Detecting Mean Reverted Patterns in Statistical Arbitrage

Kostas Triantafyllopoulos
University of Sheffield

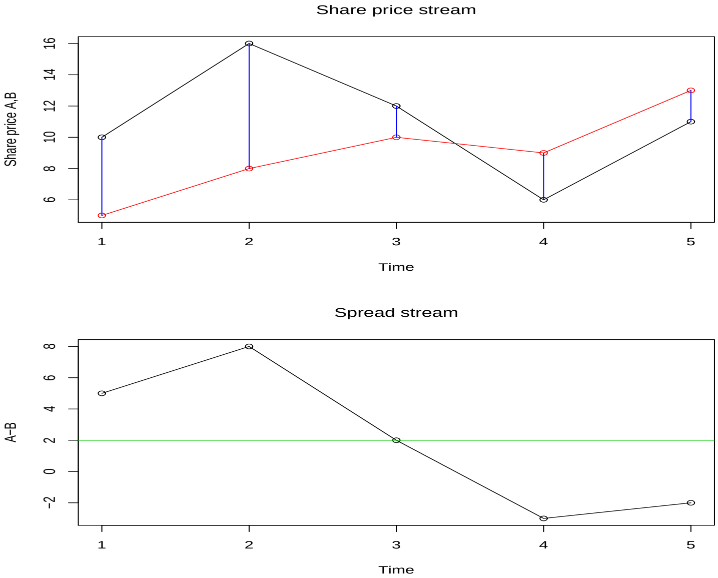
Outline

- Motivation / algorithmic pairs trading
  - Model set-up
  - Detection of local mean-reversion
- Adaptive estimation
  - 1. RLS with gradient variable forgetting factor
  - 2. RLS with Gauss-Newton variable forgetting factor
  - 3. RLS with beta-Bernoulli forgetting factor
- Trading strategy
- Pepsi and Coca Cola example

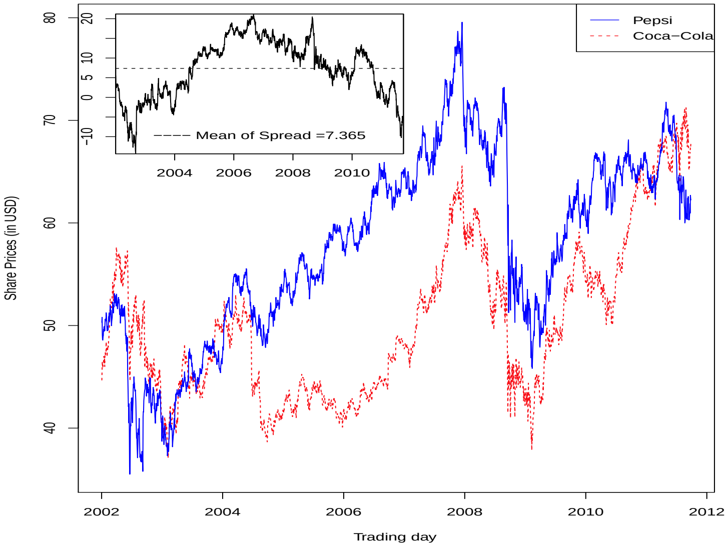
Introduction

- Statistical arbitrage.
- Algorithmic pairs trading market neutral trading.
  Buy low, sell high.
- Two assets with prices p_{At} and p_{Bt}
  - At t: if $p_{At} < p_{Bt}$, buy low (A) and sell high (B).
  - At $t+1$: if $p_{At} > p_{Bt}$, buy low (B) and sell high (A)...
- In the long run, mean reversion of spread $y_t = p_{At} - p_{Bt}$.
  If y_t goes up, y_t will go down at $t+1$.
  Take advantage of relative mispricings of A and B.

