International Journal of Theoretical and Applied Finance
Vol. 7, No. 3 (2004) 253–267

World Scientific
www.worldscientific.com

NFA FOR FACTOR NUMBER DETERMINATION IN APT

KAI-CHUN CHIU and LEI XU

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, P.R. China
kcchiu,lxu@cse.cuhk.edu.hk

Received 23 April 2003
Accepted 4 November 2003

In the context of quantitative analysis of the arbitrage pricing theory (APT) model, conventional factor analytic approaches such as maximum likelihood factor analysis (MLFA) cannot provide satisfactory answers to two important questions. The first one concerns the correct identification of factor number while the second one is related to the rotation indeterminacy of factor loadings. In the literature, MLFA followed by likelihood ratio (LR) test and the analysis of eigenvalues of sample covariance matrix were two popular approaches used to determine the appropriate number of factors. However, it was shown empirically that both of them suffered from different kinds of biases. We find the recently developed non-gaussian factor analysis (NFA) model by Xu [24] provides a new perspective for the determination of the appropriate factor number k in APT, with promising empirical results demonstrated.

Keywords: Factor analysis; arbitrage pricing theory; BYY harmony learning; model selection; statistical tests.

1. Introduction

Well-known in the finance literature, the arbitrage pricing theory (APT) assumes that the cross-sectional expected returns of securities follow a multi-factor model which is characterized by their sensitivities, usually called factor loadings, to k unknown economic factors. Pursuit to the original model [17], returns are generated under an *exact factor structure* in which the residual component of returns not explained by the factors is uncorrelated among securities. Conventional factor analytic approaches such as maximum likelihood factor analysis (MLFA) [11] were applied to recover both the factors and factor loadings and subsequent goodness-of-fit hypothesis test such as the likelihood ratio (LR) test was carried out to ascertain the minimum number of factors required to fit the model. However, two major limitations exist while applying MLFA on APT analysis and assuming the exact factor structure. First, since earlier studies showed that the number of significant factors as judged by the LR test increased with the number of securities q in the group, MLFA and LR test in general could not provide a satisfactory answer to the appropriate number of factors k for the APT model. The reason why k increased with q

might be explained by the higher probability of including securities with correlated idiosyncratic returns as q increase. Second, factor loadings as found by MLFA were not unique and were invariant to rotation.

In fact, there are many difficulties associated with the analysis of the APT model and in the literature, various efforts have been made to solve these problems. For example, it was shown in [8] that the problem of LR statistical test mentioned above could be improved by the method of cross validation. Although overfit of noise component to the model could be prevented by cross validation with out-of-sample data, it still could not guarantee to arrive at the correct factor number. The method tended to underestimate the number of factors most of the time because of the criteria of minimizing prediction error was not strictly related to finding the appropriate number of factors. Cross validation could fail in the case that several factors implicitly related were combined and appear to be a single factor. This might be a possible explanation why the number of factors determined via cross validation was consistently underestimated to be two in most cases.

On the other hand, attempt to solve both problems simultaneously by extending the exact factor structure to the so-called *approximate factor structure* was found in [5]. The rationale behind the proposition was that the key requirement for the APT that nonfactor risk be approximately eliminated through diversification could still be achieved without the assumption of a strict factor structure. The main difference between the exact factor structure and the approximate one was that the residual component of the latter being no longer uncorrelated, as revealed in [19]. For the approximate factor structure, weak correlation within the residual component was still acceptable. It has been shown in the same paper that the assumption of an approximate factor structure offers another perspective for the determination of factor number k and the unique identification of factor loadings. The use of an approximate factor model was intuitively more appealing because it made it probable that there exists some factor pertinent to a specific industry rather than to the whole market. Assuming an approximate factor structure, the LR statistic was no longer useful for the factor number determination purpose and it was proved in [5] that the analysis of eigenvalues of population covariance matrix was a suitable criterion. If k eigenvalues of the population covariance matrix increased without bound as the number of securities in the population increased, then the elements of the corresponding k eigenvectors of the covariance matrix could be used as factor sensitivities. Furthermore, it was shown in [7] that this conclusion held for the sample covariance matrix as well. Nonetheless, Brown [4] discovered that empirically the criterion typically biased towards too few factors and the result consistent with one factor might be equally consistent with k equally weighted factors that were priced. The reason was due to rotation of the original factors that minimizes the apparent number of priced factors. Moreover, Shukla and Trzcinka [20] criticized this approach on the ground that eigenvectors being used instead of statistical factor loadings in the returns-generating model for a large economy, possibly with infinitely many assets, did not necessarily imply that they

could be used in a cross-sectional model of security pricing in finite economies. The reason was that principal component analysis (PCA) was very different from factor analysis in the case of finite number of securities. PCA was more constrained than factor analysis for the application in APT because it overlooked idiosyncratic risks.

