Introduction to Algorithmic Trading Strategies
Lecture 5

Pairs Trading by Stochastic Spread Methods

Haksun Li
haksun.li@numericalmethod.com
www.numericalmethod.com

Outline

- First passage time
- Kalman filter
- Maximum likelihood estimate
- EM algorithm

References

- As the emphasis of the basic co-integration methods of most papers are on the construction of a synthetic mean-reverting asset, the stochastic spread methods focuses on the dynamic of the price of the synthetic asset.
- Most referenced academic paper: Elliot, van der Hoek, and Malcolm, 2005, Pairs Trading
  - Model the spread process as a state-space version of Ornstein-Uhlenbeck process
- Jonathan Chiu, Daniel Wijaya Lukman, Kourosh Modarresi, Avinayan Senthi Velayutham. High-frequency Trading. Stanford University. 2011
- The idea has been conceived by a lot of popular pairs trading books
  - Technical analysis and charting for the spread, Ehrman, 2005, The Handbook of Pairs Trading
  - ARMA model, HMM ARMA model, some non-parametric approach, and a Kalman filter model, Vidyamurthy, 2004, Pairs Trading: Quantitative Methods and Analysis

Spread as a Mean-Reverting Process

- $x_k - x_{k-1} = (a - b x_{k-1})\tau + \sigma\sqrt{\tau}\varepsilon_k$
- $= b\left(\frac{a}{b} - x_{k-1}\right)\tau + \sigma\sqrt{\tau}\varepsilon_k$
- The long term mean = $\frac{a}{b}$.
- The rate of mean reversion = b.

Sum of Power Series

- We note that
  - $= \sum_{i=0}^{k-1} a^i = \frac{a^k - 1}{a - 1}$

Unconditional Mean

- $E(x_k) = \mu_k = \mu_{k-1} + (a - b\mu_{k-1})\tau$
- $= a\tau + (1 - b\tau)\mu_{k-1}$
- $= a\tau + (1 - b\tau)[a\tau + (1 - b\tau)\mu_{k-2}]$
- $= a\tau + (1 - b\tau)a\tau + (1 - b\tau)^2\mu_{k-2}$
- $= \sum_{i=0}^{k-1}(1 - b\tau)^i\, a\tau + (1 - b\tau)^k\mu_0$
- $= a\tau\,\frac{1-(1-b\tau)^k}{1-(1-b\tau)} + (1 - b\tau)^k\mu_0$
- $= a\tau\,\frac{1-(1-b\tau)^k}{b\tau} + (1 - b\tau)^k\mu_0$
- $= \frac{a}{b} - \frac{a}{b}(1 - b\tau)^k + (1 - b\tau)^k\mu_0$

Long Term Mean

- $\frac{a}{b} - \frac{a}{b}(1 - b\tau)^k + (1 - b\tau)^k \mu_0$
- $\to \frac{a}{b}$

Unconditional Variance

- $\mathrm{Var}(x_k) = {\sigma_k}^2 = (1 - b\tau)^2 {\sigma_{k-1}}^2 + \sigma^2\tau$
- $= (1 - b\tau)^2 {\sigma_{k-1}}^2 + \sigma^2\tau$
- $= (1 - b\tau)^2 [(1 - b\tau)^2 {\sigma_{k-2}}^2 + \sigma^2\tau] + \sigma^2\tau$
- $= \sigma^2\tau \sum_{i=0}^{k-1} (1 - b\tau)^{2i} + (1 - b\tau)^{2k} {\sigma_0}^2$
- $= \sigma^2\tau \frac{1-(1-b\tau)^{2k}}{1-(1-b\tau)^2} + (1 - b\tau)^{2k} {\sigma_0}^2$

Long Term Variance

- $\sigma^2 \tau \frac{1-(1-b\tau)^{2k}}{1-(1-b\tau)^2} + (1 - b\tau)^{2k}\, {\sigma_0}^2$
- $\to \frac{\sigma^2 \tau}{1-(1-b\tau)^2}$

