ELSEVIER Journal of Financial Markets 2 (1999) 403–432

Journal of FINANCIAL MARKETS

www.elsevier.nl/locate/econbase

Security factors as linear combinations of economic variables☆

Guofu Zhou*

Olin School of Business, Washington University, St. Louis, MO 63130, USA

Abstract

A new framework is proposed to find the best linear combinations of economic variables that optimally forecast security factors. In particular, we obtain such combinations from Chen et al. (Journal of Business 59, 383–403, 1986) five economic variables, and obtain a new GMM test for the APT which is more robust than existing tests. In addition, by using Fama and French's (1993) five factors, we test whether fewer factors are sufficient to explain the average returns on 25 stock portfolios formed on size and book-to-market. While inconclusive in-sample, a three-factor model appears to perform better out-of-sample than both four- and five-factor models. © 1999 Elsevier Science B.V. All rights reserved.

JEL classification: G12; G11; C11; C31

Keywords: Security returns; Factors; Forecasting; Linear models; Rank; APT

1. Introduction

It is perhaps safe to say that much of the empirical asset pricing research tries to find factors to explain security returns and the associated risk premiums.

☆The author is grateful to Stephen Buser, Zhiwu Chen, Elroy Dimson, Philip Dybvig, T.W. Epps, Christopher Geczy, Campbell Harvey, Raymond Kan, Lilian Ng, Carmela Quintos, seminar participants at Ohio State University, participants of the Twenty-Fourth European Finance Conference in Vienna, and especially John Geweke, Bruce Lehmann (the Editor) and an anonymous referee for providing many insightful and detailed suggestions that substantially improved the paper.

*Tel.: +1 314 935 6384; fax: +1 314 935 6359.

E-mail address: zhou@olin.wustl.edu (G. Zhou)

PII: S1386-4181(99)00008-7

There are basically two schools of thoughts about the factors. The first takes the stand that the factors are inherently latent and unobservable directly from market data. Models, such as Ross's (1976) arbitrage pricing theory (APT), provides theoretical justification for the use of latent factors. Although Sharpe (1964) and Lintner's (1965) capital asset pricing model (CAPM) identifies the factor as the return on the market portfolio, Roll (1977) provides convincing arguments for the unobservability of the market returns. Consistent with the latent factor assumption, there are two notable methods, Connor and Korajczyk's (1986) asymptotic principal components approach and the standard factor analysis approach (see, e.g., Seber, 1984) that can be used to extract factors from asset returns data. However, these approaches appear to have two weaknesses. First, they rely on information of security returns alone, making it difficult to link the estimated factors to economic fundamentals. Second, there is an errors-in-variables problem when the estimated factors are used to test the APT.[1] As the estimated rather than the actual factors are used in the test, the procedure can potentially give rise to incorrect testing results because of the errors-in-variables problem.

The second school of thought on security factors takes a more pragmatic stand. Rather than identifying (either observable or unobservable) factors by using any asset pricing theory, this school treats factors as pre-specified economic or financial variables that appear to be related to asset returns by simple financial reasoning or plain intuition. Examples of such studies include Chen et al. (1986) and Fama and French (1993). The advantage of this approach is its flexibility in including useful economic or financial variables into the analysis. But it may run into the danger of including too many variables that are highly correlated with one another and hence redundant, resulting in overstating the number of factors and inaccurate estimation of the parameters. Moreover, as highly correlated economic or financial variables are included, say, in a linear regression system, it is difficult to interpret the associated factor risk premiums because they may simply reflect the same source of economic risk.[2]

This paper provides a new framework that applies to both schools. For the second school which uses pre-specified economic variables as factors, a procedure based on the generalized method of moments (GMM) of Hansen (1982) is proposed to estimate the minimum number of factors that is needed to explain the asset returns (this number is not necessarily the number of pre-specified

[1] One-step procedures, such as the constrained maximum likelihood and Geweke and Zhou's (1996) Bayesian approaches, will not be subject to the errors-in-variables problem, but they may not be as robust as those developed here. See Connor and Korajczyk (1995) for an excellent survey of the literature on empirical studies of the APT.

[2] Measurement error is one of the major problems in the use of economic variables such as the CPI. However, this important problem is not addressed here because this paper focuses only on the question that how the number of variables can be optimally reduced from a set of given ones.

variables). The procedure also yields the best linear combinations (minimum number) of the pre-specified economic variables to serve as the non-redundant factors. An interesting feature of this method is that the factors so obtained are orthogonal to one another. As a result, it allows the use of potentially many pre-specified economic variables to generate a few orthogonal factors. Moreover, the procedure is useful for forecasting returns which can exhibit the very general heteroskedasticity as assumed in Hansen (1982), of which the often used normality assumption is a special case.

For the first school of thought which treats the factors as unobservables, we model the latent factors explicitly as a linear function of economic variables plus a noise. Then, the proposed procedure estimates those linear combinations of the economic variables that best forecast the latent factors. There are at least three interesting aspects of our new approach. First, it estimates the latent factors by using the best forecasts from the economic variables, naturally linking the latent factors to economic fundamentals. Second, the estimation of the factors is carried out jointly with the estimation of the loading parameters. As a result, the previous errors-in-variables problem is no longer present in our test of the APT. Third, it has the flexibility of using potentially many economic variables to analyze a factor model with only a few factors. In contrast, existing procedures, such as Chen, Roll and Ross (1986), impose a five-factor model specification when they use the information of five economic variables.

As an interesting application of the methodology, we examine Fama and French's (1993) factor model by testing whether fewer factors are sufficient to explain the average returns on 25 stock portfolios formed on size and book-to-market. We find that a three-factor model is better than using all of the five factors in terms of out-of-sample explanatory power while the in-sample performance is about the same. Furthermore, this three-factor model performs better than the three-factor model suggested by Fama and French (1993) in terms of minimizing the cross-sectional pricing errors.

This paper is organized as follows. In the second section, we provide an econometric framework for the second school of thought. In the third section, we show how the proposed approach can be easily adapted to analyze the first school of thought. In the fourth section, we test for the number of factors and find the best linear combinations of the variables that optimally forecast the latent factors in the Chen, Roll and Ross's (1986) model of the APT. In the fifth section, we apply the methodology to Fama and French's (1993) factor model. Conclusions and some remarks about future research are offered in the final section.

2. The second school of thought

In this section, we explain and model formally the second school of thought. First, we discuss and set up the econometric framework for obtaining the minimum number of factors and the associated best linear combinations from a set of pre-specified economic variables. Then, in Section 2.2, we show how to obtain analytical GMM estimates of the parameters and the associated GMM tests. Finally, in Section 2.3, we generalize the model to the case in which different sets of economic variables are used to form the best linear combinations out of each of the groups.