The APT model has also attracted the attention of researchers working in other fields. Specifically, a signal processing technique called independent component analysis (ICA) was applied in [1] to recover statistically independent non-gaussian distributed components (factors) using multivariate financial time series. It was well-known that ICA could exploit higher-order statistics of the underlying distribution to overcome the problem of rotation indeterminacy. Although it was reasonable to assume non-gaussian distribution of factors, the critical weakness of applying ICA for this specific task was due to its assumption of negligible idiosyncratic risks.

Recently, a new factor analytic technique called non-gaussian factor analysis (NFA) was developed under the framework of Bayesian Ying-Yang (BYY) harmony learning. According to [24], NFA is superior to MLFA in view of its ability to overcome rotation indeterminacy as well as determine the number of hidden factors k via its model selection ability. Consequently, it may serve as an alternative tool for traditional APT analysis. In this paper, we aim to explore the application of NFA in solving the two main problems in APT discussed above and compare the results with other conventional methods.

The rest of the paper is organized as follows. Section 2 reviews the original APT model on which the analysis is based and Sec. 3 briefly introduces the NFA model and highlight its benefits in the APT analysis. Section 4 describes the methodologies. Sections 5 and 6 are devoted to a comparative study on applying different approaches to determine the factor number k, using, respectively, hypothetical and real financial data. Section 7 concludes the paper.

2. The Arbitrage Pricing Theory

The APT begins with the assumption that the $n \times 1$ vector of asset returns, $\tilde{R}_t$, is generated by a linear stochastic process with k factors:

$$R_t = \bar{R} + Af_t + e_t\,, \tag{2.1}$$

where f_t is the $k \times 1$ vector of realizations of k common factors, A is the $n \times k$ matrix of factor weights or loadings, and e_t is a $n \times 1$ vector of asset-specific risks. It is assumed that f_t and e_t have zero expected values so that $\bar{R}$ is the $n \times 1$ vector of mean returns.

The model addresses how expected returns behave in a market with no arbitrage opportunities and predicts that an asset's expected return is linearly related to the factor loadings or

$$\bar{R} = R_f + Ap\,, \tag{2.2}$$

where R_f is a $n \times 1$ vector of constants representing the risk-free return, and p is $k \times 1$ vector of risk premiums. Similar to the derivation of the Capital Asset Pricing Model (CAPM), Eq. (2.2) is based on the rationale that unsystematic risk is diversifiable and therefore should have a zero price in the market with no arbitrage opportunities.

3. Non-Gaussian Factor Analysis

3.1. *An overview of NFA*

Suppose the relationship between a state $y_t \in \mathbf{R}^k$ and an observation $x_t \in \mathbf{R}^d$ are described by the equation as follows:

$$y_t = \varepsilon_t \,, \tag{3.1}$$

$$x_t = Ay_t + e_t \,, \quad t = 1, 2, \ldots, N \,, \tag{3.2}$$

where ε_t is non-Gaussian and component-wise independent, e_t is zero-mean gaussian white noise with $E(e_i e_j) = \Sigma_e \delta_{ij}$, Σ_e is a diagonal matrix whereas δ_{ij} is the Kronecker delta function:

$$\delta_{ij} = \begin{cases} 1 \,, & \text{if } i = j \,, \\ 0 \,, & \text{otherwise.} \end{cases} \tag{3.3}$$

In the context of APT analysis, Eq. (2.1) can be obtained from Eq. (3.2) by substituting $(\tilde{R}_t - \bar{R})$ for x_t and f_t for y_t. It is worth mentioning that the NFA model is defined such that the k sources $y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(k)}$ in this state-space model are statistically independent, i.e.,

$$p(y_t) = \prod_{j=1}^{k} p(y_t^{(j)}) \,, \tag{3.4}$$

where $y_t = [y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(k)}]^T$ and $p(y_t^{(j)}) = \sum_r \alpha_{j,r} G(y_t^{(j)} | m_{j,r}, \sigma_{j,r}^2)$. The objective of NFA is to estimate the sequence of y_t's with unknown model parameters $\Theta = \{A, \Sigma_e\} \cup \{\alpha_{j,r}, m_{j,r}, \sigma_{j,r}^2\}$ through the available observations.

In implementation, an adaptive algorithm suggested in [24] is shown below.

Step 1 Fix $\{\alpha_{j,r}, m_{j,r}, \sigma_{j,r}^2\}$ and $\{A, \Sigma_e\}$, find the factor $\hat{y}_t$ via $\hat{y}_t = \arg\max_{y_t} \ln [p(x_t|y_t)p(y_t)]$ where $p(y_t)$ is given by (3.4).