Introduction



Pepsi - Coca Cola date stream



Model set-up

(Elliott et al, 2005). y_t is a noisy version of a mean-reverted process.

$$y_t = x_t + \varepsilon_t$$
$$x_t = \alpha + \beta x_{t-1} + \zeta_t$$

(Triantafyllopoulos and Montana, 2011).

$$y_t = \alpha_t + \beta_t y_{t-1} + \epsilon_t = F_t^T \theta_t + \epsilon_t,$$
$$\theta_t = \Phi \theta_{t-1} + \omega_t,$$
$$F_t = \begin{bmatrix} 1 \\ y_{t-1} \end{bmatrix} \quad \text{and} \quad \theta_t = \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix}$$

Detecting mean reversion

Define $D_t = (y_1, \ldots, y_t)$ a data stream sample.

- (Elliott et al, 2005).
  If $|\beta| < 1$, then y_t is mean-reverted.
  With D_t, get online estimates $\hat{\alpha}$ and $\hat{\beta}$. If $|\hat{\beta}| < 1$, then mean-reversion.
- (Triantafyllopoulos and Montana, 2011).
  Under some assumptions y_t is mean-reverted if $|\beta_t| < 1$, for all t. We consider mean-reversion in segments or locally.
  Test again $|\hat{\beta}_t| < 1$.

So we need online estimates $\hat{\beta}_t$.

Recursive least squares (RLS)

(Haykin, 2001).
Recursive Least Squares (RLS), if $\Phi = I$.
Find θ that minimizes the cost function

$$\sum_{j=0}^{t-1} \lambda^j (y_{t-j} - F_{t-j}^T \theta)^2$$

$0 < \lambda < 1$ a forgetting factor. Past data are "forgotten" at a rate

$$\sum_{j=0}^{t-1} \lambda^j = 1 + \lambda + \cdots + \lambda^{t-1} \to \frac{1}{1-\lambda}$$

[memory of the system]

- With $\lambda = 1$, ordinary regression, memory is ∞
- With $\lambda < 1$, RLS, memory $(1-\lambda)^{-1}$.

SS-RLS (state-space RLS)

Variable forgetting factor (Haykin, 2001, Malik, 2006).

$$\lambda_t - \lambda_{t-1} = c(t)$$

Steepest descent:

$$\lambda_t = [\lambda_{t-1} + a\nabla_\lambda(t)]_{\lambda_-}^{\lambda_+},$$

1. $\nabla_\lambda(t) \approx -e_t F_t^T \Phi \psi_{t-1}$
2. $\psi_t = \partial m_t / \partial \lambda = (I - K_t F_t^T)\Phi\psi_{t-1} + S_t F_t e_t$
3.

$$\begin{aligned} S_t &= \partial P_t / \partial \lambda = -\lambda_t^{-1} P_t + \lambda_t^{-1} K_t K_t^T \\ &\quad + \lambda_t^{-1}(I - K_t F_t^T)\Phi S_{t-1} \Phi^T (I - F_t K_t^T). \end{aligned}$$

Need starting values $m_1, P_1, \psi_1, S_1, \lambda_1$.

GN-RLS (Gauss-Newton RLS)

Song et al (2000) gave an approximate GN algorithm.

$$\lambda_t = \left[\lambda_{t-1} + a \frac{\nabla_\lambda(t)}{\nabla^2_\lambda(t)} \right]_{\lambda_-}^{\lambda_+},$$

Here $\nabla^2_\lambda(t) \approx (F_t^T \Phi \psi_{t-1})^2 - e_t F_t^T \Phi \eta_t$

$$\eta_t = \partial\psi_{t-1}/\partial\lambda = (I - K_t F_t^T)\Phi\eta_{t-1} + L_t F_t e_t - 2 S_t F_t F_t^T \Phi \psi_{t-1}$$

where

$$\begin{aligned} L_t &= \lambda_t^{-1}(I - K_t F_t^T)\Phi L_{t-1}\Phi^T(I - F_t K_t^T) \\ &\quad + \lambda_t^{-2} P_t (I - F_t K_t^T) - \lambda_t^{-1} S_t + M_t + M_t^T \\ &\quad - \lambda_t^{-2}(I - K_t F_t^T)\Phi S_{t-1}\Phi^T(I - F_t K_t^T) \end{aligned}$$

and

$$M_t = \lambda_t^{-1} S_t F_t F_t^T \{P_t - \Phi S_{t-1}\Phi^T(I - F_t K_t^T)\}.$$

GN-RLS cont.