2.1. Obtaining factors from pre-specified ones

The second school of thought treats factors as pre-specified variables that appear to be related to asset returns by simple financial reasoning or plain intuition. Assume there are M such pre-specified observable factors (either economic or financial), $X_1, \dots, X_M$. Usually, it is assumed that the returns (or returns in excess a riskfree asset) are generated by the factors,

$$r_{it} = \alpha_i + \beta_{i1} X_{1t} + \cdots + \beta_{iM} X_{Mt} + \varepsilon_{it}, \quad i = 1, \dots, N, \tag{1}$$

where ε_{it} is the model disturbance which has a zero mean conditional on available information and can have the very general conditional heteroskedasticity as spelled out by Hansen (1982). As discussed earlier, the advantage of the second school of thought is the possibility of including potentially many useful economic or financial variables into the right-hand side of (1). However, this procedure may include too many variables than necessary, so that some of the variables are correlated and hence redundant. The objective here is to minimize this and other related problems by showing how to obtain the minimum number of orthogonal factors from the M pre-specified ones. As it turns out, the minimum number of factors that are necessary in (1) depends on the rank of the regression coefficient matrix $\boldsymbol{B} = (\beta_{ij})$, $N \times M$, formed by the betas.

Indeed, if $\text{rank}(\boldsymbol{B}) = K < M$, it is easy to show that there must exist an $N \times K$ matrix $\boldsymbol{A}$ and a $K \times M$ matrix $\boldsymbol{C}$ such that $\boldsymbol{B}$ is their product,

$$\mathrm{H}_0: \quad \boldsymbol{B} = \boldsymbol{A}\boldsymbol{C}, \quad \boldsymbol{A}\colon N \times K, \quad \boldsymbol{C}\colon K \times M. \tag{2}$$

Then, Eq. (1) can be written as

$$r_{it} = \alpha_i + A_{i1} f_{1t} + \cdots + A_{iK} f_{Kt} + \varepsilon_{it}, \quad i = 1, \dots, N, \tag{3}$$

where $f_{1t}, \dots, f_{Kt}$ are new factors which are linear combinations of the pre-specified ones,

$$f_{kt} = C_{k1} X_{1t} + \cdots + C_{kM} X_{Mt}, \quad k = 1, \dots, K. \tag{4}$$

Hence, through perhaps a rotation, we can assume that the K factors as expressed in (4) are orthogonal. Eq. (3) states that, out of the M individual factors, only K linear combinations of them are sufficient to explain the returns. On the other hand, if there are $K < M$ orthogonal factors as expressed in (4), it can be shown that the rank of $\boldsymbol{B}$ must be less than or equal to K. Therefore, the determination of the minimum number of orthogonal factors becomes the problem of finding the rank of $\boldsymbol{B}$. The determination of rank ($\boldsymbol{B}$), the estimation of $\boldsymbol{A}$ and $\boldsymbol{C}$ and the associated GMM tests are provided in the next subsection.

The above procedure has the advantage of using potentially many variables into analysis, and yet keeping the number of parameters to be estimated at a minimum. Hence, as well-known in econometrics, this can enhance the estimation accuracy and the reliability of the model. From a forecasting point of view, out-of-sample performance may be improved because of fewer parameters (while the in-sample performance is more or less the same). Furthermore, the newly constructed minimum factors are likely to group correlated economic variables together so that the factor risk interpretation is more apparent from an economic standpoint.

However, given, say, ten pre-specified economic variables, how do we interpret a new factor that is a linear combination of the original ten? The linear combination is in some ways like an economic index. We can interpret it as the factor that drives the returns with the combination coefficients representing the weights or the contributions of the individual economic variables to the index. For example, if one is a true believer in the CAPM, and if there are ten pre-specified variables, then any one of the variables alone is unlikely to represent the market factor, while a linear combination of all of them is much more likely to do so. Suppose the sum of the combination coefficients is one (after scaling), and the first accounts for 80% and the last for 5%. Then, we can say that the first and the last variables contribute 80% and 5%, respectively, to the market factor.

As shown above, the minimum number of factors necessary in (1) is rank($\boldsymbol{B}$), and composition of the new factors is determined by the decomposition of $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{C}$. An important question is how this is different from looking at the eigenvalues of the matrix of the (pre-specified) factor data, by which the number of non-zero eigenvalues may suggest the number of factors, and the associated eigenvectors may serve as the new factors. The answer is that the eigenvalue approach is the standard principal components analysis which extracts the most significant factors out of the pre-specified ones to have the maximum variation. In contrast, the objective here is to extract the best factors (as linear combinations of the pre-specified ones) that best explain the asset returns. Hence, the solution here differs from that of the principal components analysis. In addition, our framework is very general. For example, Eq. (1) can be used solely for forecasting purposes. In this case, all the pre-specified factors can be any stationary processes that are available at time t for forecasting the asset returns. Then, our procedure reduces the dimensionality of the model by finding fewer factors than potentially too many, which generally yields better parameter estimates and better out-of-sample performance.

2.2. GMM estimation and tests

In this subsection, we provide a detailed derivation for both the parameter estimation and hypothesis testing problems associated with (1) and (2). Those who are interested only in applications may go directly to Eqs. (9) and (11) for the results.

To simplify the notation that follows, we write Eq. (1) in vector form

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{X}_t + \boldsymbol{U}_t, \tag{5}$$

where $\boldsymbol{R}_t$ is an N-vector of the returns, $\boldsymbol{X}_t$ and $\boldsymbol{U}_t$ are defined accordingly. Let $\boldsymbol{Z}_t$ be an L-vector of instrumental variables ($L \geqslant M$). Notice that depending on specific applications, $\boldsymbol{X}_t$ can be either time t or time $(t-1)$ variables. If all of $\boldsymbol{X}_t$ are time $(t-1)$ or earlier, Eq. (1) is used entirely for forecasting purposes. In either case, $\boldsymbol{Z}_t$ may contains a subset of a constant, $\boldsymbol{X}_t$ and its past, and the past of $\boldsymbol{R}_t$, such that we can write the moment conditions of (5) as $E[\boldsymbol{U}_t \otimes \boldsymbol{Z}_t] = \boldsymbol{0}$. Suppose the null H_0 holds that rank$(\boldsymbol{B}) = K$. Following Hansen (1982), we can obtain the GMM estimator of the parameters by minimizing a weighted quadratic form of the sample moment conditions:

$$\min Q \equiv \boldsymbol{g}_T' \boldsymbol{W}_T \boldsymbol{g}_T, \tag{6}$$

over the parameter space of $(\boldsymbol{\alpha}, \boldsymbol{B})$ satisfying rank$(\boldsymbol{B}) = K$, where $\boldsymbol{g}_T = (1/T)\sum_{t=1}^{T} \boldsymbol{f}_t, \boldsymbol{f}_t \equiv \boldsymbol{U}_t \otimes \boldsymbol{Z}_t$, an NL-vector, and $\boldsymbol{W}_T$, $NL \times NL$, is a positive-definite weighting matrix.