Observations and Hidden State Process

- The hidden state process is:
  - $x_k = x_{k-1} + (a - bx_{k-1})\tau + \sigma\sqrt{\tau}\varepsilon_k$
  - $= a\tau + (1 - b\tau)x_{k-1} + \sigma\sqrt{\tau}\varepsilon_k$
  - $= A + Bx_{k-1} + C\varepsilon_k$
  - $A \geq 0, 0 < B < 1$
- The observations:
  - $y_k = x_k + D\omega_k$
- We want to compute the *expected* state from observations.
  - $\hat{x}_k = \hat{x}_{k|k} = E[x_k|Y_k]$

First Passage Time

- Standardized Ornstein-Uhlenbeck process
  - $dZ(t) = -Z(t)dt + \sqrt{2}dW(t)$
- First passage time
  - $T_{0,c} = \inf\{t \geq 0, Z(t) = 0 | Z(0) = c\}$
- The pdf of $T_{0,c}$ has a maximum value at
  - $\hat{t} = \frac{1}{2}\ln\left[1 + \frac{1}{2}\left(\sqrt{(c^2 - 3)^2 + 4c^2} + c^2 - 3\right)\right]$

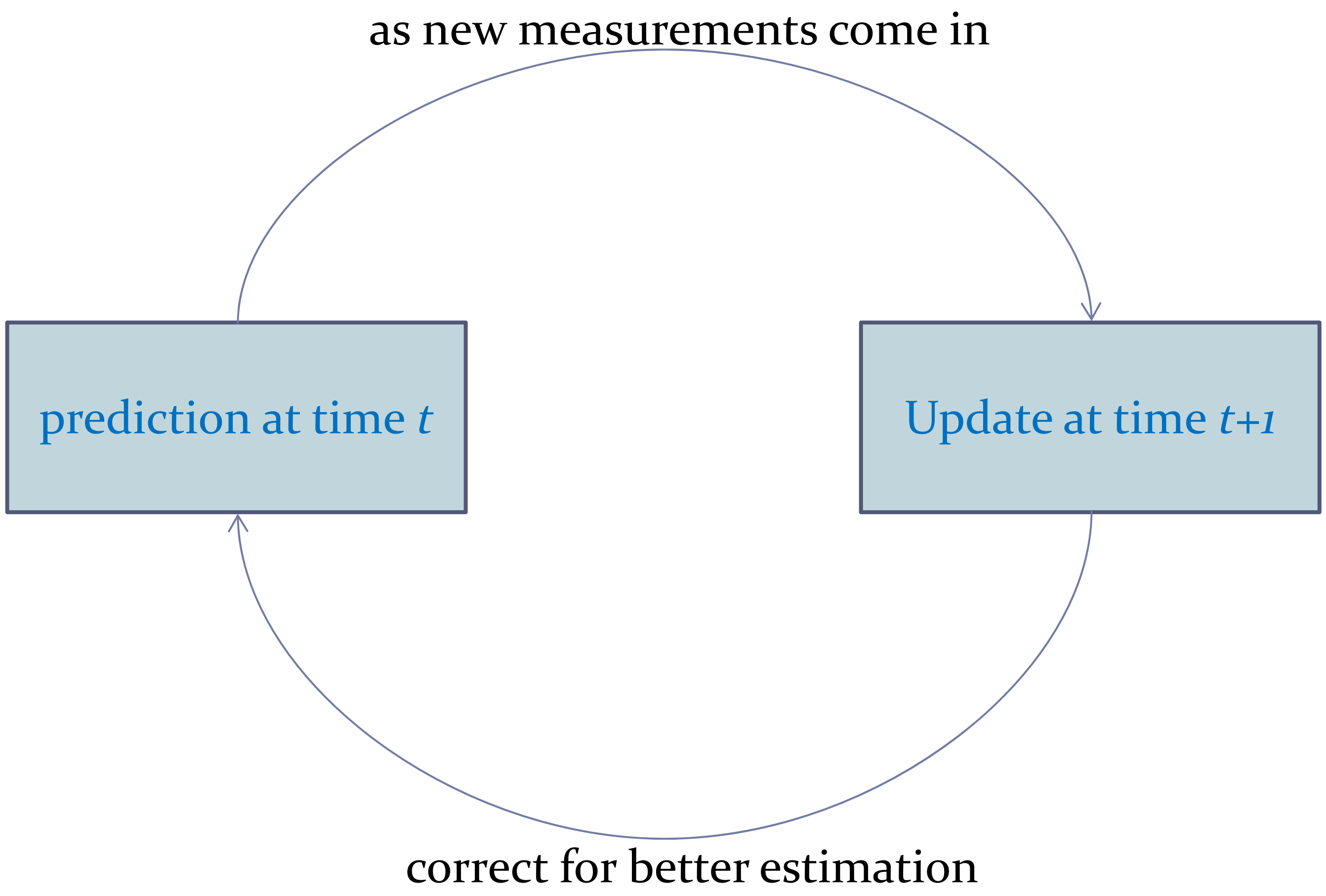
A Sample Trading Strategy

- $x_k = x_{k-1} + (a - b x_{k-1})\tau + \sigma\sqrt{\tau}\varepsilon_k$
- $dX(t) = \big(a - bX(t)\big)dt + \sigma dW(t)$
- $X(0) = \mu + c\frac{\sigma}{\sqrt{2\rho}}, X(T) = \mu$
- $T = \frac{1}{\rho}\hat{t}$
- Buy when $y_k < \mu - c\left(\frac{\sigma}{\sqrt{2\rho}}\right)$ unwind after time T
- Sell when $y_k > \mu + c\left(\frac{\sigma}{\sqrt{2\rho}}\right)$ unwind after time T

Kalman Filter

- The Kalman filter is an efficient recursive filter that estimates the state of a dynamic system from a series of incomplete and noisy measurements.