- GN-RLS creates too abrupt jumps in λ_t / too sensitive to changes.
- When smooth signal, we want SS-RLS and when noisy we want GN-RLS.
- We use

$$\lambda_t = \begin{cases} \left[\lambda_{t-1} + a\nabla_\lambda(t)\right]_{\lambda-}^{\lambda+}, & \text{if} \quad e_t^2 \leq k \\ \left[\lambda_{t-1} + a\frac{\nabla_\lambda(t)}{\nabla_\lambda^2(t)}\right]_{\lambda-}^{\lambda+}, & \text{if} \quad e_t^2 > k \end{cases}$$

BB-RLS (beta Bernoulli RLS)

In the above we ask $\lambda_- \le \lambda_t \le \lambda_+$.

- If e_t^2 small, $\lambda_t \to \lambda_+$ (smoothness)
- If e_t^2 large, $\lambda_t \to \lambda_-$ (adaptiveness)

We set

$$\lambda_t = \pi\lambda_+ + (1-\pi)\lambda_-, \quad \pi = Pr(e_t^2 \le k_t)$$

Two events small prediction error / large prediction error

$$x_t = \begin{cases} 1, & \text{if} \quad e_t^2 \le k_t, \quad \text{with probability } \pi \\ 0, & \text{if} \quad e_t^2 > k_t, \quad \text{with probability } 1-\pi \end{cases}$$

Observation model $x_t \sim \text{Bernoulli}(\pi)$.
Prior for π is beta, $\pi \sim B(c_1, c_2)$

$$p(x_t \mid \pi) = \pi^{x_t}(1 - x_t)^{1-x_t} \quad \text{(bernoulli model)}$$

$$p(\pi) \propto \pi^{c_1-1}(1 - \pi)^{c_2-1} \quad \text{(prior beta model)}$$

$$p(\pi \mid x_t) \propto p(x_t \mid \pi)p(\pi) \propto \pi^{c_1+x_t-1}(1 - \pi)^{c_2+1-x_t-1}$$

So $\pi|x_t \sim B(c_1 + x_t, c_2 - x_t + 1)$.
Sequentially: $\pi \mid x_1, \ldots, x_t \equiv \pi \mid D_{t-1} \sim Be(c_{1t}, c_{2t})$, where $c_{1t} = c_{1,t-1} + x_t$ and $c_{2t} = c_{2,t-1} - x_t + 1$.

$$\hat{\pi}_t = E(\pi \mid D_{t-1}) = c_{1t}(c_{1t} + c_{2t})^{-1}$$

$$\hat{\lambda}_t = E(\lambda_t|D_{t-1}) = \hat{\pi}_t\lambda_+ + (1 - \hat{\pi}_t)\lambda_-$$

From $Pr(e_t^2 \leq k_t) = \pi$, we have $k_t = q_t F_{\chi^2}^{-1}(\pi)$.
We use $k_t \approx q_t F_{\chi^2}^{-1}(\hat{\pi}_{t-1})$.

Key points of λ_t

- λ_t is stochastic.
- We can derive its distribution

$$p(\lambda_t|D_{t-1}) = c(\lambda_t - \lambda_-)^{c_{1t}-1}(\lambda_+ - \lambda_t)^{c_{2t}-1}$$

- We can evaluate the mode and the variance of λ_t.
- We can show

$$\hat{\lambda}_t \approx \bar{x}\lambda_+ + (1-\bar{x})\lambda_-$$

- If for many points $e_t^2 < k_t$, followed by few large $e_t^2 > k_t$, then locally $\hat{\lambda}_t$ does not work well.

Solution: intervention

Example.

1. Set $\lambda_- = 0.8$, $\lambda_+ = 0.99$
2. $e_t^2 \leq k_t$ $(t = 1, \ldots, 95)$
3. $e_t^2 > k_t$ $(t = 96, 97, 98, 99, 100)$
4. $\hat{\lambda}_t = 0.9805$ (closer to $\lambda_+ = 0.99$, than to $\lambda_- = 0.8$).