As shown in (2), under the null we have $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{C}$ for some $N \times K$ matrix $\boldsymbol{A}$ and $K \times M$ matrix $\boldsymbol{C}$. So, we need to estimate only $(\boldsymbol{\alpha}, \boldsymbol{A}, \boldsymbol{C})$. However, the estimator of $(\boldsymbol{\alpha}, \boldsymbol{A}, \boldsymbol{C})$ is not unique because for any given estimator $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{C}})$, a linear transformation of it, $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{A}}\boldsymbol{H}, \boldsymbol{H}^{-1}\tilde{\boldsymbol{C}})$ gives rise to the same estimator of $(\boldsymbol{\alpha}, \boldsymbol{B})$, where $\boldsymbol{H}$ is any $K \times K$ nonsingular matrix. Clearly, if there are no redundant assets, the first $K \times K$ submatrix of $\boldsymbol{A}$ must be nonsingular. Hence, we can use the normalization that $\boldsymbol{A}' = (\boldsymbol{I}_K, \boldsymbol{A}_2)$ and then the estimator will be unique. Let $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\alpha}, \boldsymbol{A}_2, \boldsymbol{C})$. Thus, the GMM estimator is uniquely obtained by solving the minimization problem (6) over the parameter space of $\boldsymbol{\theta}$. The number of parameters is either $q = N + (N-K)K + MK$ when the alphas are unconstrained or $q = (N - K + M)K$ when the alphas are constrained to be zeros (some other models of this paper impose such constraints).

With different choices of the weighting matrix, one obtains different estimators. Hansen (1982) suggests a weighting matrix that yields the optimal estimator which has the minimum asymptotic covariance matrix among all the

GMM estimators. Depending on the type of heteroskedasticity assumed, the optimal weighting matrix may take different forms. In general, $\boldsymbol{W}_T = \boldsymbol{S}_T^{-1}$, where $\boldsymbol{S}_T$ is a consistent estimator of the covariance matrix of the model residuals. In multivariate regression (5), the first order conditions are $E(\boldsymbol{U}_t \otimes \boldsymbol{Z}_t) = \boldsymbol{0}$. To specify an estimator for the covariance matrix of the residuals, it is natural to assume that the first order conditions continue to hold conditional on the information set $\boldsymbol{Z}_t$, then[3] a consistent estimator, $\boldsymbol{S}_T$, of the covariance matrix is, as given by MacKinlay and Richardson (1991),

$$\boldsymbol{S}_T = \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{U}_t \boldsymbol{U}_t' \otimes \boldsymbol{Z}_t \boldsymbol{Z}_t'). \tag{7}$$

In practice, a consistent estimate of $\boldsymbol{\theta}$ is often obtained first by choosing the weighting matrix as the identity matrix. Then, a weighting matrix of $\boldsymbol{W}_T = \boldsymbol{S}_T^{-1}$ is chosen to obtain the optimal GMM estimator, where $\boldsymbol{S}_T$ is evaluated at the initial estimator. However, it is generally difficult to solve the GMM optimization problem (6) under the complex nonlinear rank hypothesis for a generic weighting matrix.

Fortunately, with any weighting matrix of the following type (which includes the identity matrix as a special case),

$$\boldsymbol{W}_T \equiv \boldsymbol{W}_1 \otimes \boldsymbol{W}_2, \quad \boldsymbol{W}_1 \colon N \times N, \quad \boldsymbol{W}_2 \colon L \times L, \tag{8}$$

the estimator of $\boldsymbol{\theta}$ can be *analytically* obtained. The analytical solutions have at least four advantages compared with the numerical ones. First, the success of a numerical optimization algorithm usually depends on how close an initial estimate to its true solution, and a good initial estimate is generally not easy to obtain. Second, it is well-known that as the number of parameter increases, it is more and more difficult to obtain numerical optimizing solutions. Third, numerical algorithms often converge only to a local maximum or minimum and may not converge at all. Fourth, analytical solutions make Monte Carlo studies easy to accomplish.

To obtain the analytical GMM estimators with a weighting matrix of form (8), consider first the case where $\boldsymbol{\alpha} = \boldsymbol{0}$. In this case, we only need to estimate $\boldsymbol{A}$ and $\boldsymbol{C}$. Based on Zhou (1994), it follows that the analytical GMM estimators are

$$\tilde{\boldsymbol{C}}' = (\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}/T^2)^{-1/2}\boldsymbol{E}, \qquad \tilde{\boldsymbol{A}}' = (\tilde{\boldsymbol{C}}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\tilde{\boldsymbol{C}}')^{-1}\tilde{\boldsymbol{C}}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{R}, \tag{9}$$

[3] Other assumptions may also be made, then $\boldsymbol{S}_T$ may no longer have an expression like (7), and an estimator, such as Newey and West's (1987), may have to be used instead. However, this presents no difficulties at all because $\boldsymbol{S}_T$ enters into the GMM test only through Eq. (12) which, for an arbitrary $\boldsymbol{S}_T$, can be computed as easily as for an $\boldsymbol{S}_T$ of form (7).

where $\boldsymbol{P} \equiv \boldsymbol{Z}\boldsymbol{W}_2\boldsymbol{Z}'$, $\boldsymbol{Z}$ is a $T \times L$ matrix of the instruments, $\boldsymbol{X}$ is a $T \times M$ matrix of the pre-specified factors, $\boldsymbol{R}$ is a $T \times N$ matrix of the returns, and $\boldsymbol{E}$ is an $M \times K$ matrix stacked by the 'standardized' eigenvectors ($\boldsymbol{E}'\boldsymbol{E} = \boldsymbol{I}_K$) corresponding to the first K largest eigenvalues of the following $M \times M$ matrix:

$$\boldsymbol{\Omega}_T = (\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}/T^2)^{-1/2}(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{R}/T^2)\boldsymbol{W}_1(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{R}/T^2)'(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}/T^2)^{-1/2}. \tag{10}$$

