

Trading Basket Construction

Mean Reversion Trading

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

Speaker Profile



- Dr. Haksun Li
- CEO, Numerical Method Inc.
- (Ex-)Adjunct Professors, Industry Fellow, Advisor, Consultant with the National University of Singapore, Nanyang Technological University, Fudan University, the Hong Kong University of Science and Technology.
- Quantitative Trader/Analyst, BNPP, UBS
- PhD, Computer Sci, University of Michigan Ann Arbor
- M.S., Financial Mathematics, University of Chicago
- B.S., Mathematics, University of Chicago

References

- Pairs Trading: A Cointegration Approach. Arlen David Schmidt. University of Sydney. Finance Honours Thesis. November 2008.
- Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Soren Johansen. Oxford University Press, USA. February 1, 1996.
- Pairs Trading. Elliot, van der Hoek, and Malcolm. 2005.
- Identifying Small Mean Reverting Portfolios. A. d'Aspremont. 2008.
- High-frequency Trading. Stanford University. Jonathan Chiu, Daniel Wijaya Lukman, Kourosh Modarresi, Avinayan Senthi Velayutham. 2011.

Paris Trading

Pairs Trading

- Intuition: The thousands of market instruments are not independent. For two closely related assets, they tend to "move together" (common trend). We want to buy the cheap one and sell the expensive one.
  - Exploit short term deviation from long term equilibrium.
- Definition: trade one asset (or basket) against another asset (or basket)
  - Long one and short the other
- Try to make money from "spread".

FOMC announcements

- Rate cut!
- Bonds react.
- FX react.
- Stocks react.
- All act!

GLD vs. SLV


Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All
Nov 26, 2004 - Oct 07, 2013
GLD +185.31% SLV +56.02%
300%
250%
200%
150%
100%
50%
0%
-50%
2007
2008
2009
2010
2011
2012
2013


Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All
Dec 31, 2012 - Oct 07, 2013
GLD -20.42% SLV -25.94%
5%
0%
-5%
-10%
-15%
-20%
-25%
-30%
-35%
-40%
Jan 2013
Feb 2013
Mar 2013
Apr 2013
May 2013
Jun 2013
Jul 2013
Aug 2013
Sep 2013


Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All
Jul 08, 2013 - Oct 07, 2013
GLD +8.18% SLV +18.33%
25%
20%
15%
10%
5%
0%
Jul 12
Jul 19
Jul 26
Aug 2
Aug 9
Aug 16
Aug 23
Aug 30
Sep 6
Sep 13
Sep 20
Sep 26
Oct 4

Hows

- How to construct a pair?
- How to trade a pair?

Sample Pairs Trading Strategy

Spread

- $Z = X - \beta Y$
- β
  - hedge ratio
  - Cointegration coefficient
- How do you trade spread?
  - How much X to buy/sell?
  - How much Y to buy/sell?

Log-Spread

- $Z = \log X - \beta \log Y$
- How do you trade log-spread?
  - How much X to buy/sell?
  - How much Y to buy/sell?

Dollar Neutral Hedge

- Suppose ES (S&P500 E-mini future) is at 1220 and each point worth $50, its dollar value is about $61,000. Suppose NQ (Nasdaq 100 E-mini future) is at 1634 and each point worth $20, its dollar value is $32,680.
- $\beta = \frac{61000}{32680} = 1.87.$
- $Z = ES - 1.87 \times NQ$
- Buy Z = Buy 10 ES contracts and Sell 19 NQ contracts.
- Sell Z = Sell 10 ES contracts and Buy 19 NQ contracts.

Market Neutral Hedge

- Suppose ES has a market beta of 1.25, NQ 1.11.
- We use $\beta = \frac{1.25}{1.11} = 1.13$

Dynamic Hedge

- β changes with time, covariance, market conditions, etc.
- Periodic recalibration.

Distance Method

- The distance between two time series:
  - $d = \sum(x_i - y_j)^2$
- x_i, y_j are the normalized prices.
- We choose a pair of stocks among a collection with the smallest distance, d.