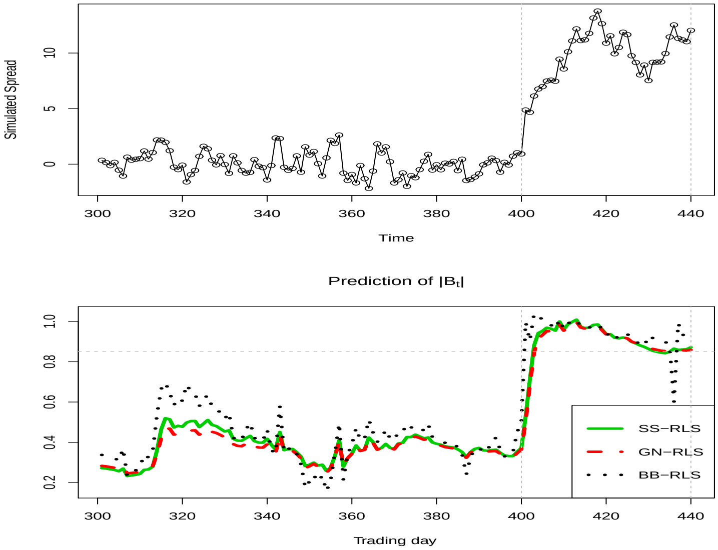
If change in x_t (from 0 to 1 or from 1 to 0), reset the priors $c_{1,t-1}$ and $c_{2,t-1}$ to the initial values $(c_{1,1} = c_{2,1} = 0.5)$.

$$\hat{\pi}_t = \frac{c_{1t}}{c_{1t} + c_{2t}} = \frac{c_{1,1} + x_t}{c_{1,1} + c_{2,1} + 1} = \frac{1 + 2x_t}{4}$$

$$\hat{\lambda}_t = \begin{cases} 0.75\lambda_+ + 0.25\lambda_-, & \text{if } x_t = 1 \\ 0.25\lambda_+ + 0.75\lambda-, & \text{if } x_t = 0 \end{cases}$$

In the example, $\hat{\lambda}_{96} = 0.25 \times 0.99 + 0.75 \times 0.8 = 0.847$.

Simulated streams



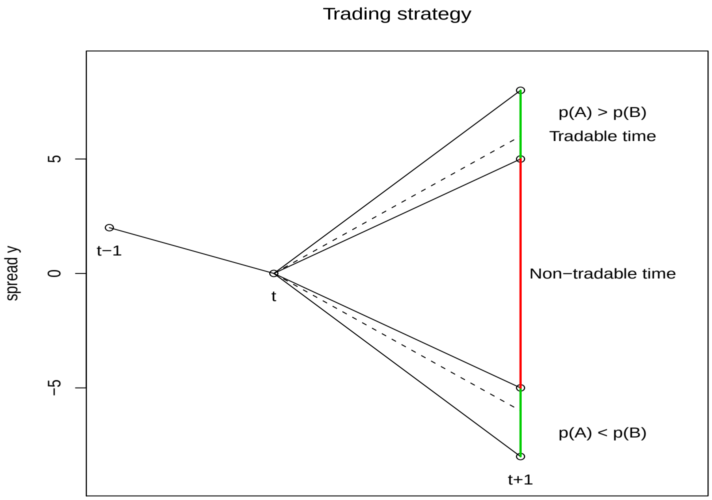
Trading strategy

Spread $y_t = p_{A,t} - p_{B,t}$.

- If y_t not mean-reverted do nothing. It is not predictable.
- If $y_t < y_{t+1}$, $p_{A,t+1} > p_{A,t}$ or $p_{B,t+1} < p_{B,t}$. Buy A / sell B.
- If $y_t > y_{t+1}$, $p_{A,t+1} < p_{A,t}$ or $p_{B,t+1} > p_{B,t}$. Buy B / sell A.
- If $y_t \approx y_{t-1}$, do nothing.

At time t, we don't know y_{t+1}, so we predict it and use $\hat{y}_{t+1}$.