Notice that the GMM estimator given in (9) is not normalized, but can easily be transformed into the normalized form by multiplying $\tilde{\boldsymbol{A}}$ by $\boldsymbol{H}$ (from the right) and $\tilde{\boldsymbol{C}}$ by $\boldsymbol{H}^{-1}$ (from the left), where $\boldsymbol{H}$ is the inverse of the first $K \times K$ submatrix of $\tilde{\boldsymbol{A}}'$. Clearly, one can start the GMM estimation with the identity weighting matrix. Then, the inverse of $(\sum_{t=1}^{T} \boldsymbol{U}_t\boldsymbol{U}_t'/T) \otimes (\sum_{t=1}^{T} \boldsymbol{Z}_t\boldsymbol{Z}_t'/T)$ is easily computed which is the optimal weighting matrix under the assumption that the residuals are independent, and identically distributed (i.i.d.). This matrix is obviously of form (8). Therefore, with $\boldsymbol{W}_T$ being either the identity weighting matrix or the inverse of $(\sum_{t=1}^{T} \boldsymbol{U}_t\boldsymbol{U}_t'/T) \otimes (\sum_{t=1}^{T} \boldsymbol{Z}_t\boldsymbol{Z}_t'/T)$, a consistent estimate of $\boldsymbol{\theta}$ is straightforward to compute.

Given any consistent estimate of $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$, and the associated weighting matrix $\boldsymbol{W}_T$, a GMM test can be constructed easily,

$$H_z \equiv T(\boldsymbol{\Psi}_T\boldsymbol{g}_T)'\boldsymbol{V}_T(\boldsymbol{\Psi}_T\boldsymbol{g}_T), \tag{11}$$

where $\boldsymbol{g}_T = \boldsymbol{g}_T(\tilde{\boldsymbol{\theta}})$, $\boldsymbol{V}_T$ is an $NL \times NL$ diagonal matrix: $\boldsymbol{V}_T = \mathrm{Diag}(1/v_1, \ldots, 1/v_d, 0, \ldots, 0)$, formed by $v_1 > \cdots > v_d > 0$ ($d = NL - q$), the positive eigenvalues of the following $NL \times NL$ semi-definite matrix:

$$\boldsymbol{\Pi}_T \equiv [\boldsymbol{I} - \boldsymbol{D}_T(\boldsymbol{D}_T'\boldsymbol{W}_T\boldsymbol{D}_T)^{-1}\boldsymbol{D}_T'\boldsymbol{W}_T]\boldsymbol{S}_T[\boldsymbol{I} - \boldsymbol{D}_T(\boldsymbol{D}_T'\boldsymbol{W}_T\boldsymbol{D}_T)^{-1}\boldsymbol{D}_T'\boldsymbol{W}_T]', \tag{12}$$

$\boldsymbol{\Psi}_T$ is an $NL \times NL$ matrix, of which the ith row is the standarized eigenvector corresponding to the ith largest eigenvalue of $\boldsymbol{\Pi}_T$ for $i = 1, \ldots, NL$; $\boldsymbol{D}_T = \boldsymbol{D}_T(\tilde{\boldsymbol{\theta}})$: an $NL \times q$ matrix of the first order derivatives of $\boldsymbol{g}_T = \boldsymbol{g}_T(\tilde{\boldsymbol{\theta}})$ with respect to the free parameters;[4] and $\boldsymbol{S}_T$ is the estimator of the underlying covariance structure of the model residuals given by (7). As shown by Zhou (1994), assuming the very general conditions of Hansen (1982), H_z is asymptotically χ^2 distributed with degrees of freedom $NL - q$. Based directly on the minimized quadratic form, Eq. (6), Jagannathan and Wang (1996) derive an alternative GMM test for an arbitrary weighting matrix. Their test has interesting economic interpretations in their applications, but is computationally more complex than ours. Since the economic interpretations of (6) are not obvious in our applications, we will use only the above GMM test in what follows. It should be noted that the above GMM test is a specification test of the null H_0 (Eq. (2)). A rejection of the moments conditions only leads to a rejection of the model. It is not a test of

[4] For the reader's convenience in implementing the computations, an explict formula for $\boldsymbol{D}_T$ is provided in Appendix A. In addition, a Fortran program and the data of the empirical studies of this paper will be available from the author through e-mail for a year after the publication of the paper.

a K-factor model versus a $(K+1)$ one. Nevertheless, as the test is analytically obtained, its empirical power against any given alternative hypothesis can be easily computed and assessed.

It should be noted that the estimator given by (9) is not the optimal GMM estimator. Its asymptotic covariance matrix is

$$\boldsymbol{\Theta}_T = (\boldsymbol{D}_T' \boldsymbol{W}_T \boldsymbol{D}_T)^{-1} \boldsymbol{D}_T' \boldsymbol{W}_T \boldsymbol{S}_T \boldsymbol{W}_T \boldsymbol{D}_T (\boldsymbol{D}_T' \boldsymbol{W}_T \boldsymbol{D}_T)^{-1}.$$

This follows straightforwardly from Hansen (1982). The expression is useful for computing the standard errors of the parameter estimates. For example, the standard error of the ith element of $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{A}}_2, \tilde{\boldsymbol{C}})$ is given by the square root of the ith diagonal element of $\boldsymbol{\Theta}_T/T$. However, one may also obtain the optimal GMM estimator by using an analytical iteration, $\tilde{\boldsymbol{\theta}}_{n+1} = \tilde{\boldsymbol{\theta}}_n - (\boldsymbol{D}_T' \boldsymbol{W}_T \boldsymbol{D}_T)^{-1} \boldsymbol{D}_T' \boldsymbol{W}_T \boldsymbol{g}_T(\tilde{\boldsymbol{\theta}}_n)$ where $\tilde{\boldsymbol{\theta}}_0 = \tilde{\boldsymbol{\theta}}$. As shown by Newey (1985), a one-step iteration of the consistent GMM estimator gives rise to an optimal GMM estimator with asymptotic covariance matrix $(\boldsymbol{D}_T' \boldsymbol{S}_T^{-1} \boldsymbol{D}_T)^{-1}$. Of course, a two or more step interations will also yield optimal GMM estimators, but iterating beyond the first step will not change the asymptotic distribution of the estimator even if continuing the process eventually will produce the traditional optimal GMM estimator (if the limit exists). Intuitively, small sample properties may be improved by iterating, but this can only be determined by further studies, say, through Monte Carlo experiments. Based on the optimal estimators, standard GMM tests which have a simpler form than (11) can be analytically obtained. However, as shown by Zhou (1994), such tests based on the one or two step iterations are not as reliable as H_z in finite sample sizes. Hence, we will use only H_z in what follows.