Step 2 Fix $\{A, \Sigma_e\}$ and y_t, update

$$\alpha_{j,r} = \frac{e^{c_{j,r}}}{\sum_l e^{c_{j,l}}} \,, \tag{3.5}$$

$$c_{j,r}^{\text{new}} = c_{j,r}^{\text{old}} + \eta \sum_l h_{j,l} [I_{l,r} - \alpha_{j,r}] \,, \tag{3.6}$$

$$m_{j,r}^{\text{new}} = m_{j,r}^{\text{old}} + \eta h_{j,r} e_{j,r} \,, \tag{3.7}$$

$$\sigma_{j,r}^{2,\text{new}} = \sigma_{j,r}^{2,\text{old}} + \eta(e_{j,r}^2 - \sigma_{j,r}^2)\,, \tag{3.8}$$

$$e_{j,r} = y_j - m_{j,r}\,, \tag{3.9}$$

$$h_{j,r} = \frac{\alpha_{j,r} G(y_t^{(j)}|m_{j,r}, \sigma_{j,r}^2)}{\sum_l \alpha_{j,l} G(y_t^{(j)}|m_{j,l}, \sigma_{j,l}^2)}\,, \tag{3.10}$$

$$I_{i,r} = \begin{cases} 1 & i = r\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.11}$$

Step 3 Fix $\{\alpha_{j,r}, m_{j,r}, \sigma_{j,r}^2\}$ and y_t, update

$$A^{\text{new}} = A^{\text{old}} + \eta e_t \hat{y}_t^T\,, \tag{3.12}$$

$$\Sigma_e^{\text{new}} = (1-\eta)\Sigma_e^{\text{old}} + \eta e_t e_t^T\,, \tag{3.13}$$

$$e_t = x_t - A y_t\,. \tag{3.14}$$

For the BYY harmony learning based NFA algorithm, the Yang machine by nature requires finding the y_t that maximizes the posterior, as in Step 1. Then, based on the sample x_t and the estimated y_t in Step 1, Steps 2 and 3 update the remaining parameters via the typical least mean square (LMS) criterion. Moreover, as discussed in [24], due to the least complexity property of the BYY harmony learning, the NFA algorithm is capable of selecting the number of factors. Furthermore, the problem of local optimization could be alleviated by the two newly introduced regularization techniques — data smoothing and normalization learning. Readers interested are referred to [24] for further details.

3.2. *Model selection versus factor number estimation*

Central to the discussion in the paper about the number of factors in APT, NFA is superior to MLFA in view of its model selection ability. In the context of APT analysis, the scale of complexity of the model is equivalent to the number of hidden factors in the original factor structure. As a result, model selection refers to deciding the appropriate number of factors in APT. According to Xu [22–24], we can achieve the aim of model selection by enumerating the cost function $J(k)$ with k incrementally and then select the appropriate k that makes $J(k)$ attain minimum:

$$\min_k J(k) = 0.5 \ln|\Sigma_e| + 0.5 \ln|A^T A| - \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{m} \ln p(y_t^{(j)})\,, \tag{3.15}$$

where Σ_e is the covariance matrix of the residual e_t, and

$$p(y_t^{(j)}) = \sum_r \alpha_{j,r} G(y_t^{(j)}|m_{j,r}, \sigma_{j,r}^2)\,.$$

3.3. *Benefits for using NFA in APT analysis*

For the analysis of non-gaussian stock return series, NFA has at least two distinct benefits over traditional factor analytic techniques. First, by considering higher order statistical information embedded in stock returns, the factors recovered are mutually independent and unique. Second, it can determine the number of hidden factors via its model selection ability. On the other hand, the typical advantage of NFA over ICA is the more realistic consideration of noise that is inherent in the model.

4. Methodologies

For hypothetical and real experiments, we compare the sensitivity of different tests on identifying the number of factors k with regard to the number of securities used. As discussed above, under the assumption of exact factor structure we will first use LR test on the results of MLFA. Then we will analyze the eigenvalues of the sample covariance matrix, assuming an underlying approximate factor structure. Finally, the results will be compared with that found by NFA's model selection criterion.