Trading strategy



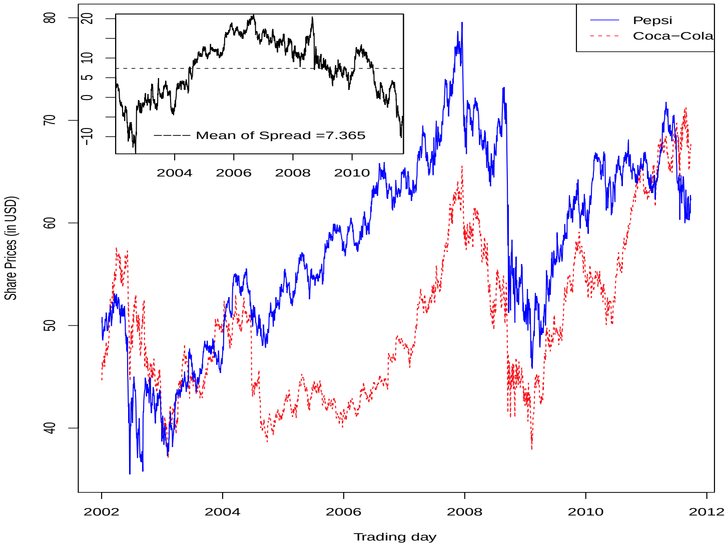
With observed spread y_t at t:

- Close the position of $t-1$ (if opened).
- If $|\hat{\beta}_{t+1}| < 0.99$, declare y_{t+1} as mean-reverted.
- If $\hat{y}_{t+1} - h > y_t$, buy A / short-sell B.
- If $\hat{y}_{t+1} + h < y_t$, buy B / short-sell A.

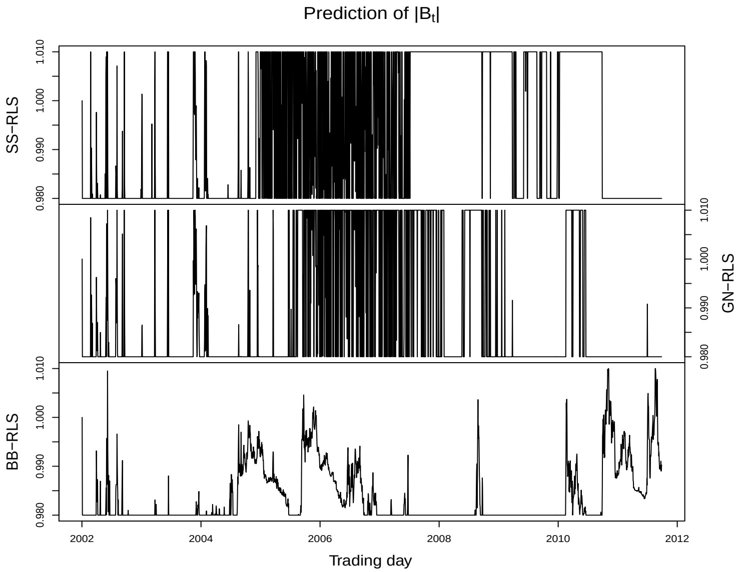
Example:

- At t: $y_t = 10$ and we project $\hat{y}_{t+1} = 11$.
- y_{t+1} can be 12 or 9 (rules change if we apply $y_t <> \hat{y}_{t+1}$.
- $y_{t+1} = 9$ can give loss, if we adopt $y_t < \hat{y}_{t+1}$ rule $(10 < 11)$.
- Take $h = 10\%$ of $\hat{y}_{t+1} = 1.1$, and $\hat{y}_{t+1} - h = 9.9 < 10 = y_t$, we do not open a position buy A and short sell B.
- As $\hat{y}_{t+1} + h = 11.1 > 10 = y_t$, we do not open a position short sell A and buy B.