Distance Trading Strategy

- Sell Z if Z is too expensive.
- Buy Z if Z is too cheap.
- How do we do the evaluation?

Z Transform

- We normalize Z.
- The normalized value is called z-score.
- $z = \frac{x - \bar{x}}{\sigma_x}$
- Other forms:
- $z = \frac{x - M \times \bar{x}}{S \times \sigma_x}$
  - M, S are proprietary functions for forecasting.

A Very Simple Distance Pairs Trading

- Sell Z when z > 2 (standard deviations).
  - Sell 10 ES contracts and Buy 19 NQ contracts.
- Buy Z when z < -2 (standard deviations).
  - Buy 10 ES contracts and Sell 19 NQ contracts.

Pros of the Distance Model

- Model free.
- No mis-specification.
- No mis-estimation.
- Distance measure intuitively captures the Law of One Price (LOP) idea.

Cons of the Distance Model

- There is no reason why the model will work (or not). There is no assumption to check against the current market conditions.
- The model is difficult to analyze mathematically.
  - Cannot predict the convergence time (expected holding time).
- The model ignores the dynamic nature of the spread process, essentially treating the spread as i.i.d.
- Using more strict criterions may work for equities. In FX trading, we don't have the luxury of throwing away many pairs.

Risks in Pairs Trading

- Long term equilibrium does not hold.
  - E.g., the company that you long goes bankrupt but the other leg does not move (one company wins over the other).
- Systematic market risk.
- Firm specific risk.
- Liquidity.

Cointegration

Stationarity

- These ad-hoc β calibration does not guarantee the single most important statistical property in trading: stationarity.
- Strong stationarity: the joint probability distribution of $\{x_t\}$ does not change over time.
- Weak stationarity: the first and second moments do not change over time.
  - Covariance stationarity

Mean Reversion

- A stationary stochastic process is mean-reverting.
- Long when the spread/portfolio/basket falls sufficiently below a long term equilibrium.
- Short when the spread/portfolio/basket rises sufficiently above a long term equilibrium.

Test for Stationarity

- An augmented Dickey–Fuller test (ADF) is a test for a unit root in a time series sample.
- It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models.
- Intuition:
  - if the series y_t is stationary, then it has a tendency to return to a constant mean. Therefore large values will tend to be followed by smaller values, and small values by larger values. Accordingly, the level of the series will be a significant predictor of next period's change, and will have a negative coefficient.
  - If, on the other hand, the series is integrated, then positive changes and negative changes will occur with probabilities that do not depend on the current level of the series.
  - In a random walk, where you are now does not affect which way you will go next.

ADF Math

- $\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p-1} \Delta y_{t-i} + \epsilon_t$
- Null hypothesis $H_0: \gamma = 0$. (y_t non-stationary)
- $\alpha = 0, \beta = 0$ models a random walk.
- $\beta = 0$ models a random walk with drift.
- Test statistics = $\frac{\hat{\gamma}}{\sigma(\hat{\gamma})}$, the more negative, the more reason to reject H_0 (hence y_t stationary).
- SuanShu: AugmentedDickeyFuller.java

Cointegration

- Cointegration: select a linear combination of assets to construct an (approximately) stationary portfolio.

Objective

- Given two I(1) price series, we want to find a linear combination such that:
  - $z_t = x_t - \beta y_t = \mu + \varepsilon_t$
- ε_t is I(0), a stationary residue.
- μ is the long term equilibrium.
- Long when $z_t < \mu - \Delta$.
- Sell when $z_t > \mu + \Delta$.

Stocks from the Same Industry

- Reduce market risk, esp., in bear market.
  - Stocks from the same industry are likely to be subject to the same systematic risk.
- Give some theoretical unpinning to the pairs trading.
  - Stocks from the same industry are likely to be driven by the same fundamental factors (common trends).

Cointegration Definition

- $X_t \sim \mathrm{CI}(d, b)$ if
  - All components of X_t are integrated of same order d.
  - There exists a β_t such that the linear combination, $\beta_t X_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_n X_{nt}$, is integrated of order $(d - b), b > 0$.
- β is the cointegrating vector, not unique.