4.1. *Test statistics and methodology*

The LR statistic proposed by Lawley and Maxwell [12] and modified by Bartlett [2] is given by

$$LR = \left(N - \frac{2q + 4k + 11}{6}\right)\left\{\ln\frac{|AA' + \Sigma|}{|S|} + \mathrm{tr}[(AA' + \Sigma)^{-1}S] - q\right\}, \qquad (4.1)$$

where N is the sample size, S is the sample return covariance matrix and q is the total number of securities. The first and second terms in the sum of LR refer to the variance and bias components, respectively, of the statistic. Lawley and Maxwell [12] has shown that the maximum likelihood factor estimates are unbiased, and consequently, the bias component will converge asymptotically to zero. As a result, the LR statistic measures the overall error in the factor estimates of the sample covariances by comparing the generalized variances. When the normality assumptions apply, general properties of the LR statistic establish that it has an asymptotic central χ^2 distribution with $[(q-k)^2 - (q+k)]/2$ degrees of freedom. The minimum number of factors k can be inferred from the computed p value at a specific level of significance, which is 5% in this paper.

4.2. *Eigenvalues analysis*

Eigenvalues of the sample correlation or covariance matrix can be obtained either by direct calculation or indirectly via performing PCA. According to Chamberlain and Rothschild [5], for an approximate k-factor structure, the first k eigenvalues of

Table 1. Sensitivities of LR statistics to the number of securities.

	30 Securities			60 Securities			90 Securities		
k	D.f.	LR Stat.	p-Value	D.f.	LR Stat.	p-Value	D.f.	LR Stat.	p-Value
1	405	20355.77	0.0000	1710	50213.04	0.0000	3915	81222.22	0.0000
2	376	14165.34	0.0000	1651	35342.23	0.0000	3826	60980.38	0.0000
3	348	7804.46	0.0000	1593	20904.84	0.0000	3738	40183.88	0.0000
4	321	2557.15	0.0000	1536	7795.64	0.0000	3651	18283.34	0.0000
5	295	393.67	0.0001	1480	1975.91	0.0000	3565	4685.65	0.0000
6	270	336.46	0.0037	1425	1833.47	0.0000	3480	4470.17	0.0000
7	246	292.51	0.0224	1371	1717.06	0.0000	3396	4287.48	0.0000
8	223	231.96	0.3262	1318	1606.57	0.0000	3313	4122.15	0.0000
9	201	193.45	0.6361	1266	1503.81	0.0000	3231	3959.43	0.0000
10	180	158.99	0.8682	1215	1413.01	0.0001	3150	3804.32	0.0000
11	160	128.05	0.9702	1165	1327.76	0.0006	3070	3647.94	0.0000
12	141	102.82	0.9934	1116	1245.64	0.0039	2991	3499.17	0.0000
13	123	79.66	0.9991	1068	1167.69	0.0175	2913	3360.09	0.0000
14	106	63.05	0.9997	1021	1085.90	0.0776	2836	3227.70	0.0000
15	90	47.58	0.9999	975	1007.82	0.2266	2760	3100.28	0.0000
16	75	34.63	1.0000	930	938.24	0.4184	2685	2970.40	0.0001
17				886	872.65	0.6190	2611	2847.79	0.0007
18				843	809.76	0.7894	2538	2733.42	0.0036
19				801	750.14	0.9001	2466	2624.71	0.0131
20				760	691.98	0.9627	2395	2518.23	0.0392
21				720	636.04	0.9889	2325	2410.92	0.1048
22				681	585.90	0.9964	2256	2315.51	0.1872
23				643	541.86	0.9985	2188	2225.68	0.2822
24				606	497.72	0.9995	2121	2127.34	0.4572
25				570	412.22	1.0000	2055	2027.89	0.6607
26							1990	1932.17	0.8199
27							1926	1839.30	0.9204
28							1863	1758.81	0.9581
29							1801	1674.56	0.9841
30							1740	1589.23	0.9956
31							1680	1515.69	0.9983
32							1621	1424.41	0.9998

the covariance matrix of returns grow without limit as the number of securities, q, increases, while the remaining $q - k$ stay constant. However, for limited number of securities, we can only determine the factor number heuristically via counting the number of relatively large eigenvalues.

5. Determination of Factor Number Using Hypothetical Data

In this experiment, we assume returns of 30, 60 and 90 securities being generated randomly via a fixed number of factors 5 by the NFA model in Eqs. (3.1) and (3.2). The parameters used to generate $N = 1000$ data points are predetermined as follows:

A: A $p \times 5$ matrix where $q = 30$, 60 or 90.

$\varepsilon_{\mathbf{t}}$: Randomly generated by five componentwise independent non-gaussian densities $\prod_{i=1}^{5} p(y_t^{(i)})$.

$\mathbf{e_t}$: Randomly generated with pdf $G(e_t|0, \Sigma)$ where Σ is a $q \times q$ matrix with diagonal elements σ_{ii} and off-diagonal elements σ_{ij}, with $\sigma_{ii} \sim$ U(0.1,0.25), $\sigma_{ij} \sim$ U(0,0.01), where $U(r, s)$ denotes uniform distribution in the interval $[r, s]$.