Now consider the case where the alphas are unconstrained. The idea is to solve first for the estimator of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{C}$, and then solve for $\boldsymbol{A}$ and $\boldsymbol{C}$. It can be shown that the GMM estimator of $\boldsymbol{A}$ and $\boldsymbol{C}$ will have the same formulas as (9) except that the previous $\boldsymbol{P}$ is now replaced by $\boldsymbol{P}^* = \boldsymbol{P} - \boldsymbol{P}\boldsymbol{1}_T (\boldsymbol{1}_T' \boldsymbol{P} \boldsymbol{1}_T)^{-1} \boldsymbol{1}_T' \boldsymbol{P}$, and the estimator of $\boldsymbol{\alpha}'$ is $\tilde{\boldsymbol{\alpha}}' = (\boldsymbol{1}_T' \boldsymbol{P} \boldsymbol{1}_T)^{-1} \boldsymbol{1}_T' \boldsymbol{P}(\boldsymbol{R} - \boldsymbol{X}\boldsymbol{B}')$, where $\boldsymbol{1}_T$ is a T-vector of ones. The GMM test also has the same form as the previous one with $\boldsymbol{P}$ replaced by $\boldsymbol{P}^*$. However, as there are now N additional parameters, the degrees of freedom of the asymptotic χ^2 distribution should be adjusted down by N.

There is an alternative approach that applies to the case where the alphas are unconstrained. It can be shown that the alphas are zero for a de-meaned version of (5), i.e., a new model of form (5) with the asset returns and factors subtracting out their time series means. Clearly all asymptotic properties of the parameter estimates and tests will be unchanged with the de-meaning. Optimal estimates of the betas can be obtained from a regression of the de-meaned returns on the de-meaned factors. Because of this, the procedure used for the first case applies to the second case as well. The estimates of the alphas are then given by the asset means minus the product of the betas with the means of the pre-specified factors.

2.3. Extension to multiple sets of variables

To motivate, consider the case where we have two sets of pre-specified factors, $\boldsymbol{X}_{1t}$ of dimension m_1 and $\boldsymbol{X}_{2t}$ of dimension m_2, so that we have in total $m_1 + m_2 = M$ variables. Suppose that we are interested in a two-factor model. Our previous procedure allows one to extract two factors from the best linear combinations of both $\boldsymbol{X}_{1t}$ and $\boldsymbol{X}_{2t}$. Assume now that variables in $\boldsymbol{X}_{1t}$ measure the macroeconomic output risk, and those in $\boldsymbol{X}_{2t}$ measure the interest rate risk. For easier economic interpretations, it may be desirable to extract the first factor from the best linear combination of variables in $\boldsymbol{X}_{1t}$, and the second factor from the best linear combination of variables in $\boldsymbol{X}_{2t}$. Hence, as an extension of (1) or (5), we have now a multivariate regression of asset returns on two sets of economic variables,

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + \boldsymbol{B}_1\boldsymbol{X}_{1t} + \boldsymbol{B}_2\boldsymbol{X}_{2t} + \boldsymbol{U}_t, \tag{13}$$

where there are two rank restrictions, one on $\boldsymbol{B}_1$, an $N \times m_1$ matrix, and another one on $\boldsymbol{B}_2$, an $N \times m_2$ matrix. As there is only one factor to be extracted out of $\boldsymbol{X}_{1t}$ and $\boldsymbol{X}_{2t}$ separately, the rank of both $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are restricted to be one.

Brown et al. (1997) propose an idea of forming factors by equal-weighting variables out of each set individually from a few given sets of portfolios. The above extension compliments this idea by allowing the weights be not necessarily equal, but chosen to maximize the explanatory power of the model. In general, there may be p sets of observable economic variables, $\boldsymbol{X}_{1t}, \dots, \boldsymbol{X}_{pt}$, each of which has m_i variables that have related or similar economic interpretations. There are in total $m_1 + \cdots + m_p = M$ observables. To keep the same economic interpretations, it is of interest now to extract K_i factors ($K_i \leqslant m_i$) from each set of them, resulting in a total of $K_1 + \cdots + K_p = K$ factors. In this case, there are p rank restrictions on the coefficient matrices of the multivariate asset returns regression,

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + \boldsymbol{B}_1\boldsymbol{X}_{1t} + \cdots + \boldsymbol{B}_p\boldsymbol{X}_{pt} + \boldsymbol{U}_t, \tag{14}$$

where $\boldsymbol{R}_t$ is an N-vector of asset returns, $\boldsymbol{B}_i$ is an $N \times m_i$ matrix of regression coefficients and $\boldsymbol{U}_t$ is the model residual. The rank restrictions state that the rank of $\boldsymbol{B}_i$ is K_i for $i = 1, \dots, p$. Parameter estimation and tests for this general model is unfortunately complex. Unlike the case of a single rank restriction, analytical solutions to the GMM estimation and tests are no longer available. Nevertheless, we in what follows provide formulas (see Appendix B for proofs) so that the estimators and tests can be obtained by using analytical iterations, making numerical solutions feasible in practice. In contrast, numerical optimization routines may be very difficult to implement directly as there are now too many parameters (easily over one hundred in common empirical applications).

Under the multiple rank restrictions, we can write $\boldsymbol{B}_i$ as $\boldsymbol{B}_i = \boldsymbol{A}_i\boldsymbol{C}_i$, where $\boldsymbol{A}_i$ and $\boldsymbol{C}_i$ are $N \times K_i$ and $K_i \times m_i$ matrices, respectively. Conditional on C_i

($i = 1, \dots, p$), the GMM estimators of $\boldsymbol{\alpha}$ and the A's are obtained from

$$\boldsymbol{a} = (\boldsymbol{X}^{**\prime}\boldsymbol{W}_T\boldsymbol{X}^{**})^{-1}\boldsymbol{X}^{**\prime}\boldsymbol{W}_T\boldsymbol{y}, \tag{15}$$

where $\boldsymbol{a} = \mathrm{vec}[\boldsymbol{\alpha}, \boldsymbol{A}_1, \dots, \boldsymbol{A}_p]'$, $\boldsymbol{X}^{**} = \boldsymbol{I}_N \otimes \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Z}_t(1, \boldsymbol{X}'_{1t}\boldsymbol{C}'_1, \dots, \boldsymbol{X}'_{pt}\boldsymbol{C}'_p)'$, $\boldsymbol{y} = (1/T)\sum_{t=1}^{T}\boldsymbol{R}_t \otimes \boldsymbol{Z}_t$, and, as in Section 2.2, $\boldsymbol{W}_T$ and $\boldsymbol{Z}_t$ are the GMM weighting matrix and instruments. On the other hand, conditional on $\boldsymbol{\alpha}$ and the A's, the C's are given by