Conceptual Diagram



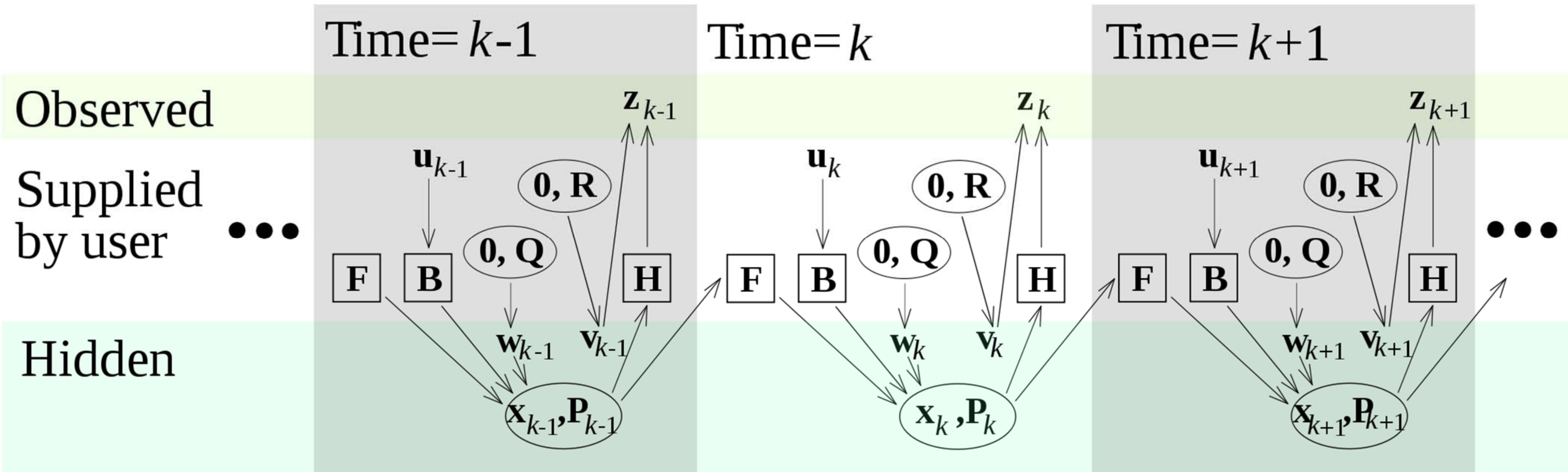
A Linear Discrete System

- $x_k = F_k x_{k-1} + B_k u_k + \omega_k$
- F_k: the state transition model applied to the previous state
- B_k: the control-input model applied to control vectors
- $\omega_k \sim N(0, Q_k)$: the noise process drawn from multivariate Normal distribution

Observations and Noises

- $z_k = H_k x_k + v_k$
- H_k: the observation model mapping the true states to observations
- $v_k \sim N(0, R_k)$: the observation noise

Discrete System Diagram



Prediction

- predicted a prior state estimate
  - $\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + B_k u_k$
- predicted a prior estimate covariance
  - $P_{k|k-1} = F_k P_{k-1|k-1} {F_k}^T + Q_k$

Update

- measurement residual
  - $\tilde{y}_k = z_k - H_k \hat{x}_{k|k-1}$
- residual covariance
  - $S_k = H_k P_{k|k-1} {H_k}^T + R_k$
- optimal Kalman gain
  - $K_k = P_{k|k-1} {H_k}^T {S_k}^{-1}$
- updated a posteriori state estimate
  - $\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \tilde{y}_k$
- updated a posteriori estimate covariance
  - $P_{k|k} = (I - K_k H_k) P_{k|k-1}$

Computing the 'Best' State Estimate

- Given A, B, C, D, we define the conditional variance
  - $R_k = \Sigma_{k|k} \equiv \mathrm{E}[(x_k - \hat{x}_k)^2 | Y_k]$
- Start with $\hat{x}_{0|0} = y_0$, $R_0 = D^2$.

Predicted (a Priori) State Estimation

- $\hat{x}_{k+1|k}$
- $= \mathrm{E}[x_{k+1}|Y_k]$
- $= \mathrm{E}[A + Bx_k + C\varepsilon_{k+1}|Y_k]$
- $= \mathrm{E}[A + Bx_k|Y_k]$
- $= A + B\,\mathrm{E}[x_k|Y_k]$
- $= A + B\hat{x}_{k|k}$

Predicted (a Priori) Variance

- $\Sigma_{k+1|k}$
- $= \mathrm{E}[(x_{k+1} - \hat{x}_{k+1})^2 | Y_k]$
- $= \mathrm{E}[(A + Bx_k + C\varepsilon_{k+1} - \hat{x}_{k+1})^2 | Y_k]$
- $= \mathrm{E}\left[\left(A + Bx_k + C\varepsilon_{k+1} - A - B\hat{x}_{k|k}\right)^2 | Y_k\right]$
- $= \mathrm{E}\left[\left(Bx_k - B\hat{x}_{k|k} + C\varepsilon_{k+1}\right)^2 | Y_k\right]$
- $= \mathrm{E}\left[\left(Bx_k - B\hat{x}_{k|k}\right)^2 + C^2\varepsilon^2{}_{k+1} | Y_k\right]$
- $= B^2\Sigma_{k|k} + C^2$

Minimize Posteriori Variance

- Let the Kalman updating formula be
  - $\hat{x}_{k+1} = \hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K\left[y_{k+1} - \hat{x}_{k+1|k}\right]$
- We want to solve for K such that the conditional variance is minimized.
  - $\Sigma_{k+1|k} = \mathrm{E}[(x_{k+1} - \hat{x}_{k+1})^2|Y_k]$