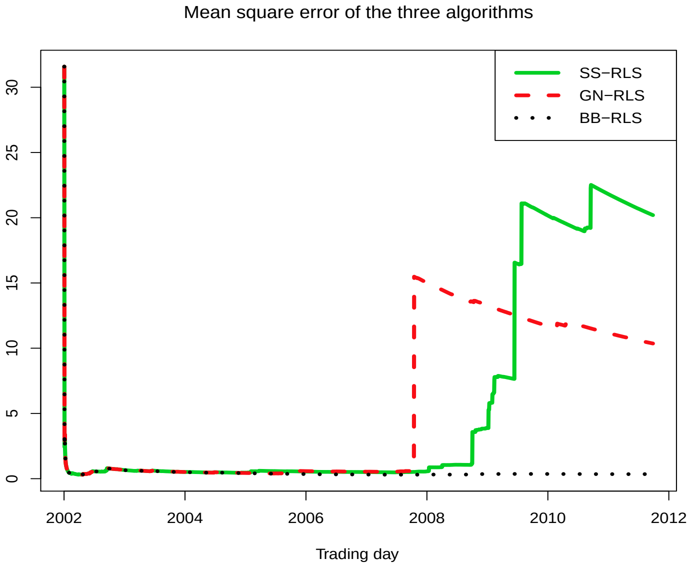
Pepsi - Coca Cola date stream



Detection of mean reverted patterns



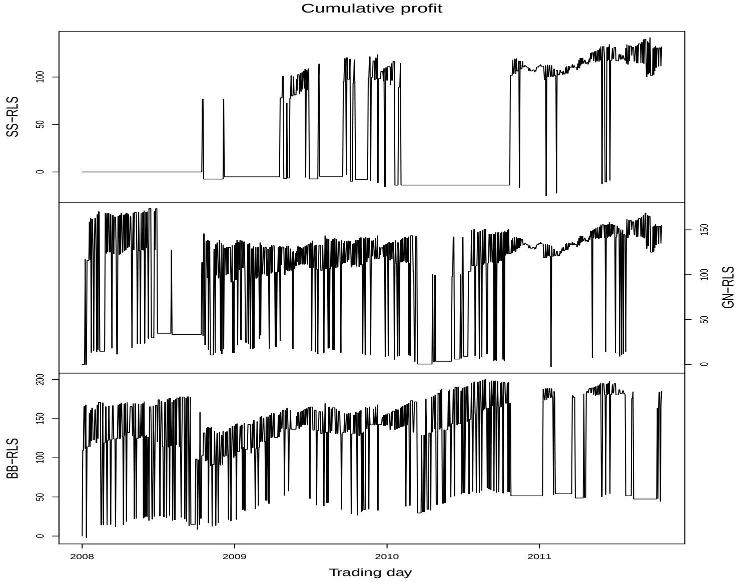
MSE over time



Trading performance

	h		
Mean	1%	3%	5%
SS-RLS	35.758	30.860	30.655
GN-RLS	101.885	92.819	80.091
BB-RLS	119.935	97.140	64.146
STD	1%	3%	5%
SS-RLS	57.210	55.638	55.357
GN-RLS	53.799	59.603	63.687
BB-RLS	54.890	60.223	58.654
FB	1%	3%	5%
SS-RLS	131.85	129.29	127.85
GN-RLS	155.19	159.44	172.29
BB-RLS	185.28	178.58	155.00

Trading performance



Closing remarks

- Algorithmic pairs trading / statistical arbitrage require online machine learning methods.
- Pattern recognition methods for mean reversion / segments of stationarity.
- We develop variable forgetting factors for online learning.
- Other methods include sequential Monte Carlo.
- Need to take into account the shape of the distribution of the data stream.
- Larger data streams / complex portfolios / many pairs to consider simultaneously.
- Trading strategy can be improved.

References

- Elliott, R., Van Der Hoek, J., and Malcolm, W. (2005). Pairs trading. Quantitative Finance, 5:271-276.
- Haykin, S. (2001). Adaptive Filter Theory. Prentice Hall, 4th edition.
- Malik, M. B. (2006). State-space recursive least-squares with adaptive memory. Signal Processing, 86:1365-1374.
- Song, S., Lim, J.-S., Baek, S., and Sung, K.-M. (2000). Gauss-Newton variable forgetting factor recursive least squares for time varying parameter tracking. Electronics Letters, 36:988-990.
- Triantafyllopoulos, K. and Montana, G. (2011). Dynamic modeling of mean-reverting spreads for statistical arbitrage. Computational Management Science, 8:23-49.