Experimental results showing the number of factors identified by LR statistics are shown in Table 1 and that by eigenvalues analysis and $J(k)$ of NFA are shown in Table 2. As shown in Table 1, the minimum number of factors identified by MLFA and LR statistics is very sensitive to the number of securities under test and increases progressively with the number of securities. The acceptable number of factors at 5% level of significance is 8, 14 and 21 for 30, 60 and 90 securities respectively. On the other hand, evidence in Table 2 based on eigenvalues of sample covariance matrices seems to support the one-factor structure. Clearly, decision based on LR statistics tends to overestimate the number of factors while that based on eigenvalues tends to understate the same. The cost function $J(k)$ is not only insensitive to the number of securities, but also estimates the number of factors correctly in all three cases. Figure 1 plots the values of $J(k)$ against the number of factors for different number of securities.

6. Determination of Factor Number Using Real Financial Data

In this section, similar methodology discussed in the last section will be applied on historical stock data for the analysis of APT.

Table 2. Sensitivities of eigenvalues and $J(k)$ to the number of securities for factor number determination.

	Eigenvalues		
k	30 Sec.	60 Sec.	90 Sec.
1	36.4533	65.2531	141.2334
2	4.2125	12.5418	15.8765
3	3.5930	7.8534	9.0989
4	2.8207	6.2982	7.6715
5	1.9409	4.7302	6.2121
6	0.7865	1.7834	2.6785
7	0.2417	0.3184	0.3026
8	0.2242	0.3061	0.2920
9	0.2129	0.2938	0.2831
10	0.2035	0.2866	0.2702
11	0.1936	0.2733	0.2606
12	0.1871	0.2589	0.2582

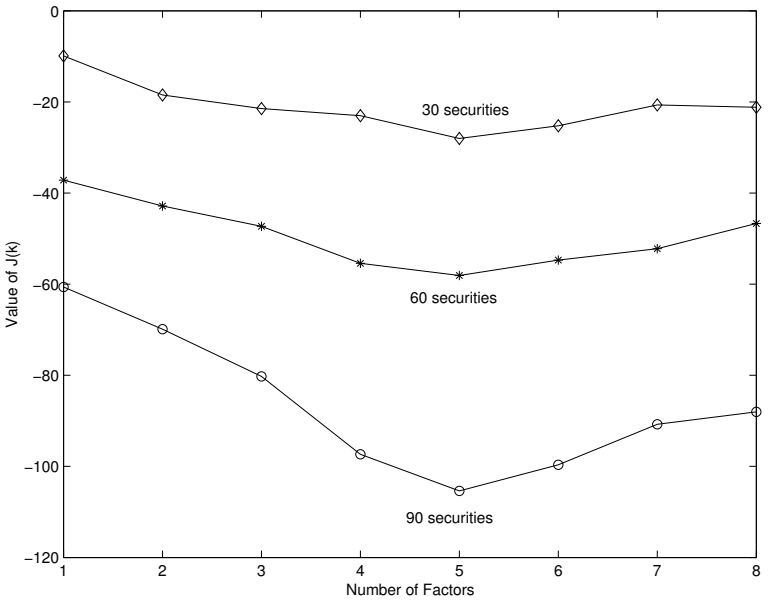

Fig. 1. $J(k)$ for different number of securities using hypothetical data.

6.1. *Data considerations*

We have carried out our analysis using past stock price and return data of Hong Kong. Daily closing prices of 86 actively trading stocks covering the period from January 1, 1998 to December 31, 1999 are used. The number of trading days throughout this period is 522. These stocks can be subdivided into three main categories according to different indices they constitute. Of the 86 equities, 30 of them belong to the Hang Seng Index (HSI) constituents, 32 are Hang Seng China-Affiliated Corporations Index (HSCCI) constituents and the remaining 24 are Hang Seng China Enterprises Index (HSCEI) constituents. In Hong Kong, stocks constituting HSI are known as "blue chips" and they consist of conglomerates with large market capitalizations. Most of them are domestic companies with a large proportion of business in Hong Kong. HSCCI constituents are composed of large companies with China affiliation but incorporated in Hong Kong. They are commonly referred to as "red chips" and they put approximately equal emphasis on business in both Hong Kong and China. HSCEI companies, also called "H-Shares" companies in Hong Kong, are mainly China incorporated and later become listed in Hong Kong. Their main business lines are located in China and they become listed in Hong Kong largely for the purpose of raising capital. Since companies constituting different indices have vastly dissimilar backgrounds, our intuition is that their stock prices may be affected by different number of factors. As a result, our analysis will be index-based as well as on all securities.