Illustration for Trading

- Suppose we have two assets, both reasonably I(1), we want to find β such that
  - $Z = X + \beta Y$ is I(0), i.e., stationary.
- In this case, we have $d = 1, b = 1$.

A Simple VAR Example

- $y_t = a_{11}y_{t-1} + a_{12}z_{t-1} + \varepsilon_{yt}$
- $z_t = a_{21}y_{t-1} + a_{22}z_{t-1} + \varepsilon_{zt}$
- Theorem 4.2, Johansen, places certain restrictions on the coefficients for the VAR to be cointegrated.
  - The roots of the characteristics equation lie on or outside the unit disc.

Coefficient Restrictions

- $a_{11} = \frac{(1-a_{22})-a_{12}a_{21}}{1-a_{22}}$
- $a_{22} > -1$
- $a_{12}a_{21} + a_{22} < 1$

VECM (1)

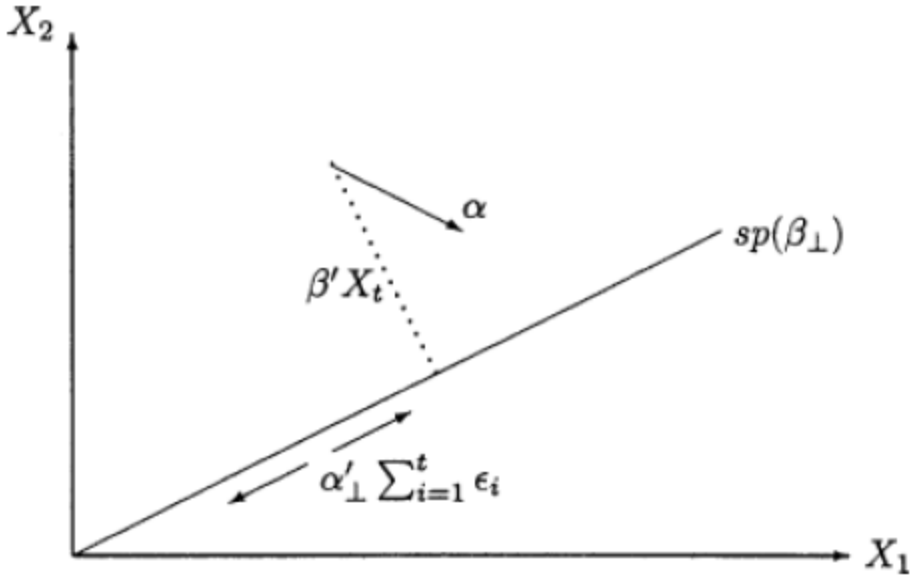
- Taking differences
  - $y_t - y_{t-1} = (a_{11} - 1)y_{t-1} + a_{12}z_{t-1} + \varepsilon_{yt}$
  - $z_t - z_{t-1} = a_{21}y_{t-1} + (a_{22} - 1)z_{t-1} + \varepsilon_{zt}$
- $\begin{bmatrix}\Delta y_t \\ \Delta z_t\end{bmatrix} = \begin{bmatrix}a_{11} - 1 & a_{12} \\ a_{21} & a_{22} - 1\end{bmatrix}\begin{bmatrix}y_{t-1} \\ z_{t-1}\end{bmatrix} + \begin{bmatrix}\varepsilon_{yt} \\ \varepsilon_{zt}\end{bmatrix}$
- Substitution of a_{11}
  - $\begin{bmatrix}\Delta y_t \\ \Delta z_t\end{bmatrix} = \begin{bmatrix}\frac{-a_{12}a_{21}}{1-a_{22}} & a_{12} \\ a_{21} & a_{22} - 1\end{bmatrix}\begin{bmatrix}y_{t-1} \\ z_{t-1}\end{bmatrix} + \begin{bmatrix}\varepsilon_{yt} \\ \varepsilon_{zt}\end{bmatrix}$