Solve for K

- $\mathrm{E}[(x_{k+1} - \hat{x}_{k+1})^2 | Y_k]$
- $= \mathrm{E}\left[\left(x_{k+1} - \hat{x}_{k+1|k} - K\left[y_{k+1} - \hat{x}_{k+1|k}\right]\right)^2 | Y_k\right]$
- $= \mathrm{E}\left[\left(x_{k+1} - \hat{x}_{k+1|k} - K\left[x_{k+1} - \hat{x}_{k+1|k} + D\omega_{k+1}\right]\right)^2 | Y_k\right]$
- $= \mathrm{E}\left[\left[(1 - K)\left(x_{k+1} - \hat{x}_{k+1|k}\right) - KD\omega_{k+1}\right]^2 | Y_k\right]$
- $= (1 - K)^2\, \mathrm{E}\left[\left(x_{k+1} - \hat{x}_{k+1|k}\right)^2 | Y_k\right] + K^2 D^2$
- $= (1 - K)^2\, \Sigma_{k+1|k} + K^2 D^2$

First Order Condition for k

- $\frac{d}{dK}(1-K)^2\,\Sigma_{k+1|k} + K^2D^2$
- $= \frac{d}{dK}(1-2K+K^2)\,\Sigma_{k+1|k} + K^2D^2$
- $= (-2+2K)\,\Sigma_{k+1|k} + 2KD^2$
- $= 0$

Optimal Kalman Filter

- $K_{k+1} = \frac{\Sigma_{k+1|k}}{\Sigma_{k+1|k}+D^2}$

Updated (a Posteriori) State Estimation

- So, we have the "optimal" Kalman updating rule.
  - $\hat{x}_{k+1} = \hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K\left[y_{k+1} - \hat{x}_{k+1|k}\right]$
  - $= \hat{x}_{k+1|k} + \frac{\Sigma_{k+1|k}}{\Sigma_{k+1|k}+D^2}\left[y_{k+1} - \hat{x}_{k+1|k}\right]$

Updated (a Posteriori) Variance

- $R_{k+1} = \Sigma_{k+1|k} = \mathrm{E}[(x_{k+1} - \hat{x}_{k+1})^2 | Y_{k+1}] = (1 - K)^2\, \Sigma_{k+1|k} + K^2 D^2$
- $= \left(1 - \frac{\Sigma_{k+1|k}}{\Sigma_{k+1|k}+D^2}\right)^2 \Sigma_{k+1|k} + \left(\frac{\Sigma_{k+1|k}}{\Sigma_{k+1|k}+D^2}\right)^2 D^2$
- $= \left(\frac{D^2}{\Sigma_{k+1|k}+D^2}\right)^2 \Sigma_{k+1|k} + \left(\frac{\Sigma_{k+1|k}}{\Sigma_{k+1|k}+D^2}\right)^2 D^2$
- $= \frac{D^4 \Sigma_{k+1|k} + D^2 {\Sigma_{k+1|k}}^2}{\left(\Sigma_{k+1|k}+D^2\right)^2}$
- $= \frac{D^4 \Sigma_{k+1|k} + D^2 {\Sigma_{k+1|k}}^2}{\left(\Sigma_{k+1|k}+D^2\right)^2}$
- $= \frac{\left(\Sigma_{k+1|k} D^2\right)\left(D^2 + \Sigma_{k+1|k} D^2\right)}{\left(\Sigma_{k+1|k}+D^2\right)^2}$
- $= \Sigma_{k+1|k} D^2$

Parameter Estimation

- We need to estimate the parameters $\vartheta = \{A, B, C, D\}$ from the observable data before we can use the Kalman filter model.
- We need to write down the likelihood function in terms of ϑ, and then maximize w.r.t. ϑ.

Likelihood Function

- A likelihood function (often simply the likelihood) is a function of the parameters of a statistical model, defined as follows: the likelihood of a set of parameter values given some observed outcomes is equal to the probability of those observed outcomes given those parameter values.
- $L(\vartheta; Y) = p(Y|\vartheta)$

Maximum Likelihood Estimate

- We find ϑ such that $L(\vartheta; Y)$ is maximized given the observation.