We do not adopt random sampling in the stock selection process so as to avoid the small-firm effect. This is because there are lots of small-sized listed companies in Hong Kong, many even very inactive. Obviously this kind of stocks is not representative of the whole market and including them will adversely affect the validity of any conclusions drawn from our analysis.

Regarding the observation frequency, we has considered using either daily or monthly equity returns. The potential benefit associated with the use of daily data in the estimation of variances and covariances is enormous. This is because the more frequent the observation, the more precise the parameter estimates. Moreover, for equal number of daily and monthly data points, the period spanned by daily data will be much shorter. As a result, our analysis will be just focused on a small period and is therefore less susceptible to large fluctuations such as structural break.

6.2. *Data preprocessing*

Before carrying out the analysis, the stock prices must be converted to stationary stock returns. The transformation applied can be described in four steps as shown below.

Step 1 Transform the raw prices to returns by $R_t = \frac{p_t - p_{t-1}}{p_{t-1}}$.
Step 2 Calculate the mean return $\bar{R}$ by $\frac{1}{N}\sum_{t=1}^{N} R_t$.
Step 3 Subtract $\bar{R}$ from R_t to get the zero-mean return.
Step 4 Let the result of above transformation be the adjusted return $\widetilde{R}_t$.

Table 3. Empirical results of factor number determination using real stock data: 30 HSI constituents.

k	D.f.	LR Statistics	p-Value	Eigenvalue
1	405	1753.21	0.0000	0.0180
2	376	1372.65	0.0000	0.0020
3	348	1051.90	0.0000	0.0014
4	321	746.88	0.0000	0.0013
5	295	575.57	0.0000	0.0012
6	270	463.34	0.0000	0.0011
7	246	396.55	0.0000	0.0009
8	223	320.69	0.0000	0.0008
9	201	265.05	0.0016	0.0008
10	180	212.93	0.0470	0.0008
11	160	175.97	0.1836	0.0007
12	141	146.20	0.3649	0.0006
13	123	112.61	0.7387	0.0006
14	106	91.02	0.8497	0.0006
15	90	66.69	0.9689	0.0005
16	75	50.61	0.9863	0.0005
17	61	34.31	0.9977	0.0004
18	48	23.10	0.9991	0.0004

6.3. *Empirical test results*

The aim of our experiments are to examine the relationship between the number of factors affecting stocks of various indices as well as the whole market. Table 7 gives an overview of the final results based on the details shown in Tables 3–6. From Table 7, we can see that the number of factors k determined based on the

Table 4. Empirical results of factor number determination using real stock data: 32 HSCCI constituents.

k	D.f.	LR Statistics	p-Value	Eigenvalue
1	464	2132.32	0.0000	0.0427
2	433	1597.10	0.0000	0.0048
3	403	1257.86	0.0000	0.0043
4	374	991.13	0.0000	0.0026
5	346	748.25	0.0000	0.0023
6	319	624.20	0.0000	0.0019
7	293	495.61	0.0000	0.0019
8	268	406.72	0.0000	0.0018
9	244	339.17	0.0001	0.0017
10	221	286.07	0.0021	0.0016
11	199	237.61	0.0318	0.0015
12	178	201.73	0.1074	0.0014
13	158	168.74	0.2649	0.0012
14	139	141.90	0.4158	0.0012
15	121	113.73	0.6678	0.0011
16	104	88.18	0.8668	0.0010
17	88	75.11	0.8346	0.0009
18	73	48.25	0.9888	0.0009
19	59	33.05	0.9975	0.0008
20	46	22.80	0.9984	0.0008
21	34	14.20	0.9989	0.0007

Table 5. Empirical results of factor number determination using real stock data: 24 HSCEI constituents.

k	D.f.	LR Statistics	p-Value	Eigenvalue
1	252	1155.54	0.0000	0.0294
2	229	707.41	0.0000	0.0034
3	207	519.18	0.0000	0.0027
4	186	402.98	0.0000	0.0022
5	166	335.53	0.0000	0.0020
6	147	273.94	0.0000	0.0018
7	129	223.13	0.0000	0.0015
8	112	171.50	0.0003	0.0015
9	96	135.15	0.0052	0.0014
10	81	105.64	0.0344	0.0012
11	67	77.31	0.1826	0.0012
12	54	47.45	0.7234	0.0011
13	42	29.45	0.9280	0.0011
14	31	16.74	0.9826	0.0010

Table 6. Empirical results of factor number determination using real stock data: All 86 securities.