VECM (2)

- $\Delta y_t = \alpha_y(y_{t-1} - \beta z_{t-1}) + \epsilon_{yt}$
- $\Delta z_t = \alpha_z(y_{t-1} - \beta z_{t-1}) + \epsilon_{zt}$
- $\alpha_y = \frac{-a_{12}a_{21}}{1-a_{22}}$
- $\alpha_z = a_{21}$
- $\beta = \frac{1-a_{22}}{a_{21}}$, the cointegrating coefficient
- $y_{t-1} - \beta z_{t-1}$ is the long run equilibrium, I(0).
- α_y, α_z are the speed of adjustment parameters.

Interpretation

- Suppose the long run equilibrium is 0,
  - Δy_t, Δz_t responds only to shocks.
- Suppose $\alpha_y < 0$, $\alpha_z > 0$,
  - $\{y_t\}$ decreases in response to a +ve deviation.
  - $\{z_t\}$ increases in response to a +ve deviation.



Granger Representation Theorem

- If X_t is cointegrated, an VECM form exists.
- The increments can be expressed as a functions of the dis-equilibrium, and the lagged increments.
- $\Delta X_t = \alpha\beta' X_{t-1} + \sum c_t \Delta X_{t-1} + \varepsilon_t$
- In our simple example, we have
  - $\begin{bmatrix}\Delta y_t \\ \Delta z_t\end{bmatrix} = \begin{bmatrix}\alpha_y \\ \alpha_z\end{bmatrix} [1 \quad -\beta] \begin{bmatrix}y_{t-1} \\ z_{t-1}\end{bmatrix} + \begin{bmatrix}\varepsilon_{yt} \\ \varepsilon_{zt}\end{bmatrix}$

Granger Causality

- $\{z_t\}$ does not Granger Cause $\{y_t\}$ if lagged values of $\{\Delta z_{t-i}\}$ do not enter the Δy_t equation.
- $\{y_t\}$ does not Granger Cause $\{z_t\}$ if lagged values of $\{\Delta y_{t-i}\}$ do not enter the Δz_t equation.

Engle-Granger Two Step Approach

- Estimate either
  - $y_t = \beta_{10} + \beta_{11} z_t + e_{1t}$
  - $z_t = \beta_{20} + \beta_{21} y_t + e_{2t}$
  - As the sample size increase indefinitely, asymptotically a test for a unit root in $\{e_{1t}\}$ and $\{e_{2t}\}$ are equivalent, but not for small sample sizes.
- Test for unit root using ADF on either $\{e_{1t}\}$ and $\{e_{2t}\}$.
- If $\{y_t\}$ and $\{z_t\}$ are cointegrated, $\{\beta\}$ super converges.

Engle-Granger Pros and Cons

- Pros:
  - simple
- Cons:
  - This approach is subject to twice the estimation errors. Any errors introduced in the first step carry over to the second step.
  - Work only for two I(1) time series.

Testing for Cointegration

- Note that in the VECM, the rows in the coefficient, Π, are **NOT** linearly independent.
  - $\begin{bmatrix}\Delta y_t \\ \Delta z_t\end{bmatrix} = \begin{bmatrix}\frac{-a_{12}a_{21}}{1-a_{22}} & a_{12} \\ a_{21} & a_{22}-1\end{bmatrix}\begin{bmatrix}y_{t-1} \\ z_{t-1}\end{bmatrix} + \begin{bmatrix}\varepsilon_{yt} \\ \varepsilon_{zt}\end{bmatrix}$
- $\begin{bmatrix}\frac{-a_{12}a_{21}}{1-a_{22}} & a_{12}\end{bmatrix} \times \frac{-(1-a_{22})}{a_{12}} = [a_{21} \quad a_{22}-1]$
- The rank of Π determine whether the two assets $\{y_t\}$ and $\{z_t\}$ are cointegrated.

