Estimating Latent Factors Based on Statistical Data Analysis

Guoqing Xu and Guoxiao Yang*
School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China

*Corresponding author email: yanggx@bit.edu.cn

Abstract. In recent years, statistical methods have been widely used to estimate latent risk factors that affect the prices of financial assets. This paper develops new estimators for asset pricing factors by introducing dependence measure--distance covariance, that can identify nonlinear dependence. We combined distance covariance with Principal Component Analysis (PCA) and Risk-Premium PCA (RPPCA) and made contrast analysis based on Chinese market data. RPPCA, as a new method, shows strong applicability and detects factors with high Sharpe-ratio efficiently. Moreover, distance covariance produces better performance than covariance in PCA as a factor estimator, which illustrates the superiority of the distance covariance. Finally, the most striking results revealed by the study is that RPPCA including distance covariance of residuals outperforms others with a smaller pricing error and a significantly large Sharpe-ratio.

Keywords: Factor analysis; Distance covariance; Risk-premium PCA.

1. Introduction

Factor analysis has also been extensively studied in economics in the last decades. Approximate factor models assuming weakly correlated error terms (a.k.a. idiosyncratic components) have become the most popular tools for estimating latent factors. Bai and Ng carried on large dimensional factor analysis by an approximate factor structure [1]. Stock and Watson discussed the results respecting the use of factors in macroeconomic forecasting with approximate assumption [2]. Ludvigson and Ng analyzed the relation between bond excess returns and the macro economy in the approximate factor augmented regression framework [3].

Principal component analysis (PCA) is typically used to evaluate the unknown factors and loadings as a classical technique dealing with factor models. Bai and Ng adopted PCA to establish the number of latent factors in approximate factor models [4]. Breitung and Tenhofenthe compared the GLS estimators with usual principal components estimator [5]. Lam and Yao similarly utilized PCA to handle the factor models for high-dimensional time series [6].

The key application of factor analysis is to estimate asset pricing factors. Asset pricing has always been one of the most important topics in finance. Arbitrage Pricing Theory (APT) shows that prices of financial assets is explained by systematic risk factors[7]. In recent years, Factor models as a category of powerful statistical models has been widely used for financial asset pricing. Thus, how to choose factors is of vital importance. However, more than 300 potential factors have been proposed at present[8]. Large number and uncertain dependence between factors make choosing factors become very difficult. We hope to use fewer, uncorrelated, and most important factors to explain asset returns. In order to achieve this goal, many techniques have been developed to estimate statistical factors from large-dimensional panel market return data.

Initially, Connor and Korajczyk employed PCA to find asset pricing factors [9]-[10]. Based on factor analysis for S&P 500 index, Fan, Liao and Wang introduced the approach called Projected-PCA,
which wielded PCA to the smoothed data matrix onto the linear space spanned by covariates [11]. Kelly, Pruitt and Su proposed instrumented principal component analysis (IPCA) to figure out latent factors in application to equity asset pricing [12]. All these methods with different version of PCA don’t take into account expected returns. Lettau and Pelger generalized PCA with a penalty term to which was an estimator called Risk-Premium PCA (RPPCA) for factors accounting for expected excess returns [13]. RPPCA efficiently combines first and second moments and turns out to improve the factors estimation significantly [14]. We perform PCA and RPPCA to estimate latent factors with excess returns data of Chinese stock market in this study.

Distance covariance (dCov) and distance correlation (dCor) are new measures proposed by Székely et al. for pairwise dependence analogous to classical Pearson product-moment covariance and correlation[15]. Conventional covariance can only capture linear information between random variables, however, dCov tests both show significant results in some nonlinear and nonmonotone cases in Székely and Rizzo’s empirical applications [16]. Considering that distance covariance includes more dependence information than the covariance, we employs PCA to the sample distance covariance matrix replacing the sample covariance matrix to promote the explanation power for variation in the data. Moreover, inspired by the separation between linear and nonlinear components of bivariate dependence in EXAMPLE 5 in Székely and Rizzo’s paper[16], this study generalizes the framework of PCA and RPPCA to involve the distance covariance with respect to residual, which aims to make factors contain more dependence information and get better estimation.

The organization of the paper is as follows. Section 2 gives a brief description of distance dependence measures. Third section introduces the approximate factor models and supplies an intuition for methods. Section 4 details an implementation of factor estimator for Chinese market data. In the last section, the findings of study are discussed.