$$\boldsymbol{c} = (\tilde{\boldsymbol{X}}'\boldsymbol{W}_T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{W}_T\tilde{\boldsymbol{y}}, \tag{16}$$

where $\boldsymbol{c} = [(\mathrm{vec}\boldsymbol{C}_1)', \dots, (\mathrm{vec}\boldsymbol{C}_p)']'$, $\tilde{\boldsymbol{X}} = (1/T)\sum_{t=1}^{T}(\boldsymbol{X}'_{1t}\otimes\boldsymbol{A}_1, \dots, \boldsymbol{X}'_{pt}\otimes\boldsymbol{A}_p)\otimes\boldsymbol{Z}_t$ and $\tilde{\boldsymbol{y}} = (1/T)\sum_{t=1}^{T}(\boldsymbol{R}_t - \boldsymbol{\alpha})\otimes\boldsymbol{Z}_t$. Analytical iterations between (15) and (16) give rise to a series of estimators, whose limit is the GMM estimator of the underlying parameters. However, as in the single rank restriction case, the estimators will not be unique unless certain identification conditions are imposed. Following the earlier discussions, one can easily impose any normalization conditions, such as $\boldsymbol{C}'_i\boldsymbol{C}_i = \boldsymbol{I}_{K_i}$ to uniquely identify the parameters (in the present case, imposing normalization conditions on $\boldsymbol{C}$ is simpler than doing so on $\boldsymbol{A}$). In the special case where $K_i = m_i$ for some i, i.e., no rank restrictions are imposed on the ith set of observables, the above estimation can clearly be simplified. In this case, the C_i in (15) can be taken as the identity matrix, and that in (16) can be deleted. But $\tilde{\boldsymbol{y}}$ needs be adjusted by replacing $(\boldsymbol{R}_t - \boldsymbol{\alpha})$ with $(\boldsymbol{R}_t - \boldsymbol{\alpha} - \boldsymbol{B}_i\boldsymbol{X}_{it})$. If there are no rank restrictions on several sets or all of them, similar simplifications obviously hold. Hence, with the provided analytical iteration formulas, it is an easy matter to program to obtain the parameter estimates. Then, with the estimates, it is straightforward to obtain the standard chi-squared tests for the various rank restrictions of the model as the optimal weighting matrix can be used in the analytical iterations directly.

3. The first school of thought

In this section, we explain and model formally the idea of the first school of thought. In Section 3.1, we illustrate that the first school of thought can yield the same econometric model as the second school, although they are conceptually different. Then, in Section 3.2, we discuss the implications for testing the CAPM. Finally, in Section 3.3, we show how similar ideas can be applied to extracting the latent factors of the APT model,[5] and to transforming the APT asset pricing restrictions into testable rank restrictions. As a result, we provide a new test for the APT that naturally links the latent factors to economic fundamentals. This

[5] Independently, Bakshi and Chen (1996) provide an interesting way of extracting latent factors from equilibrium models of asset prices.

test allows for robust distributional assumptions and is free from the errors-in-variables problem present in many existing studies.

3.1. An equivalent econometric model

The first school of thought regards the factors as inherently latent. For simplicity, consider the simple one factor case where we have a one-factor model for asset returns,

$$R_{it} = \alpha_i + \beta_i f_t + \varepsilon_{it}, \quad i = 1, \dots, N, \tag{17}$$

where R_{it} is the return of asset i in excess of the risk-free rate, f_t is the latent factor and ε_{it} is the model residual with zero mean. Without loss of generality, f_t is usually assumed to have zero mean with unit variance, but its realizations are *unobservable* directly from market data.

As the factor is unknown, Connor and Korajczyk's (1986) asymptotic principal components approach and the standard factor analysis approach (see, e.g, Seber, 1984) may be used to extract it from Eq. (17). But, as pointed earlier, it is difficult to relate the estimated factor to $X_1, \dots, X_M$, a set of M given observable economic or financial variables which have zero expected values. To overcome this difficulty, we can always *project* f_t onto $X_1, \dots, X_M$ to obtain

$$f_t = C_1 X_{1t} + \cdots + C_M X_{Mt} + v_t, \tag{18}$$

where v_t is the projection error which is uncorrelated with $X_1, \dots, X_M$. Assume ε_{it} is also uncorrelated with $X_1, \dots, X_M$, i.e., $\mathrm{E}[X_m \varepsilon_{it}] = 0$. This is not so restrictive because ε_{it} measures the idiosyncratic risk, whereas both f_t and the observables are pervasive risks that are common to all assets. The condition $\mathrm{E}[X_m \varepsilon_{it}] = 0$ will be used later in the GMM estimation of (17) or Eq. (19) below.

The advantage of the projection formulation, Eq. (18), is that the latent factors are modelled explicitly as a linear function of observables plus a noise.[6] By estimating the parameters (C_1 through C_M), we know the contributions of the known variables to the latent factor. Furthermore, both estimation and asset pricing tests can be carried out in one-step so that there will no longer exist the errors-in-variables problem.

Combining Eq. (18) with (17), we get a regression of the asset returns on the observables, $R_{it} = \alpha_i + \beta_i (C_1 X_{1t} + \cdots + C_M X_{Mt}) + u_{it}$, where $u_{it} = \beta_i v_t + \varepsilon_{it}$. In vector form, this can be written as

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{X}_t + \boldsymbol{u}_t, \tag{19}$$

[6] In the latent variable literature (see, e.g., Gibbons and Ferson, 1985), the latent variables are expected returns associated with factors, not the factors themselves.

where $\boldsymbol{R}_t$ is an N-vector of the returns, $\boldsymbol{\alpha}, \boldsymbol{X}_t$ and $\boldsymbol{u}_t$ are defined similarly, and $\boldsymbol{B}$ is an $N \times M$ matrix of regression coefficients,

$$\boldsymbol{B} = \boldsymbol{\beta}\boldsymbol{C}', \tag{20}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_N)'$, $N \times 1$, and $\boldsymbol{C} = (C_1, \dots, C_M)'$, $M \times 1$. It is seen that the multivariate regression (19) cannot be arbitrary, but with the regression coefficients matrix restricted to be of rank one in the form of (20).

In comparison (19) with (5), here we have the same multivariate regression models for the asset returns. Furthermore, both have rank restrictions on $\boldsymbol{B}$. Hence, the GMM estimation and tests can be done by using exactly the same procedures as described in Section 2.2. However, Eqs. (19) and (5) are conceptually different. Eq. (5) is a model assumed at the outset to be true. The objective there is to get the minimum number of factors necessary from the pre-specified ones. In contrast, Eq. (19) is a model derived from (17) and (18). In Eq. (17), one views the factors as latent and has a prior belief on the number of latent factors, whereas Eq. (18) simply links the latent factors to the observables. For example, if one really believes in a one-factor model (motivated perhaps by the CAPM), then one is only interested in testing (19) for the rank one restriction. But for those of the second school of thought, they may test (5) for rank restrictions $(M-1)$, $(M-2)$, etc., to find the minimum number of factors necessary.