Example Using the Normal Distribution

- We want to estimate the mean of a sample of size N drawn from a Normal distribution.
  - $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$
- $\vartheta = \{\mu, \sigma\}$
- $L_N(\vartheta; Y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\}$

Log-Likelihood

- $\log L_N(\vartheta; Y) = \sum_{i=1}^{N} \left\{ \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i - \mu)^2}{2\sigma^2} \right\}$
- Maximizing the log-likelihood is equivalent to maximizing the following.
  - $-\sum_{i=1}^{N} \{(x_i - \mu)^2\}$
- First order condition w.r.t.,μ
  - $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

Nelder-Mead

- After we write down the likelihood function for the Kalman model in terms of $\vartheta = \{A, B, C, D\}$, we can run any multivariate optimization algorithm, e.g., Nelder-Mead, to search for ϑ.
  - $\max_{\vartheta} L(\vartheta; Y)$
- The disadvantage is that it may not converge well, hence not landing close to the optimal solution.

Marginal Likelihood

- For the set of hidden states, $\{X_t\}$, we write
  - $L(\vartheta; Y) = p(Y|\vartheta) = \sum_X p(Y, X|\vartheta)$
- Assume we know the conditional distribution of X, we could instead maximize the following.
  - $\max_{\vartheta} \mathop{\mathrm{E}}_{X}[L(\vartheta|Y, X)]$, or
  - $\max_{\vartheta} \mathop{\mathrm{E}}_{X}[\log L(\vartheta|Y, X)]$
- The expectation is a weighted sum of the (log-) likelihoods weighted by the probability of the hidden states.

The Q-Function

- Where do we get the conditional distribution of $\{X_t\}$ from?
- Suppose we somehow have an (initial) estimation of the parameters, ϑ_0. Then the model has no unknowns. We can compute the distribution of $\{X_t\}$.
- $Q\left(\vartheta|\vartheta^{(t)}\right) = \underset{X|Y,\vartheta}{\mathrm{E}}[\log L(\vartheta|Y,X)]$

EM Intuition

- Suppose we know ϑ, we know completely about the mode; we can find X.
- Suppose we know X, we can estimate ϑ, by, e.g., maximum likelihood.
- What do we do if we don't know both ϑ and X?

Expectation-Maximization Algorithm

- Expectation step (E-step): compute the expected value of the log-likelihood function, w.r.t., the conditional distribution of X under Y and ϑ.
  - $Q\left(\vartheta|\vartheta^{(t)}\right) = \underset{X|Y,\vartheta}{\mathrm{E}}[\log L(\vartheta|Y,X)]$
- Maximization step (M-step): find the parameters, ϑ, that maximize the Q-value.
  - $\vartheta^{(t+1)} = \underset{\vartheta}{\operatorname{argmax}}\, Q\left(\vartheta|\vartheta^{(t)}\right)$

EM Algorithms for Kalman Filter

- Offline: Shumway and Stoffer smoother approach, 1982
- Online: Elliott and Krishnamurthy filter approach, 1999

A Trading Algorithm

- From $y_0, y_1, \ldots, y_N$, we estimate $\hat{\vartheta}(N)$.
- Decide whether to make a trade at $t = N$, unwind at $t = N + 1$, or some time later, e.g., $t = N + T$.
- As y_{N+1} arrives, estimate $\hat{\vartheta}(N + 1)$.
- Repeat.