2. Distance Dependence Measures

2.1. Distance Covariance (dCov) and Distance Correlation (dCor)

Distance covariance (dCov) is a new approach to measure the dependence, which is described by the distance between the joint characteristic function of two random vectors and product of the respective characteristic function. It is similar to the classic Pearson product-moment covariance, but it is an extension and generalization.

Distance correlation (dCor) is a standardized form of distance covariance. Unlike the classical correlation, that distance correlation is 0 describes the independence between pairwise random vectors, and two random vectors can take arbitrary dimension without the same dimension necessarily.

Suppose that $X$ and $Y$ are two random vectors, $X \sim R^p$, $Y \sim R^q$, $E|X|_p < \infty$, $E|Y|_q < \infty$, $f_X$ and $f_Y$ denotes the characteristic functions of $X$ and $Y$ respectively. $f_{X,Y}$ denotes the joint characteristic function. The definition is as follows.

Distance covariance $V(X,Y) \in [0, \infty)$ is defined by

$$V^2(X,Y) = \left\| f_{x,y}(t,s) - f_X(t)f_Y(s) \right\|^2 = \frac{1}{c_p c_q} \int_{R^p \times R^q} \left\| f_{x,y}(t,s) - f_X(t)f_Y(s) \right\|^2 dt ds,$$

(1)

where $c = C(d,1) = \frac{p^{0.5d} \gamma^2}{G((1 + d)/2)}$, $|x|_p$ is the Euclidean norm in $R^p$.

Similarly, distance variance $V(X) \in [0, \infty)$ is given by
\[ V^2(X) = \left\| f_{X,s}(t,s) f_{X,s}(s) \right\|^2 \] (2)

Obviously,
\[ V(X, Y) = 0 \iff X \text{ and } Y \text{ are independent}. \] (3)

Distance correlation (dCor) between \( X \) and \( Y \) is denoted as \( R(X, Y) \), whose square form is:
\[ R^2(X, Y) = \begin{cases} \frac{V^2(X,Y)}{\sqrt{V^2(X)V^2(Y)}}, & V^2(X)V^2(Y) > 0, \\ 0, & V^2(X)V^2(Y) = 0. \end{cases} \] (4)

Distance correlation enjoys the property:
\[ 0 \leq R(X, Y) \leq 1, \quad \text{and} \quad R(X, Y) = 0 \iff X \text{ and } Y \text{ are independent}. \] (5)

Classical covariance and correlation can only capture the linear dependence between pairwise random variables. Distance covariance extends the ability of measuring dependence to various types of relations, which is no longer limited to linearity. Székely et al. provides the practical message that distance covariance test still has powerful performance in the case of nonmonotonic and nonlinear dependence [16]. The numerical experiments in Székely et al. [16] implies that distance covariance and correlation coefficient have more flexible application in all types of dependence, and they probably include additional information of dependence which is not detected by covariance and correlation. This is the reason why we draw distance covariance into factor analysis. Further, the sample versions of distance dependence measures are used in empirical application.

2.2. Distance Dependence Statistics
Let \( \{X', Y' : k = 1, K, n\} \) be an observed sample from the joint distribution of random vectors \( X \overset{\text{iid}}{\sim} \mathbb{R}^p \) and \( Y \overset{\text{iid}}{\sim} \mathbb{R}^q \), \( | \cdot | \) expresses the Euclidean distance, we note that
\[ a_{ui} = |X_i - X|, \quad \bar{a}_{ui} = \frac{1}{n} \sum_{i=1}^{n} a_{ui}, \quad \bar{a}_{*} = \frac{1}{n} \sum_{k,l=1}^{n} a_{ui}, \]
\[ b_{ui} = |Y_i - Y|, \quad \bar{b}_{ui} = \frac{1}{n} \sum_{i=1}^{n} b_{ui}, \quad \bar{b}_{*} = \frac{1}{n} \sum_{k,l=1}^{n} b_{ui}, \]
\[ A_{ui} = a_{ui} - \bar{a}_{*} - \bar{a}_{ui} + \bar{a}_{*}, \]
\[ B_{ui} = b_{ui} - \bar{b}_{*} - \bar{b}_{ui} + \bar{b}_{*}. \]

The form of the sample distance covariance \( V_n(X', Y') \) is as follows:
\[ V_n^2(X', Y') = \frac{1}{n^2} \sum_{k,j=1}^{n} A_{ui} B_{ui}. \] (6)