VAR & VECM

- In general, we can write convert a VAR to an VECM.
- VAR (from numerical estimation by, e.g., OLS):
  - $X_t = \sum_{i=1}^{p} A_i X_{t-i} + \varepsilon_t$
- Transitory form of VECM (reduced form)
  - $\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + \varepsilon_t$
- Long run form of VECM
  - $\Delta X_t = \sum_{i=1}^{p-1} \Upsilon_i \Delta X_{t-i} + \Pi X_{t-p} + \varepsilon_t$

The Π Matrix

- Rank(Π) = n, full rank
  - The system is already stationary; a standard VAR model in levels.
- Rank(Π) = 0
  - There exists NO cointegrating relations among the time series.
- 0 < Rank(Π) < n
  - $\Pi = \alpha\beta'$
  - β is the cointegrating vector
  - α is the speed of adjustment.

Rank Determination

- Determining the rank of Π is amount to determining the number of non-zero eigenvalues of Π.
  - Π is usually obtained from (numerical VAR) estimation.
  - Eigenvalues are computed using a numerical procedure.

Trace Statistics

- Suppose the eigenvalues of Π are:$\lambda_1 > \lambda_2 > \cdots > \lambda_n$.
- For the o eigenvalues, $\ln(1 - \lambda_i) = 0$.
- For the (big) non-zero eigenvalues, $\ln(1 - \lambda_i)$ is (very negative).
- The likelihood ratio test statistics
  - $Q\big(H(r)|H(n)\big) = -T\sum_{i=r+1}^{p}\log(1 - \lambda_i)$
  - Ho: rank ≤ r; there are at most r cointegrating β.

Test Procedure

- int r = 0;//rank
- for (; r <= n; ++r) { // loop until the null is accepted
  - compute Q = $Q\big(H(r)|H(n)\big)$;
  - If (Q > c.v.) { // compare against a critical value
    - break; // fail to reject the null hypothesis; rank found
  - }
- }
- r is the rank found

Decomposing Π

- Suppose the rank of $\Pi = r$.
- $\Pi = \alpha\beta'$.
- Π is $n \times n$.
- α is $n \times r$.
- $\beta'$is r $\times n$.

Estimating β

- β can estimated by maximizing the log-likelihood function in Chapter 6, Johansen.
  - $\log L(\Psi, \alpha, \beta, \Omega)$
- Theorem 6.1, Johansen: β is found by solving the following eigenvalue problem:
  - $\left|\lambda S_{11} - S_{10} {S_{00}}^{-1} S_{01}\right| = 0$

β

- Each non-zero eigenvalue λ corresponds to a cointegrating vector, which is its eigenvector.
- $\beta = (v_1, v_2, \cdots, v_r)$
- β spans the cointegrating space.
- For two cointegrating asset, there are only one β (v_1) so it is unequivocal.
- When there are multiple β, we need to add economic restrictions to identify β.

Trading the Pairs

- Given a space of (liquid) assets, we compute the pairwise cointegrating relationships.
- For each pair, we validate stationarity by performing the ADF test.
- For the strongly mean-reverting pairs, we can design trading strategies around them.

Problems with Using Cointegration

- The assets may be cointegrated sometimes but not always.
  - What do you do when it is not cointegrated but you are already in the market?
- Cointegration creates a dense basket – it includes every asset in the time series analyzed.
  - Incur huge transaction cost.
  - Reduce the significance of the structural relationships.
- Optimal mean reverting portfolios behave like noise and vary well inside the bid-ask spreads, hence not meaningful statistical arbitrage opportunities.
  - What about not so optimal ones?

