Computational Statistics and Data Analysis 128 (2018) 34–47

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Factor-adjusted multiple testing of correlations[☆]

Lilun Du, Wei Lan [*], Ronghua Luo, Pingshou Zhong

Hong Kong University of Science and Technology, Hong Kong
Southwestern University of Finance and Economics, China
Michigan State University, United States

ARTICLE INFO

Article history:
Received 4 December 2016
Received in revised form 8 April 2018
Accepted 3 June 2018
Available online 21 June 2018

Keywords:
Factor-adjusted correlation learning
False discovery rate
Pairs trading
Model selection consistency

ABSTRACT

Both global and multiple testing procedures have previously been proposed to untangle the correlation structures among high-dimensional data. In this article, we extend the results of both tests to learn the correlations of the factor-adjusted residuals in an approximate factor model, which can be used to simultaneously detect the highly matched pairs of stocks in finance. The factor-adjusted residuals are not observed and estimated using the method of principal components. We theoretically investigate the effects of estimating the factor-adjusted residuals on the subsequent global and multiple testing procedures. Furthermore, we demonstrate that the correlation structure of the factor-adjusted residuals can be recovered if appropriate thresholds are used in the proposed multiple testing procedure. Extensive simulation studies and a real data analysis are presented in which the proposed method is applied to select stock pairs in China's stock market.

1. Introduction

Pairs trading is one of the most notable short-term speculation strategies and proprietary statistical arbitrage tools are currently being used by hedge funds as well as many investment banks (Gatev et al., 2006). Pairs trading has been practiced for at least 20 years and is known to be effective (Gatev et al., 2006; Do and Faff, 2012). The idea of pairs trading is very simple. It involves two steps. The first step is selecting two comparable and highly associated stocks; the second step is trading in the selected pair based on their relative performance. Investors calculate the mean ratio (or spread) of their prices based on their historical performance. If the current price ratio (or spread) is significantly larger or smaller than the historical values, a portfolio can be constructed by buying the underpriced stock and selling the overpriced one. The idea is that in a long run there is a profit opportunity because the price ratio (or spread) is expected to revert to the mean.

The most fundamental and essential step in implementing pairs trading is to match stocks, say S_1 and S_2, whose price ratio (or price spread) will eventually return to "normal". That is, the price ratio (or price spread) is likely to display mean-reversion. If S_1 and S_2 are strongly associated, the stock price movement trends should behave similarly in a long run. As a result, if S_1 increases temporarily in price but S_2 does not, then we may expect that either the price of S_1 will decrease or the price of S_2 will increase in the future. That allows a portfolio to be constructed short on S_1 and long on S_2. The most important question is identifying suitable, similar pairs. From a statistical point of view, this amounts to finding two stocks with returns which are strongly positively correlated (i.e., $\rho_{S_1S_2} = \mathrm{corr}(S_1, S_2) > \rho_0$ for some pre-specified ρ_0). In practice, the investors

[☆] Wei Lan's research was supported by National Natural Science Foundation of China (NSFC, 11401482, 71532001). Lilun Du's research was partially supported by IGN15BM04, SBI16BM01, and Hong Kong RGC ECS26301216. Ping-Shou Zhong's research was supported by a National Science Foundation grant DMS 1309156. The authors are grateful to the Editor, the AE, and the anonymous referees for their insightful comments and constructive suggestions.

[*] Corresponding author at: Southwestern University of Finance and Economics, China.
E-mail address: facelw@gmail.com (W. Lan).

https://doi.org/10.1016/j.csda.2018.06.001
0167-9473/© 2018 Elsevier B.V. All rights reserved.

use a naive thresholding method to sort out stock pairs according to price movement correlations. They then select the most highly correlated pairs whose sample correlations exceed a pre-specified threshold level $\widehat{\rho}_0$. It is simple, but the procedure might be misleading in practice for two reasons. First, it has long been recognized theoretically and empirically that stock returns can be affected by some common factor (Sharpe, 1964; Fama and French, 1993; Fan et al., 2008, 2011). As a result, two stocks can be correlated due to common shock affecting the entire market. And such correlations can vary over time for unexplained reasons. So using the magnitude of the correlations directly may be misleading. A second consideration is that when a very large number of stocks is being considered, the number of possible stock pairs is extremely large. Testing the correlation of each stock pair separately can generate type I error, leading to a non-trivial multiple testing effect (Benjamini and Hochberg, 1995).