Jöreskog and Goldberger (1975) seem the first to estimate a model of form (17) and (18) by using the maximum likelihood approach. Their approach is, however, complex and difficult to implement numerically. In contrast, our approach based on the generalized method of moments (GMM) is easy to apply because both the estimation and tests can be solved analytically for both the above one-factor model and for a general K-factor one (see Section 2.2). The benefits of the analytical solutions are especially important when there are many assets and many factors. In contrast, the estimation is difficult even in the one-factor case by using the maximum likelihood approach, and almost impossible in the K-factor case. However, the GMM method is asymptotically less efficient than the maximum likelihood approach, but more robust to the underlying distributions, such as non-normality.

3.2. *Implications for testing the CAPM*

As a special case of the one-factor model (18), the usual market model regression states that

$$R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it}, \quad i = 1, \dots, N, \tag{21}$$

where $R_{mt} = r_{mt} - r_{ft}$ is the return on the market portfolio in excess of the risk-free rate. It is well-known that a test of the CAPM is a test of whether or not the intercepts are zero. The difficulty is that return on the market portfolio, r_{mt}, is unobservable. As a result, most studies assume that r_{mt} is some pre-specified stock index.[7] Clearly, the pre-specified index may not be the market portfolio exactly, and an errors-in-variables problem is introduced. Based on our analysis here, we may project $r_{mt} - r_{ft}$ onto a few known economic variables, including stock indices, in the same way as in Eq. (18). Then, a test of the CAPM becomes a test of whether or not the intercepts in (19) are zero. This can be easily carried out by using the GMM test provided in Section 2.2.

An alternative method of testing the CAPM is as follows. As an extension to the practice of using a proxy, such as the value-weighted or equal-weighted stock index, to replace r_{mt}, we assume that the market portfolio is a portfolio of the two,[8]

$$r_{mt} = \gamma_1 r_{vt} + \gamma_2 r_{et}, \tag{22}$$

where r_{vt} and r_{et} are returns on the value-weighted or equal-weighted stock indices, and γ_1 and γ_2 are portfolio weights such that $\gamma_1 + \gamma_2 = 1$. Intuitively, a portfolio of the two indices should better mimic the unknown market portfolio than using either of them individually. In addition, the equal-weighted index may have more to do with firm's capital size, and hence the portfolio may capture some of the size effects found in empirical studies. Stambaugh (1982) provides many additional economic indices, such as real estate and consumer durables, for the composition of the market portfolio. While his weights of the various indices in the market portfolio are chosen based on economic intuition and suffer from some arbitrariness, an assumption (or a projection) like (22) by using those indices provides a formal way of estimating the weights from the data.

With the market portfolio generated from (22), the market model can be written as $R_{it} = \alpha_i + \beta_i(\gamma_1 r_{vt} + \gamma_2 r_{et} - r_{ft}) + \varepsilon_{it}$. Clearly, a test of the CAPM is still a test of whether or not the intercepts are zero. In other words, if the CAPM is true, a multivariate regression of the excess returns on $r_{vt} - r_{et}$ and $r_{et} - r_{ft}$,

$$R_{it} = \alpha_i + \beta_{i1}(r_{vt} - r_{ft}) + \beta_{i2}(r_{et} - r_{ft}) + u_{it}, \tag{23}$$

must have zero intercepts and the rank of the matrix formed by the betas must be one. Hence, the earlier GMM estimation and test procedures are straightforwardly adapted to test for these CAPM restrictions.[9]

[7] After Roll's (1977) forceful arguments for the unobservability of the market portfolio, a test of the CAPM becomes a test for efficiency of a given portfolio. Shanken (1996) provides an excellent survey of recent developments.

[8] An assumption like $r_{mt} = \gamma_1 r_{vt} + \gamma_2 r_{et} + v_t$, may also be used, where v_t is an error term with zero mean conditional on the indices. But this will neither affect the point estimate of γ_1 and γ_2, nor the way the GMM test is computed.

[9] Nardari and Zhou (1999) provides a detailed study for the decomposition of the market portfolio.

3.3. Implications for testing the APT

In the APT, it is usually assumed that the returns on a vector of N assets are related to K pervasive and latent factors by a K-factor model:

$$r_{it} = \mathrm{E}[r_{it}] + \beta_{i1} f_{1t} + \cdots + \beta_{iK} f_{Kt} + \varepsilon_{it}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{24}$$

where r_{it} is the return on asset i at time t, f_{kt} the kth pervasive factor at time t, ε_{it} the idiosyncratic factor of asset i at time t, β_{ik} the beta or factor loading of the kth factor for asset i, N the number of assets, and T the number of periods.

In what follows, it will be convenient for us to work with the vector form of the model:

$$\boldsymbol{r}_t = \mathrm{E}[\boldsymbol{r}_t] + \boldsymbol{\beta} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{25}$$

where $\boldsymbol{r}_t$ is an $N \times 1$ vector of returns, $\mathrm{E}[\boldsymbol{r}_t]$ is the expected returns conditional on $\boldsymbol{f}_t$, $\boldsymbol{\beta}$, $N \times K$, $\boldsymbol{f}_t$ and $\boldsymbol{\varepsilon}_t$ are defined accordingly. The standard assumptions on the factor model are:

$$\mathrm{E}[\boldsymbol{f}_t] = \boldsymbol{0}, \quad \mathrm{E}[\boldsymbol{f}_t \boldsymbol{f}_t'] = \boldsymbol{I}, \quad \mathrm{E}[\boldsymbol{\varepsilon}_t | \boldsymbol{f}_t] = \boldsymbol{0}, \quad \mathrm{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' | \boldsymbol{f}_t] = \boldsymbol{\Sigma}. \tag{26}$$

With the factor model, the classic APT is derived under the assumption that the residual covariance matrix $\boldsymbol{\Sigma}$ is diagonal, but subsequent studies replace this assumption by a much weaker one (see, e.g., Shanken, 1992). Consistent with these developments, our methodology below does not require $\boldsymbol{\Sigma}$ be diagonal as long as it is positive definite.

