Non-Parametric Entropy

(II) Multiple Testing Procedures and Empirical Analysis with Applications to Pairs Trading Strategies

Kuan-Lun WANG [1]

National Taiwan University
https://www.csie.ntu.edu.tw/∼d06922002/

July 2, 2018

[0] This talk is the ITCT Final Presentation.
[1] E-mail: polyphonicared@gmail.com

Short Biodata

Kuan-Lun Wang is a doctoral student majoring in generalized pairs trading. The main goal of his research is to develop an algorithmic trading mechanism based on statistical arbitrage. His areas of expertise include automatic search procedures for model selection, multivariate co-integration approach, and structural change test.

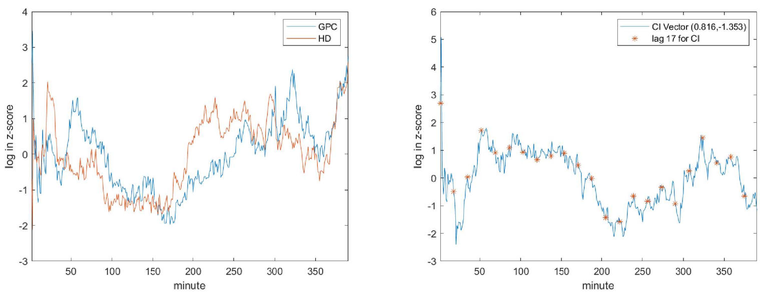
Research Interests

Kuan-Lun Wang's research interests comprise time series models, simulation modeling, and portfolio choice. The central themes of his application are the study of multivariate pairs trading in real time, search for assets with a long-run equilibrium, and building of riskless portfolios. Much of his current work involves conducting structural change analysis and co-integration test of the finite order vector autoregressive process and estimating the probability of mean reversion. Such methods are important in a variety of applications, including economic indicators and hedging. One such application is index funds being tied to indexes with very low costs and risks.

Outline

Should you have any questions, feel free to contact us.
(Don't believe me! Just check it by yourself!!)

Previously on the Last Talk (1/3)



(a) Log Prices Process (b) Mean Reversion Process

Figure: Genuine Parts Company and Hewlett-Packard Company

Previously on the Last Talk (2/3)

We consider a n-dimensional vector autoregressive process:

$$y_t = v + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t.$$

Moreover, the process has VECM representation:

$$\begin{aligned} &\Delta y_t \\ = {} & \alpha\beta' \begin{pmatrix} y_{t-1} \\ 1 \end{pmatrix} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p-1} + u_t \\ = {} & \text{initial} + \text{information}_{\text{lag}} + \text{information}_t. \end{aligned}$$

Previously on the Last Talk (3/3)

Assumption 1: Roots

The characteristic polynomial $A(z)$ satisfies the condition that if $|A(z)| = 0$, then either $|z| > 1$ or $z = 1$.

We will

1. estimate and test vector β [4];
2. normality test [5];
3. Multiple testing procedures [1];
4. I.I.D. Test for $\beta' y_t$.

Empirical analysis

- Minute bar of S&P 500 at 1.4.2016;
  - The number of stocks is 420;
  - Quandl sells data but I haven't money.
- Univariate pairs trading;
  - The number of PTS is 88,410;
  - My master thesis proposed a search algorithm.
- Only observation not moving windows;
  - The moving windows has some problem. e.g., structural change;
  - A structural change case is solved by my master thesis.

Estimate and Test Vector β (1/5)

Definition: VAR(p) [6]

Let y_t be a K-dimensional process as in

$$y_t = \nu + \sum_{i=1}^{p} A_i y_{t-i} + u_t, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where ν is a $(K \times 1)$ constant vector, $A_1, \ldots, A_p$ are $(K \times K)$ constant matries with $A_p \neq 0$, and u_t is independent identically distributed errors that are $\mathcal{N}(0, \Sigma_u)$. Then the process y_t is a VAR(p) process.

In this section, We consider u_t is independent identically distributed errors that are $\mathcal{N}(0, \Sigma_u)$.

Estimate and Test Vector β (2/5)

Definition: VECM Representation [6]

The VAR(p) process y_t has a VECM representation as in

$$\Delta y_t = \alpha\beta' y^*_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + u_t,\ \ t = 0, \pm 1, \pm 2, \dots$$

where $\alpha\beta' = (-(I_K - \sum_{i=1}^{p} A_i)', \nu')'$, $y^*_{t-1} = (y'_{t-1}, 1)'$, and $\Gamma_i = -\sum_{j=1}^{p-i} A_{i+j}$ for all $i = 1, \dots, p-1$, and the loading matrix and the cointegration matrix of VECM are α and β, respectively.

Assumption 2: The Constant Trend in the Cointegration Relations

The equation $\alpha U = \nu$ has solutions. [a]

[a] The assumption can be tested but we don't test in this talk [3].