Similarly, the sample distance variance \( V_n^2(X') \) is the square root of
\[ V_n^2(X') = V_n^2(X', X') = \frac{1}{n^2} \sum_{k,j=1}^{n} A_{ui}^2. \] (7)
The sample distance correlation $R_n^2(X', Y')$ is defined by

$$R_n^2(X', Y') = \begin{cases} \frac{V_n^2(X', Y')}{\sqrt{V_n^2(X')}\sqrt{V_n^2(Y')}}, & V_n^2(X')V_n^2(Y') > 0, \\ 0, & V_n^2(X')V_n^2(Y') = 0. \end{cases}$$

(8)

The joint empirical characteristic function and the marginal empirical characteristic function are successively denoted as $f_{X',Y'}^n(t,s), f_X^n(t), f_Y^n(s)$, the scalar product of vectors $t$ and $s$ is denoted by $\langle t,s \rangle$. Naturally,

$$f_{X',Y'}^n(t,s) = \frac{1}{n} \sum_{k=1}^{n} \exp \{i \langle t, X_k \rangle + i \langle s, Y_k \rangle \}, \quad f_X^n(t) = \frac{1}{n} \sum_{k=1}^{n} \exp \{i \langle t, X_k \rangle \}, \quad f_Y^n(s) = \frac{1}{n} \sum_{k=1}^{n} \exp \{i \langle s, Y_k \rangle \}.$$

The following theorems hold, the norm $\| x \|$ is defined by (1).

**THEOREM 1.** If $(X', Y')$ is a sample from the joint distribution of $(X, Y)$, then

$$V_n^2(X', Y') = \left\| f_{X',Y'}^n(t,s) f_X^n(t) f_Y^n(s) \right\|^2.$$

**THEOREM 2.** If $E|X| < \infty$ and $E|Y| < \infty$, then

$$\lim_{n \to \infty} V_n^2(X', Y') = V(X, Y) \text{ a.s.}$$

**COROLLARY.** If $E\left( |X| + |Y| \right) < \infty$, then

$$\lim_{n \to \infty} R_n(X', Y') = R(X, Y) \text{ a.s.}$$

The proofs can be obtained in Székely et al. [15]. These results reveal that distance dependence statistics $V_n$ and $R_n$ possess almost sure convergence and are easy to compute, which shows that they are good sample statistics of distance dependence measures. R package energy provides necessary functions for the use of sample distance measures in empirical applications [17].

3. Factor Analysis

3.1. Approximate factor model

According to the arbitrage pricing theory, prices of financial assets depend on systematic risk factors [18]. It is presumed that excess returns of assets abide an approximate factor model, which means that excess returns have a K-dimensional factor structure with idiosyncratic components weakly correlated. We have a observation for excess returns of N assets throughout T time periods. To concretize the model, let $i = 1, \ldots, N$ index assets and $t = 1, \ldots, T$ index time. The approximate factor model is expressed by

$$X_i = F_{\infty}^{T}B_{\infty} + e_i$$

(9)

The matrix form is written as

$$X = \underbrace{F_{T \times N}}_{T \times K} B_{K \times N} + e_{T \times N}$$

(10)
where \( X \) represents the excess return matrix of \( N \) assets, \( F \) is the factor matrix composed of \( K \) factor vectors, \( B \) is the loading matrix and \( e \) stands for the weakly dependent idiosyncratic components (a.k.a. error terms or residuals). \( e = (e_1, \ldots, e_n) \) is assumed to be uncorrelated with \( F \) and have a zero mean value. The latent factors \( F \), the loadings \( B \) and residuals \( e \) are unknown. We stress that, only \( X \) can be observed in practice, \( F \) and \( B \) are both extracted from it by statistical techniques.

### 3.2. Methods

Lots of methods have been proposed for estimating statistical factors. Precisely, the most representative one is principal component analysis (PCA). PCA transforms the correlated original variables into several orthogonal comprehensive variables (principal components), which can achieve the purpose of dimensionality reduction [19]. The factors acquired by PCA can offer an explanation for a large proportion of the time-series variation, that is, systematic risk in asset pricing.

Risk-Premium principal component analysis (RPPCA) is a new technique for constructing factors, which adds the penalty term of first moment to PCA to ensure economic returns, considering that expected excess return can be explained by risk premium of factors [20]. The information incorporated in the first and second moments enables RPPCA to receive effective factor estimation.

Since distance measures can capture more types of dependence information (such as nonlinear dependence, etc.), this paper proposes several new methods, which combine the sample distance covariance with PCA and RPPCA respectively.