k	D.f.	LR Statistics	p-Value	Eigenvalue
1	3569	13836.27	0.0000	0.0794
2	3484	8974.26	0.0000	0.0090
3	3400	8013.02	0.0000	0.0055
4	3317	7322.74	0.0000	0.0049
5	3235	6748.97	0.0000	0.0042
6	3154	6206.89	0.0000	0.0033
7	3074	5734.00	0.0000	0.0030
8	2995	5386.65	0.0000	0.0028
9	2917	5013.30	0.0000	0.0024
10	2840	4715.36	0.0000	0.0022
11	2764	4477.92	0.0000	0.0021
12	2689	4246.48	0.0000	0.0021
13	2615	4027.50	0.0000	0.0020
14	2542	3817.85	0.0000	0.0019
15	2470	3637.25	0.0000	0.0018
16	2399	3461.50	0.0000	0.0018
17	2329	3294.01	0.0000	0.0017
18	2260	3149.45	0.0000	0.0016
19	2192	3001.69	0.0000	0.0016
20	2125	2848.76	0.0000	0.0015
21	2059	2708.99	0.0000	0.0014
22	1994	2557.03	0.0000	0.0014
23	1930	2425.27	0.0000	0.0013
24	1867	2299.82	0.0000	0.0013
25	1805	2184.25	0.0000	0.0013
26	1744	2072.12	0.0000	0.0012
27	1684	1973.43	0.0000	0.0012
28	1625	1869.59	0.0000	0.0012
29	1567	1774.60	0.0002	0.0011
30	1510	1681.16	0.0013	0.0011
31	1454	1586.46	0.0082	0.0011
32	1399	1502.31	0.0275	0.0010
33	1345	1418.03	0.0813	0.0010
34	1292	1342.88	0.1584	0.0010
35	1240	1270.46	0.2676	0.0009
36	1189	1199.01	0.4136	0.0009
37	1139	1138.35	0.4999	0.0009
38	1090	1068.94	0.6699	0.0009
39	1042	999.59	0.8231	0.0008
40	995	919.63	0.9572	0.0008

Table 7. Results summary of factor number k determined based on real financial data.

Stock index	Total number of securities	MLFA	Eigenvalue	$J(k)$
HSI	30	11	1	4
HSCCI	32	12	1	4
HSCEI	24	9	1	5
All Securities	86	33	1	5

methodology of MLFA increases progressively with the number of securities included in a particular group. According to MLFA, there are 11 factors for HSI constituents, 12 for HSCCI constituents, 9 for HSCEI constituents and 33 for all market securities as a whole. On the other hand, the number of factors as revealed through the analysis of eigenvalues of sample covariance matrix is 1 irrespective of indices. The findings by the previous two methods can be contrasted with that discovered by the model selection criterion of NFA. Since the unique k associated with the minimum value of the cost function $J(k)$ corresponds to the appropriate factor number in APT, the factor numbers are 4 for both HSI and HSCCI and 5 for both HSCEI and all 86 securities. Figure 2 shows the plots of $J(k)$.

6.4. *Results interpretation*

The correct determination of factor number is critical for APT analysis. However, the issue of the appropriate number of factors has been the subject of some controversy in the literature such as Roll and Ross [14, 16, 15], Dhrymes, Friend and Gultekin [9], Trzcinka [21], Conway and Reinganum [8], and Brown [4]. Although Roll and Ross [14] believed that the number of factors is not more than five based on some empirical research findings, it was still far from conclusive because the tool on which the APT analysis was based suffers from various indeterminacies discussed in the previous sections. Interestingly, the factor number determined via the cost function $J(k)$ and NFA appears to agree with what suggested by Roll and Ross.

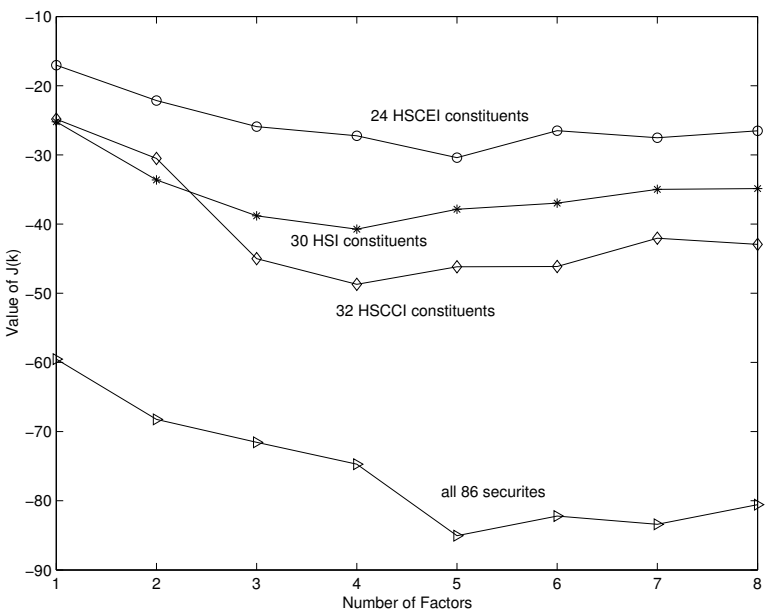

Fig. 2. $J(k)$ using different index constituents.