Estimate and Test Vector β (3/5)

Denote $\hat{\cdot}$ be the maximized likelihood (ML) estimator for VECM. e.g., $\hat{\Sigma}_u$.
Denote r be the rank of β. e.g.,$\hat{\beta}_{r_0}$, $\hat{\Sigma}_u(r_0)$.
Then we have the following log-likelihood function value:

$$\log L(\hat{\beta}_{r_0}) = -\frac{T}{2} \log \left|\hat{\Sigma}_u(r_0)\right| - \frac{KT}{2}(1 + \log 2\pi),$$

and the likelihood ratio test statistic subject to the restrictions specified under null hypothesis $r = r_0$, is

$$\text{statistic}_r = 2\Big(\log L(\hat{\beta}_{r_0+1}) - \log L(\hat{\beta}_{r_0})\Big).$$

Estimate and Test Vector β (4/5)

However,

$$\text{statistic}_r \overset{d}{\nrightarrow} \chi^2_\cdot$$

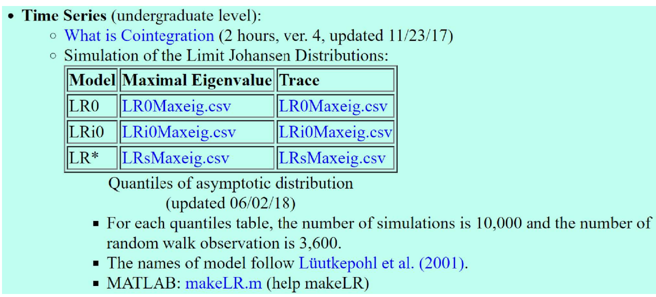
but

$$\text{statistic}_r \overset{d}{\to} \operatorname{tr}\left(\left(\int_0^1 (\mathrm{d}\,B_u)F_u'\right)\left(\int_0^1 F_uF_u'\right)^{-1}\int_0^1 F_u(\mathrm{d}\,B_u)'\right)$$

where B is the $(K - r_0)$-dimensional Brownian motion, and $F = (B', 1)'$. So, we shall simulate the asymptotic distribution.

Estimate and Test Vector β (5/5)

For each quantiles table, the number of simulations is 10,000 and the number of random walk observation is 3,600.



- **Time Series** (undergraduate level):
  - What is Cointegration (2 hours, ver. 4, updated 11/23/17)
  - Simulation of the Limit Johansen Distributions:

Model	Maximal Eigenvalue	Trace
LR0	LR0Maxeig.csv	LR0Maxeig.csv
LRi0	LRi0Maxeig.csv	LRi0Maxeig.csv
LR*	LRsMaxeig.csv	LRsMaxeig.csv

Quantiles of asymptotic distribution (updated 06/02/18)

- For each quantiles table, the number of simulations is 10,000 and the number of random walk observation is 3,600.
- The names of model follow Lüutkepohl et al. (2001).
- MATLAB: makeLR.m (help makeLR)

Figure: The sources are on my web.

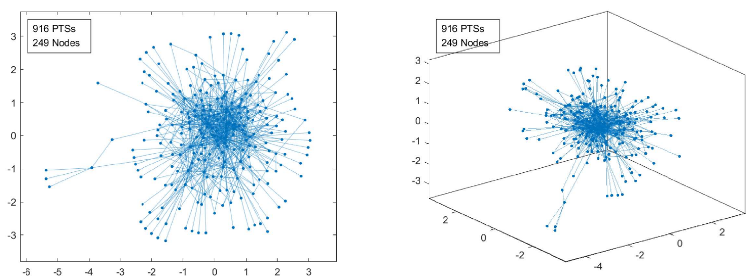
Normality test (1/1)

Denote $\tilde{\cdot}$ be the least square (LS) estimator for VAR(p). e.g., $\tilde{\Sigma}_u$.
Denote $\tilde{P}$ be the choleski decomposition fo $\tilde{\Sigma}_u$.
Then Jarque-Bera statistic is

$$\text{statistic}_J \equiv \frac{T\tilde{b}_1'\tilde{b}_1}{6} + \frac{T(\tilde{b}_2 - 3_K)'(\tilde{b}_2 - 3_K)}{24} \xrightarrow{d} \chi^2_{2K}$$

where $\tilde{b}_1 = \text{mean}((\tilde{P}^{-1}u_t)^{3_{K\times 1}})$ and $\tilde{b}_2 = \text{mean}((\tilde{P}^{-1}u_t)^{4_{K\times 1}})$. [5]

Empirical analysis



(a) 2-D Force-Directed Layout (b) 3-D Force-Directed Layout

Figure: Plot Graph Nodes and Edges

Multiple Testing Procedures (1/3)

Consider two hypotheses $H_\cdot^a$ and $H_\cdot^b$.
Consider the null hypothesis:

$$H_0^a \wedge H_0^b$$

Moreover, we have the following table:

	H_0^a	H_1^a	H_0^b	H_1^b
H_0^a	$1-\alpha$	α	X	X
H_0^b	X	X	$1-\alpha$	α

Table: The type I of multiple test

We see at once that

$$\alpha(2-\alpha) \geq \text{error}_\text{I} \geq \alpha.$$

Multiple Testing Procedures (2/3)

First, we use bootstrap estimation procedure. [1]