Then, we will elaborate on the above methods. The particulars are:

1. **Conventional PCA**: Principal component analysis is applied to sample covariance matrix, that is

   \[
   \frac{1}{T} X^T X - \overline{X} \overline{X}^T, \]

   where \( \overline{X} \) is the sample mean of excess return \( X \). The estimation of loadings \( \hat{B} \) is obtained by calculating the eigenvectors corresponding to the first \( K \) largest eigenvalues. Furthermore, we can estimate factors \( \hat{F} \) as

   \[
   \hat{F} = X \hat{B} \left( \hat{B}^T \hat{B} \right)^{-1}
   \]

2. **RPPCA**: The estimator utilizes PCA to sample covariance with weighted mean

   \[
   \frac{1}{T} X^T X + \gamma \overline{X} \overline{X}^T, \]

   where \( \gamma \) is called risk-premium weight. \( \hat{B} \) and \( \hat{F} \) are computed in the same way as (1).

3. **DCOVPCA**: The sample distance covariance matrix of excess return \( X \) is employed as a substitution for the sample covariance matrix in conventional PCA, whose computational means is shown in equation (6) in the second section.

4. **Residual DCOV+COV**: The first step is to estimate statistical factors through PCA, and then perform time series regression to obtain residuals \( e \). Next, PCA as an estimator for eventual potential factors is wielded to sample covariance plus sample distance covariance of residuals \( e \).

5. **Residual DCOV+RP**: It is almost the same as the method in (4), the only distinction is that PCA in the two steps is replaced by RPPCA.

The idea of (4) and (5) which hopes for more valid estimates is to use covariance to separate the linear information, and then add additional dependence through the sample distance covariance of residual.

In section 4, we estimate asset pricing factors based on real market return data by these methods.
4. Empirical Application

4.1. Data
The experimental data in our paper is monthly return data from August 1995 to May 2020 \(T = 298\) for 25 \(N = 25\) double-sorted portfolios according to size and book-to-market, which are constructed by all A-share stocks listed on the Shanghai and Shenzhen exchanges based on the approach of Fama and French [21]. The portfolios are total market value weighted. The monthly returns and risk-free interest rate which are used for acquiring excess return matrix \(X\) are both from RESSET database.

4.2. Estimation for Asset Pricing Factors
This paper subtracts the risk-free rate at the corresponding time from the monthly return data for 25 portfolios to obtain the monthly excess return matrix \(X\) and then establishes an approximate factor model to estimate its statistical factors. The methods include PCA, RPPCA with \(\gamma=10\) and \(\gamma=100\), DCOVPCA, Residual DCOV+COV, Residual DCOV+RP , which are mentioned in section 3.2. Root Mean Square Pricing Error \((RMS\alpha)\) and maximum Sharpe-ratio (SR) are used as performance measurements. \(RMS\alpha\) is formulated as Equation 12 \((\alpha_i\) is intercept term of the \(i\)-th asset time series regression).

\[
RMS\alpha = \frac{1}{N} \sum_{i=1}^{N} \alpha_i^2
\]  

(12)

SR is an indicator to measure the quality of factors in this study, where \(F\) denotes factors and weights \(\omega\) is \(\Sigma_f \mu_f\).

\[
SR = \frac{E(\omega^TF)}{\text{std}(\omega^TF\omega)}
\]  

(13)

4.3. Analysis of Experimental Results

| Methods                  | RMS\(\alpha\) | SR       |
|--------------------------|--------------|----------|
| PCA                      | 0.2464117    | 0.2323055|
| RPPCA \((\gamma=10)\)   | 0.2391588    | 0.2562936|
| RPPCA \((\gamma=100)\)  | 0.2397152    | 0.2806552|
| DCOVPCA                  | 0.2416015    | 0.2423794|
| Residual DCOV+COV       | 0.2544979    | 0.2184756|
| Residual DCOV+RP \((\gamma=100)\) | 0.2376864 | 0.2794347|
| Fama-French 3factor      | 0.2580313    | 0.2171651|

The results for \(K=3\) factors with different methods are shown by Table 1. As it is seen from the table, RPPCA is still effective under the Chinese market data, which shows better estimation ability for latent factors compared with PCA and DCOVPCA. Moreover, for \(\gamma = 100\) the Sharpe-ratio of RPPCA is the largest. It reveals that RPPCA may be apt to select factors with high Sharpe-ratio. Which is worthy of our attention, from the comparison between PCA and DCOVPCA, DCOVPCA has better performance with smaller pricing error and higher Sharpe-ratio. This implies us distance covariance reliably includes more possibilities and deserves to be explored further. Residual DCOV+RP provides the smallest root-mean-square pricing error among all methods and the second largest Sharpe-ratio, which indicates that adding the sample distance covariance of residual improves the pricing performance. A rational explanation is that after isolating the linear dependence, the distance covariance of residual supplement additional dependence information and a better factors
estimation is obtained. This further shows that the distance covariance can indeed capture nonlinear dependence from data and sufficiently verifies that the introduction of distance covariance is meaningful to estimate asset pricing factors. Meanwhile, factors obtained by statistical methods outperform Fama-French 3 factors.

