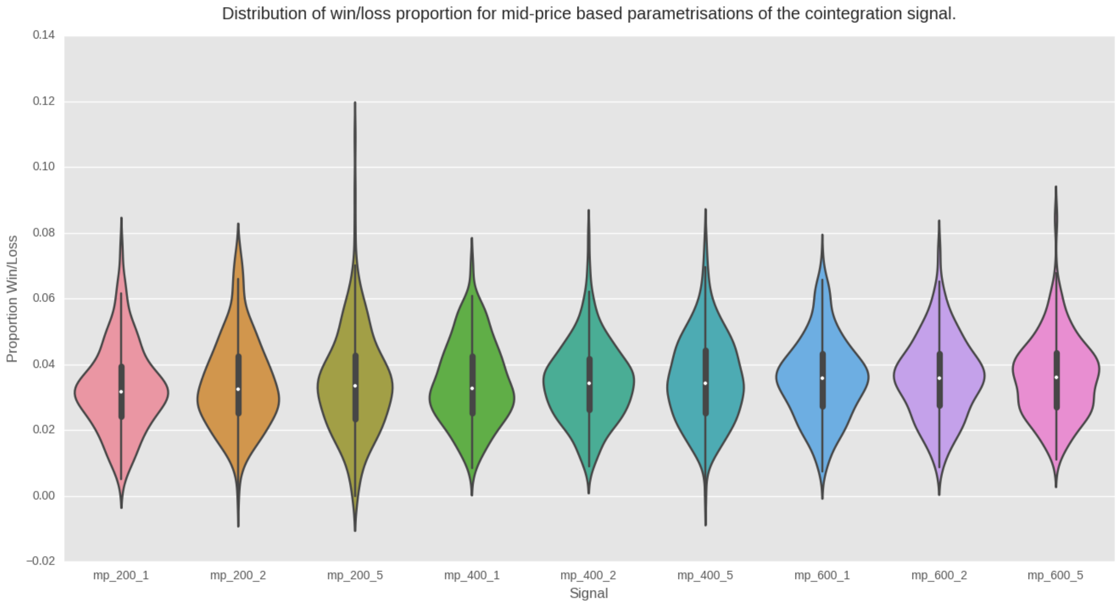


Figure A.5

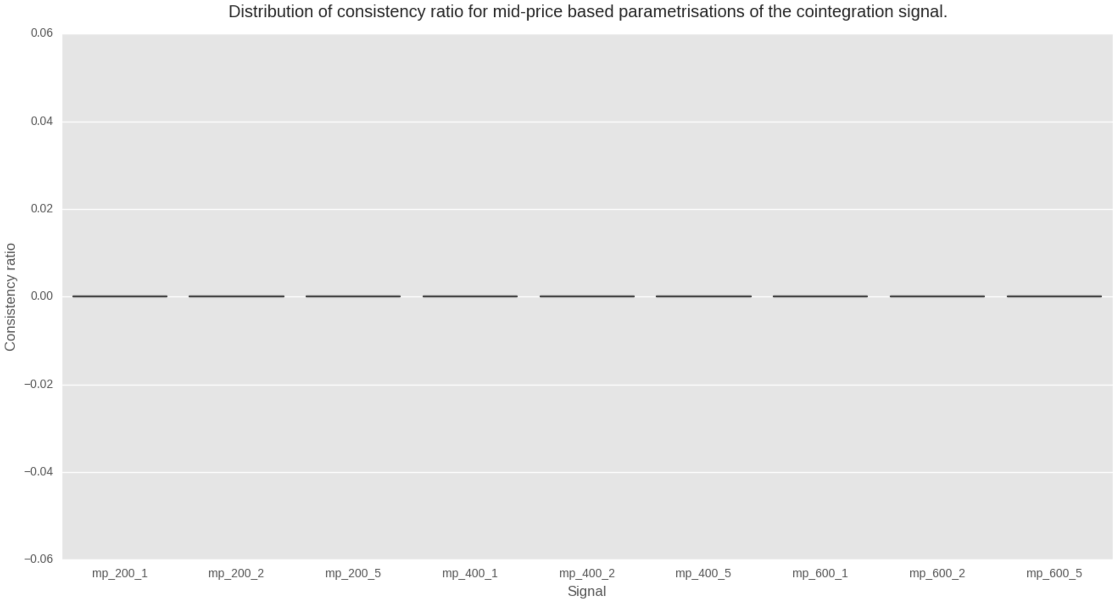

Distribution of consistency ratio for mid-price based parametrisations of the cointegration signal.
Consistency ratio
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
mp_200_1
mp_200_2
mp_200_5
mp_400_1
mp_400_2
mp_400_5
mp_600_1
mp_600_2
mp_600_5
Signal

Figure A.6