7. Conclusion

Two major obstacles of APT analysis via MLFA exist owing to its inability to determine the appropriate number of factors k and rotation indeterminacy. A comparative study on the effectiveness of different approaches towards identifying the factor number has been conducted. We find that LR test on results of MLFA is biased towards more factors while the identification via eigenvalues of sample covariance matrix tends to bias towards a smaller factor number. On the other hand, NFA not only can overcome rotation indeterminacy, but also provide a solution to determine the number of factors via a simple cost function $J(k)$ and its model selection ability.

Acknowledgments

The authors would like to express their gratitude to the anonymous reviewers for their comments and suggestions that improve the original manuscript. The work described in this paper was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (Project No: CUHK 4184/03E).

References

[1] A. D. Back and A. S. Weigend, A first application of independent component analysis to extracting structure from stock returns, *Int. J. Neural Syst.* **8** (1997) 473–484.
[2] M. S. Bartlett, Tests of significance in factor analysis, *Brit. J. Math. Stat. Psy.* **3** (1950) 77–85.
[3] M. Berry, E. Burmeister and M. McElroy, Sorting out risks using known apt factors, *Financ. Anal. J.* **44** (1988) 29–42.
[4] S. Brown, The number of factors in security returns, *J. Financ.* **44** (1989) 1247–1262.
[5] G. Chamberlain and M. Rothschild, Arbitrage and mean variance analysis on large asset markets, *Econometrica* **51** (1983) 1281–1304.
[6] N. F. Chen, R. Roll and S. Ross, Economic forces and the stock market, *J. Business* **59** (1986) 383–403.
[7] G. Connor and R. Korajczyk, Risk and return in an equilibrium APT: application of a new methodology, *J. Financ. Econ.* **21** (1988) 255–289.
[8] D. Conway and M. Reinganum, Stable factors in security returns: identification using cross-validation, *J. Bus. Econ. Stat.* **6** (1988) 1–15.
[9] P. Dhrymes, I. Friend and N. Gultekin, A critical reexamination of the empirical evidence on the arbitrag pricing theory, *J. Financ.* **39** (1984) 323–346.
[10] T. Estep, N. Hansen and C. Johnson, Sources of value and risk in common stocks, *J. Portfolio Manage.* **9** (1983) 5–13.
[11] K. G. Jöreskog, A general approach to confirmatory maximum likelihood factor analysis, *Psychometrika* **34** (1969) 183–202.
[12] D. N. Lawley and A. E. Maxwell, *Factor Analysis as a Statistical Method* (Butterworth, London, 1963).
[13] B. N. Lehmann and D. M. Modest, The empirical foundations of the arbitrage pricing theory, *J. Financ. Econ.* **21** (1988) 213–254.
[14] R. Roll and S. Ross, An empirical investigation of the arbitrage pricing theory, *J. Financ.* **35** (1980) 1073–1103.

[15] R. Roll and S. Ross, A critical reexamination of the empirical evidence on the arbitrage pricing theory: a reply, *J. Financ.* **39** (1984) 347–350.
[16] R. Roll and S. Ross, The arbitrage pricing theory approach to strategic portfolio planning, *Financ. Anal. J.* **40** (1984) 14–26.
[17] S. Ross, The arbitrage theory of capital asset pricing, *J. Econ. Theory* **13** (1976) 341–360.
[18] J. Shanken, The arbitrage pricing theory: is it testable, *J. Financ.* **37** (1982) 1129–1140.
[19] J. Shanken, The current state of arbitrage pricing theory, *J. Finance* **47** (1992) 1569–1574.
[20] R. Shukla and C. Trzcinka, Sequential tests of the arbitrage pricing theory: a comparison of principal components and maximum likelihood factors, *J. Financ.* **45** (1990) 1541–1564.
[21] C. Trzcinka, On the number of factors in the arbitrage pricing model, *J. Financ.* **41** (1986) 347–368.
[22] L. Xu, Bayesian Ying-Yang learning theory for data dimension reduction and determination, *J. Comput. Intelligence Financ.* **6** (1998) 6–18.
[23] L. Xu, Temporal BYY learning for state space approach, hidden markov model and blind source separation, *IEEE Trans. Signal Proces.* **48** (2000) 2132–2144.
[24] L. Xu, BYY harmony learning, independent state space and generalized APT financial analyses, *IEEE Trans. Neural Networ.* **12** (2001) 822–849.