Results (1)

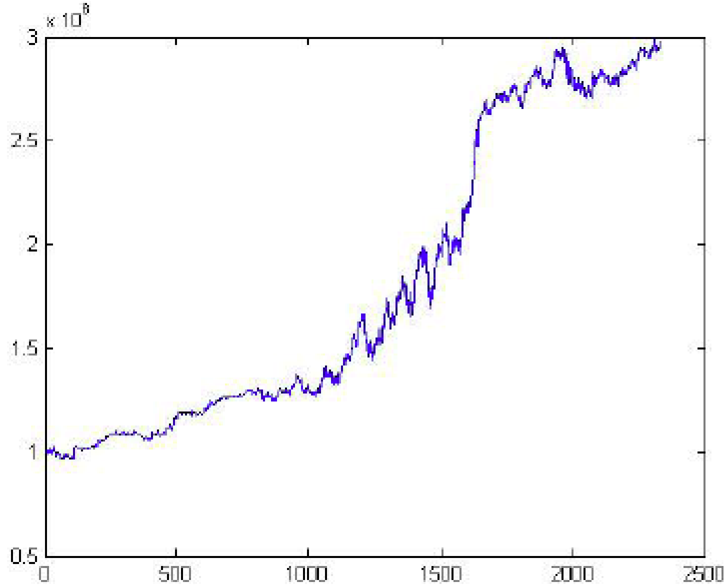


Figure 4: Backtesting Result with Optimization for weights on each pair

Results (2)

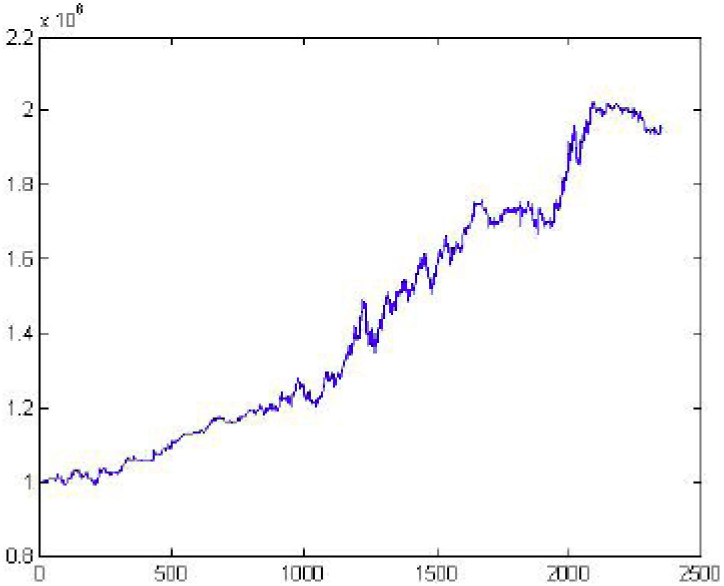


Figure 5: Backtesting Result Using Equal Weight Portfolio

Results (3)

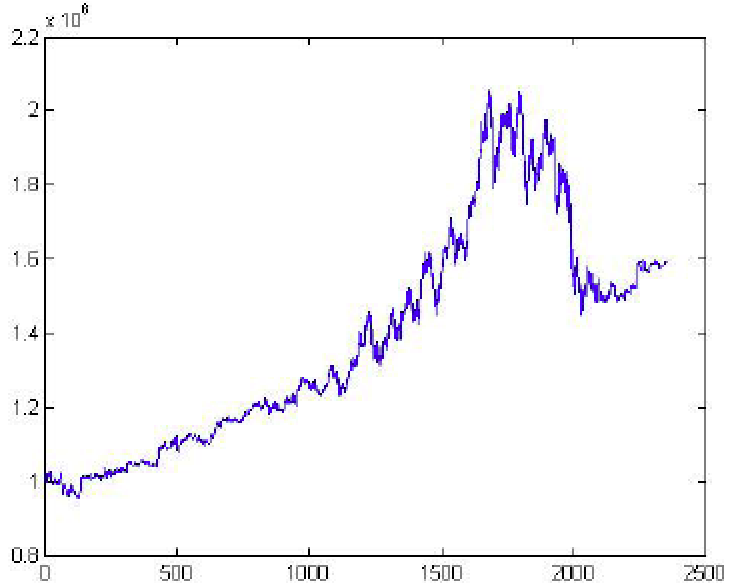


Figure 6: Backtesting Results Using 30-days Rolling Window Size