Stochastic Spread

Ornstein–Uhlenbeck Process

- $z_t = x_t - \beta y_t$
- $dz_t = \theta(\mu - z_t)dt + \sigma dW_t$

Spread as a Mean-Reverting Process

- $x_k - x_{k-1} = (a - bx_{k-1})\tau + \sigma\sqrt{\tau}\varepsilon_k$
- $= b\left(\frac{a}{b} - x_{k-1}\right)\tau + \sigma\sqrt{\tau}\varepsilon_k$
- 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

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

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

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

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.
  - $\hat{x}_k = \hat{x}_{k|k} = E[x_k|Y_k]$

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

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

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.
  - $\max_{\vartheta} \mathrm{E}_X[L(\vartheta|Y, X)]$, or
  - $\max_{\vartheta} \mathrm{E}_X[\log L(\vartheta|Y, X)]$
- 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

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
  - $\hat{t} = \frac{1}{2}\ln\left[1 + \frac{1}{2}\left(\sqrt{(c^2-3)^2 + 4c^2} + c^2 - 3\right)\right]$

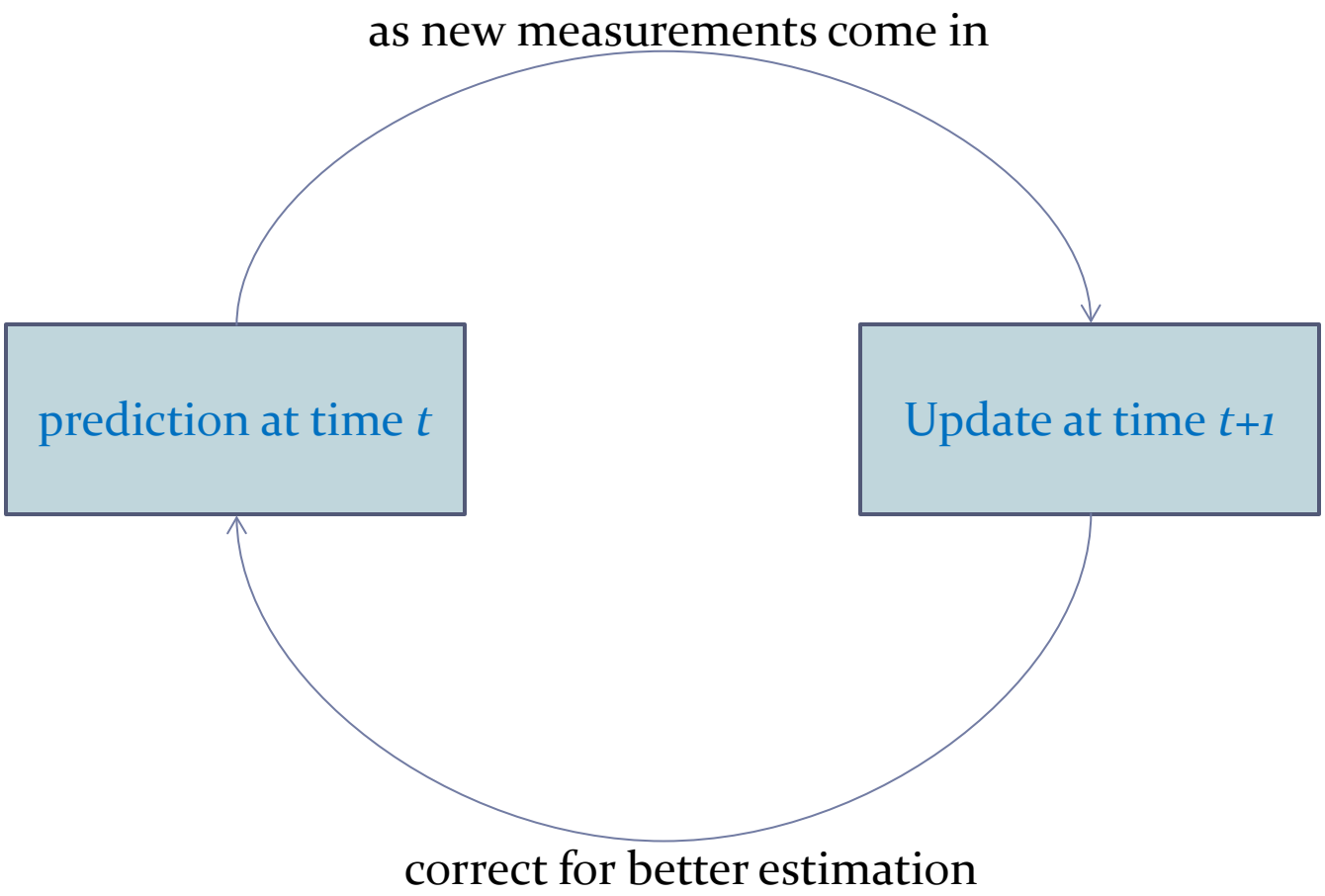
A Sample Trading Strategy

- $x_k = x_{k-1} + (a - bx_{k-1})\tau + \sigma\sqrt{\tau}\varepsilon_k$
- $dX(t) = \big(a - bX(t)\big)dt + \sigma dW(t)$
- $X(0) = \mu + c\frac{\sigma}{\sqrt{2\rho}}$, $X(T) = \mu$
- $T = \frac{1}{\rho}\hat{t}$
- 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



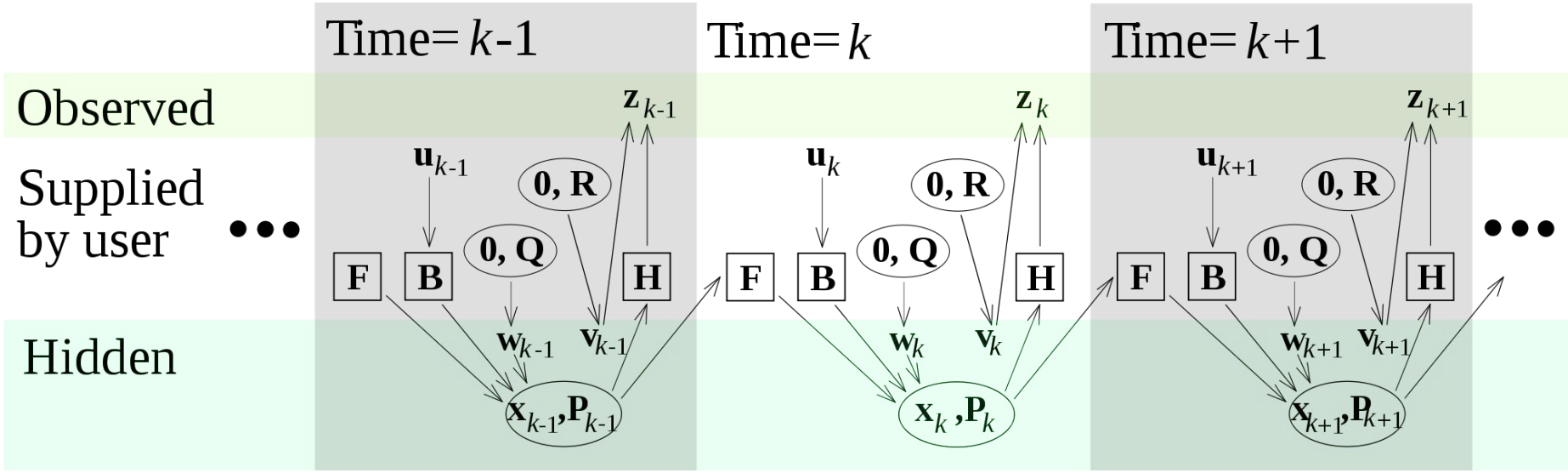
A Linear Discrete System

- $x_k = F_k x_{k-1} + B_k u_k + \omega_k$
- 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

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

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)