![Figure 1. Cross-sectional for sorted portfolios (Size and BM).](image1)

The cross-sectional $\alpha$ for sorted portfolios of several methods can be seen from Figure 1. Figure 2 shows the loadings for different factors, which suggests methods adding distance covariance can detect factors well.

**5. Conclusion**

New techniques for latent asset pricing factors by involving distance covariance are proposed in this paper. Several distinct methods are carried out for latent factors based on Chinese market data. Briefly the findings are as follows. First of all, RPPCA has the same good applicability to Chinese market and is capable of picking up factors with high Sharpe-ratio. On the other hand, the experimental results
show that DCOVPCA performs better than PCA, which illustrates the superiority of the distance covariance. Strikingly, RPPCA including distance covariance of residuals dominates other methods with more efficient estimation for factors. Therefore, we highlight the significance of distance covariance respect to extracting asset pricing factors in a sense.

Whether RPPCA can be extended to other fields is worth researching. Including more dependence information offers the distance covariance more possibilities, and pretty well performance in the study also inspires us to make deeper research on the distance covariance for factors in the future work.

Acknowledgments

We are grateful for helpful support from School of Mathematics and Statistics, Beijing Institute of Technology.

References

[1] Bai J, Ng S. Large dimensional factor analysis [J]. Foundations and trends in econometrics, 2008, 3(2): 89-163.
[2] Stock J, Watson M. Macroeconomic forecasting using many predictors [M]. Handbook of Economic Forecasting, 2006.
[3] Ludvigson S, Ng S. A factor analysis of bond risk premia [M]. Handbook of the Economics of Finance, 2010.
[4] Bai J, Ng S. Determining the number of factors in approximate factor models [J]. Econometrica, 2002, 70: 191-221.
[5] Breitung J, Tenhofen J. GLS estimation of dynamic factor models [J]. J. Amer. Statist. Assoc, 2011, 106: 1150-1166.
[6] Lam C, Yao Q. Factor modeling for high-dimensional time series: Inference for the number of factors [J]. Ann. Statist, 2012, 40: 694-726.
[7] Chamberlain G, Rothschild M. Factor structure and mean-variance analysis in large asset markets [J]. Econometrica, 1983, 51(5).
[8] Harvey Campbell R, Liu H, Zhu. Editor's Choice: and the Cross-Section of Expected Returns[J]. The Review of Financial Studies, 2016, 29: 5-68.
[9] Connor G, Korajczyk R. Risk and return in an equilibrium apt: Application to a new test methodology [J]. Journal of Financial Economics, 1988, 21: 255-289.
[10] Connor G, Korajczyk R. A test for the number of factors in an approximate factor model [J]. Journal of Finance, 1993, 58: 1263-1291.
[11] Fan, J., Liao, Y., Wang, W., 2016. Projected principal component analysis in factor models. The Annals of Statistics 44 (1), 219–254.
[12] Kelly, B., Pruitt, S., Su, Y., 2017. Instrumented principal component analysis. Working Paper.
[13] Lettau M, Pelger M. Factors that fit the time series and cross-section of stock returns. Working paper, 2018.
[14] Lettau M, Pelger M. Estimating latent asset-pricing factors [J]. Social Science Electronic Publishing, 2018.
[15] Székely, Gábor J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances [J]. Annals of Statistics, 2007, 35(6):2769-2794.
[16] Székely, Gábor J, Rizzo M L. Brownian distance covariance [J]. The Annals of Applied Statistics, 2009, 3(4), 1236–1265.
[17] Rizzo M, G Székely. Energy: E-statistics (energy statistics). R package version 1.6.2, 2015.
[18] Chamberlain G, Rothschild M. Arbitrage, factor structure, and mean-variance analysis on large asset markets [J]. Scholarly Articles, 1982.
[19] Hotelling H. Analysis of a complex of statistical variables into principal components [J]. Journal of educational psychology,1933, 24: 417.
[20] Lettau M, Pelger M. Factors that fit the time series and cross-section of stock returns [J]. The Review of Financial Studies, 2020.
[21] Fama E F, French K R. Common risk factors in the returns on stocks and bonds [J]. Journal of Financial Economics, 1993, 33(1):3-56.