This study tested another way to choose $\widehat{\rho}_0$. It involves a two-step procedure for determining the factor-adjusted residuals using an approximate factor model (Fan et al., 2013). The first step involves testing if any pairs are correlated globally. If we reject the null hypothesis of global significance in the first step, in the second step factor-adjusted correlation learning (FACL) is applied to identify positively-correlated pairs with the false discovery rate (FDR) controlled for (Benjamini and Hochberg, 1995; Storey et al., 2004). The proposed two-step procedure shares similarities with those suggested by Cai and Jiang (2011) and Cai and Liu (2016), but with one key difference. The factor-adjusted residuals are estimated rather than observed. They are estimated using the method of principal components (Wang, 2012). This study has shown theoretically that the effects of estimating the factor-adjusted residuals on the subsequent global and multiple testing procedures vanish as the dimensionality increases, so that the results developed by Cai and Jiang (2011) and by Cai and Liu (2016) continue to hold. Note that many shrinkage and thresholding methods might be used to select and estimate the non-zero components of a large covariance (e.g., Bickel and Levina (2008a, b), Rothman et al. (2009), Cai and Liu (2011), Fan et al. (2013)). The procedure just described can select non-zero correlated pairs consistently if the tuning parameters are appropriately chosen, but it is very challenging to analyze the relationship between the FDR and the parameters. Compared with these methods, the proposed FACL procedure has two major advantages. First, the existing regularization approaches depend on the choice of tuning parameters, whereas the proposed testing procedure is tuning-free. Then the proposed FACL procedure hybridizes *variable selection* with *multiple testing*, not only providing consistent model selection under some sparsity conditions, but also controlling the false discovery rate to any pre-specified nominal level. These features are quite important, especially with finite samples (see, for example, Wasserman and Roeder (2009) and Meinshausen and Bühlmann (2010) for detailed discussions). It is worth mentioning too that the stability selection method of Meinshausen and Bühlmann (2010) is able to control the false positive rate asymptotically to a pre-specified level, but their method still involves selecting unknown tuning parameters, and the error control is usually quite conservative in finite samples.

Extensive simulation studies demonstrate that the two-step procedure can deliver satisfactory performance. To illustrate its practical utility, it was applied in this study to pairs trading in China's stock markets. The empirical findings show that FACL-based pairs trading can earn significantly higher returns and generate better Sharp ratios than the naive thresholding method, and is more reliable in practice. In the rest of this article most of the technical conditions and details are relegated to the Appendix.

2. Testing framework

2.1. Models and assumptions

Consider the following approximate factor model which has been much-investigated in economics and empirical finance (Fama and French, 1993; Bai, 2003; Wang, 2012; Fan et al., 2013),

$$Y_{it} = \boldsymbol{\beta}_i^\top \boldsymbol{X}_t + \varepsilon_{it}, \quad i = 1, \dots, N,\ t = 1, \dots, T. \tag{2.1}$$

Y_{it} here is the univariate observed response (stock return) for the ith stock at time t. Without loss of generality, assume that Y_{it} has been centralized such that $E(Y_{it}) = 0$ for $1 \le i \le N$, $1 \le t \le T$ and $\boldsymbol{X}_t$ is a $d \times 1$ vector (where d is unknown but a fixed number) of unobservable common factors with mean zero. The $\boldsymbol{\beta}_i \in \mathbb{R}^d$ are the associated unknown factor loadings, and ε_{it} is an error term, usually called the idiosyncratic component. Assume further that ε_{it} is independent of $\boldsymbol{X}_t$. For identification purposes, assume $\mathrm{cov}(\boldsymbol{X}_t) = \mathbf{I}_d$, where $\mathbf{I}_d$ is the identity matrix of dimension d. Unless explicitly stated otherwise, assume that the number of units N is much larger than the sample size T, and $N \to \infty$ as $T \to \infty$ for asymptotic analysis.

Model (2.1) can be represented in a matrix form as

$$\boldsymbol{Y}_t = \mathbf{B}\boldsymbol{X}_t + \mathcal{E}_t, \tag{2.2}$$

where $\boldsymbol{Y}_t = (Y_{1t}, \dots, Y_{Nt})^\top \in \mathbb{R}^N$, $\mathbf{B} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N)^\top \in \mathbb{R}^{N\times d}$ and $\mathcal{E}_t = (\varepsilon_{1t}, \dots, \varepsilon_{Nt})^\top \in \mathbb{R}^N$ $(1 \le t \le T)$ is the idiosyncratic error which follows a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{j_1 j_2}) \in \mathbb{R}^{N\times N}$. The corresponding correlation matrix is defined as $\mathbf{R} = (\rho_{j_1 j_2}) \in \mathbb{R}^{N\times N}$ with $\rho_{j_1 j_2} = \sigma_{j_1 j_2}/\{\sigma_{j_1 j_1}\sigma_{j_2 j_2}\}^{1/2}$.

Of particular interest is the structure of $\mathbf{R}$, the correlation matrix of the idiosyncratic errors. More specifically, the aim is to select pairs of stocks (e.g., j_1 and j_2 in $\{1, \dots, N\}$) that are highly correlated such that $\rho_{j_1 j_2} > \rho_0$ for some pre-specified positive constant ρ_0. A meaningful approach is to test all of the pairs $\rho_{j_1 j_2} > \rho_0$ simultaneously to select the significant pairs