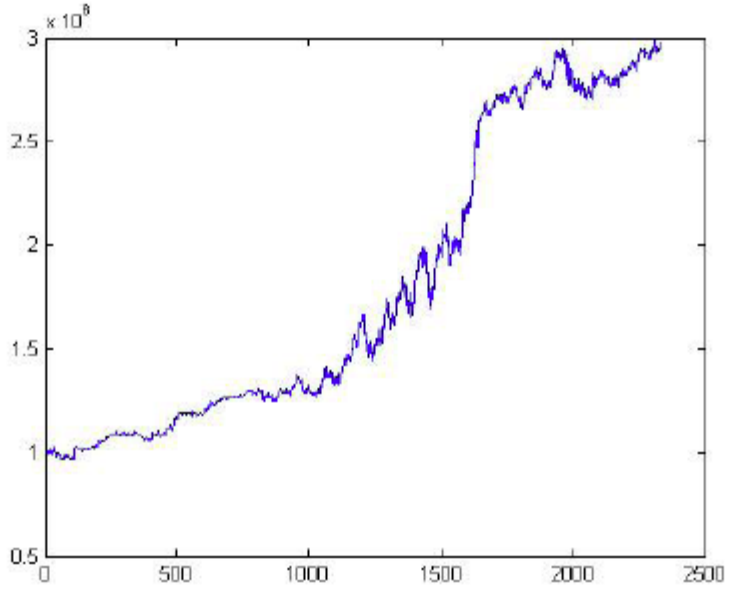


Figure 4: Backtesting Result with Optimization for weights on each pair

Results (2)

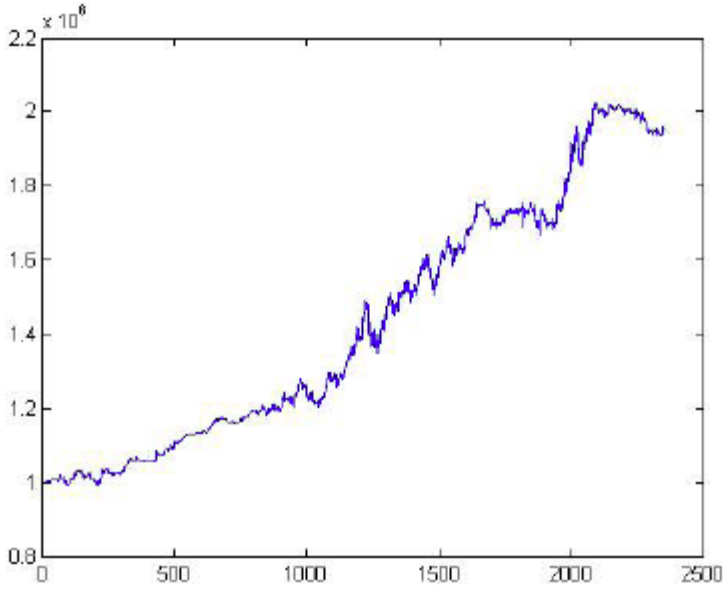


Figure 5: Backtesting Result Using Equal Weight Portfolio

Results (3)

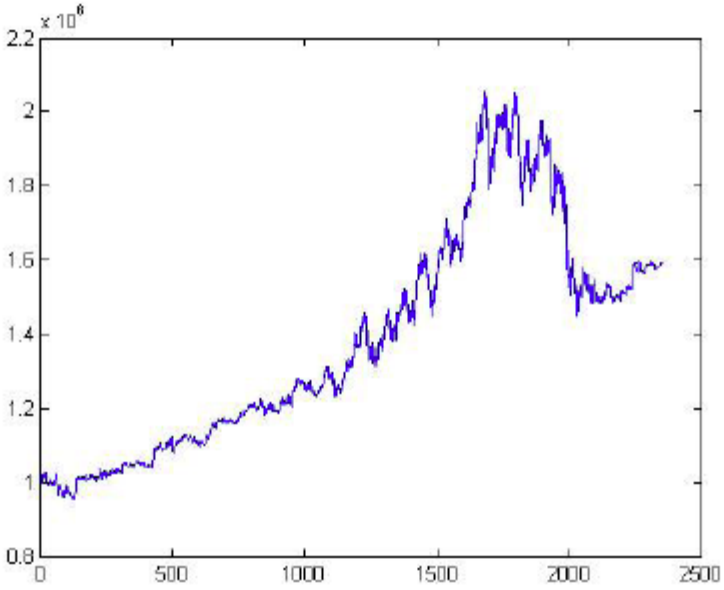


Figure 6: Backtesting Results Using 30-days Rolling Window Size