1. Generate B bootstrap sample under null hypotheses;
2. For each bootstrap sample, compute an test statistics $T_{m,b}$ for hypothesis m;
3. For each m, compute mean and variance of $T_{m,b}$;
4. Compute the null shift and scale-transformed test statistics null distribution:

$$Z_{m,b} \equiv \sqrt{\min\left\{1, \frac{\tau_{0,m}}{\text{Var}[T_{m,b}]}\right\}}(T_{m,b} - \text{E}[T_{m,b}]) + \lambda_{m,0},$$

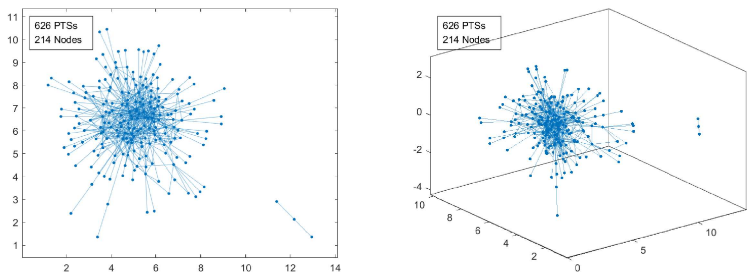
where $\lambda_{0,m}$ and $\tau_{0,m}$ are null mean and null variance, respectively.

Multiple Testing Procedures (3/3)

Second, step-down minP procedure show the following estimated unadjusted p-values: [1]

$$p_m = \Pr(Z_{m,b} \geq \text{statistic}_m) = \frac{1}{B}\sum_{b=1}^{B} I(Z_{m,b} \geq t_m).$$

Empirical analysis



(a) 2-D Force-Directed Layout (b) 3-D Force-Directed Layout

Figure: Plot Graph Nodes and Edges

Independent and Identically Distribution Test (1/1)

Denote $\hat{\mathcal{I}}_j$ be the nonparametric entropy estimator for Kullback-Leibler information criterion of lag j.
Denote $k(\cdot)$ and $k_b(\cdot)$ be quartic kernel and jackknife kernel, respectively.
Then Hong and White (2005) show

$$2hn\hat{\mathcal{I}}_j + hd \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

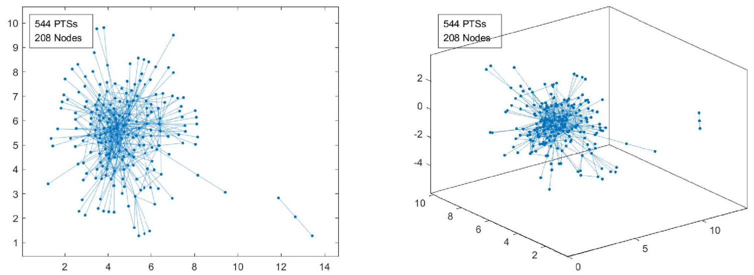
where h is the bandwidth for entropy estimator,

$$d \equiv \left(\left(\frac{1}{h} - 2 \right) \int_{-1}^{1} k^2(u) \, \mathrm{d}\, u + 2 \int_{0}^{1} \int_{-1}^{b} k_b^2(u) \, \mathrm{d}\, u \, \mathrm{d}\, b - 1 \right)^2,$$

and

$$\sigma^2 \equiv 2 \int_{-1}^{1} \int_{-1}^{1} \left[2k(u)k(u') - \int_{-1}^{1} k(u+v)k(v) \, \mathrm{d}\, v \times \int_{-1}^{1} k(u'+v')k(v') \, \mathrm{d}\, v' \right]^2 \mathrm{d}\, u \, \mathrm{d}\, u'.$$

Empirical analysis (1/2)



(a) 2-D Force-Directed Layout (b) 3-D Force-Directed Layout

Figure: Plot Graph Nodes and Edges

Empirical analysis (2/2)

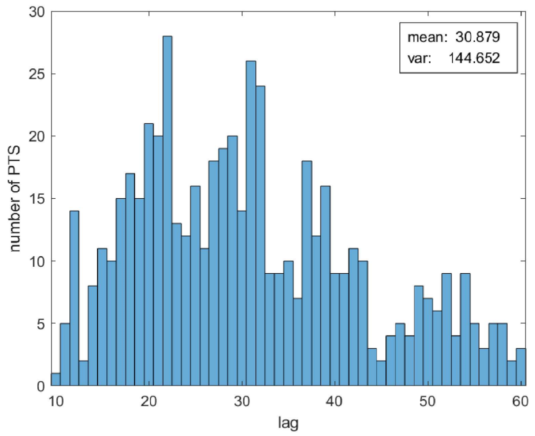


Figure: Histogram of Minimum IID-Lag PTS

Thank you for your time.

References (1/2)

[1] S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer New York, New York, NY, 2008.

[2] Y. Hong and H. White. Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica*, 73(3):837–901, may 2005.

[3] S. Johansen. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580, nov 1991.

References (2/2)

[4] S. Johansen. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, New York, 1995.

[5] L. Kilian and U. Demiroglu. Residual-based tests for normality in autoregressions: asymptotic theory and simulation evidence. *Journal of Business & Economic Statistics*, 18(1):40–50, jan 2000.

[6] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1 edition, 2005.