Consider the following restrictions of the APT:

$$\mathrm{E}[\boldsymbol{r}_t | \boldsymbol{f}_t] = r_{ft} \boldsymbol{1}_N + \boldsymbol{\beta} \boldsymbol{\lambda}, \tag{27}$$

where r_{ft} is the risk-free rate of return, $\boldsymbol{1}_N$ is an N-vector of ones and $\boldsymbol{\lambda}$ is a K-vector of the risk premiums. Eq. (27) is the implication, for example, of the equilibrium version of the APT (Connor, 1984). The factor model (25) and the APT restrictions (27) imply that

$$\boldsymbol{r}_t - r_{ft} \boldsymbol{1}_N = \boldsymbol{\beta}(\boldsymbol{f}_t + \boldsymbol{\lambda}) + \boldsymbol{\varepsilon}_t. \tag{28}$$

Since the factors are unobservable, neither are $\boldsymbol{f}_t + \boldsymbol{\lambda}$. Because of this, most of the existing procedures (e.g., Lehmann and Modest, 1998, and Connor and Korajczyk, 1995) estimate $\boldsymbol{f}_t + \boldsymbol{\lambda}$ first, and then a regression of the excess returns on the estimated factors is run to test whether the intercepts are zero. As the estimated rather than the actual factors are used in the regression, the errors-in-variables problem is introduced. To overcome this problem, we project the latent factors, $\boldsymbol{f}_t + \boldsymbol{\lambda}$, onto M economic variables $\boldsymbol{X}_t$,

$$\boldsymbol{f}_t + \boldsymbol{\lambda} = \boldsymbol{\mu} + \boldsymbol{\phi} \boldsymbol{X}_t + \boldsymbol{v}_t. \tag{29}$$

This states that the factors (plus risk premiums) are a linear function of $\boldsymbol{X}_t$ plus a random noise $\boldsymbol{v}_t$. In other words, the factors are forecasted by using $\boldsymbol{X}_t$. The $\boldsymbol{v}_t$ term represents the forecasting error having zero expectation conditional on $\boldsymbol{X}_t$.

Combining the factor forecasting Eq. (29) with (28), we obtain a regression of the excess returns on the economic variables,

$$\boldsymbol{r}_t - r_{ft}\mathbf{1}_N = \boldsymbol{\beta\mu} + \boldsymbol{\beta\phi X}_t + \boldsymbol{U}_t, \tag{30}$$

where $\boldsymbol{U}_t = \boldsymbol{\beta v}_t + \boldsymbol{\varepsilon}_t$. Let $\boldsymbol{R}_t$ be an N-vector of the returns in excess of the risk-free rate. In comparison (30) with the following unrestricted multivariate regression of the excess returns on the economic variables,

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + \boldsymbol{B X}_t + \boldsymbol{U}_t, \tag{31}$$

the K-factor APT implies that there exist a K-vector $\boldsymbol{\mu}$, an $N \times K$ matrix $\boldsymbol{\beta}$ and a $K \times M$ matrix $\boldsymbol{\phi}$ such that

$$\mathrm{H}_0\colon \quad \boldsymbol{\alpha} = \boldsymbol{\beta\mu}, \quad \boldsymbol{B} = \boldsymbol{\beta\phi}. \tag{32}$$

Hence, a test of H_0 in Eq. (31) is a test of the K-factor APT.

Intuitively, if there are truly K factors, a regression run on M ($M > K$) variables should reveal how these variables contribute to the K unknown factors, and the contributions are summarized here by the linear combinations in Eq. (30). Though based on a factor model, our analysis here differs from the standard factor analysis in many important ways. It ties the factors directly to the economic variables whose observations play a direct role in the estimates of the factors. In contrast, the standard factor analysis uses only the returns data to arrive at its factor estimates. In terms of hypothesis testing, one tests in a factor analysis the null hypothesis of a K-factor model against an unrestricted model since a ($K + 1$)-factor model is unidentified under the null of a K-factor model, whereas our framework has restrictions on the projection coefficients which are identified under both the null and the ($K + 1$)-factor alternative.

Econometrically, there are four important aspects worth noting. First, it should be pointed out that we need moment conditions $\mathrm{E}[\boldsymbol{U}_t \otimes \boldsymbol{Z}_t] = \mathbf{0}$ to estimate (31) and test (32). To guarantee (26) and (29), we may have to assume that the asset returns and factors be jointly i.i.d. normally distributed. However, the normality assumption may be relaxed to allow for only i.i.d. assumption because $\mathrm{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' | \boldsymbol{f}_t] = \boldsymbol{\Sigma}$ in (26) can potentially depend on $\boldsymbol{f}_t$ as long as the APT theoretical conclusion (27) holds. In this case, we must assume $\mathrm{E}[\boldsymbol{\varepsilon}_t \boldsymbol{X}_t] = \mathbf{0}$ to guarantee $\mathrm{E}[\boldsymbol{U}_t \otimes \boldsymbol{Z}_t] = \mathbf{0}$ with $\boldsymbol{Z}_t = (1, \boldsymbol{X}_t')$. The condition $\mathrm{E}[\boldsymbol{\varepsilon}_t \boldsymbol{X}_t] = \mathbf{0}$ does not seem to be too restrictive because $\mathrm{E}[\boldsymbol{\varepsilon}_t | \boldsymbol{f}_t] = \mathbf{0}$ and both $\boldsymbol{f}_t$ and $\boldsymbol{X}_t$ are pervasive variables.

Second, as usual, the factor loadings are not uniquely identified. However, various identification conditions can be easily imposed to obtain unique loadings. Third, the parameter constraints can be transformed into the previous rank restrictions on the regression parameter matrix. Indeed, (31) can be written as $\boldsymbol{R}_t = \boldsymbol{B}^*\boldsymbol{X}_t^* + \boldsymbol{U}_t$, where $\boldsymbol{B}^* \equiv [\boldsymbol{\alpha}, \boldsymbol{B}]$ and $\boldsymbol{X}_t^* = (1, \boldsymbol{X}_t')'$. Then, it is easy to verify that constraints (32) are equivalent to the rank restriction that $\mathrm{rank}(\boldsymbol{B}^*) = K$. For example, in a one-factor model, this imposes a rank one restriction on the N by $(M + 1)$ matrix of the regression coefficients $\boldsymbol{B}^*$. Hence, all the GMM estimation and testing procedures of this paper can be unified in the same framework provided earlier in Section 2.2. Finally, the tests here and those of the latent variable approach are very similar because they all end up testing rank restrictions (Velu and Zhou, 1999, provide related tests). However, there are important differences in both the assumed return generating processes and the interpretation of the parameters. For example, the return generating process of the latent variable approach is a conditional forecasting equation that restricts the instruments $\boldsymbol{Z}_t$ be the information which investors use to forecast returns. More importantly, the latent factors themselves cannot be identified with the latent variable approach, while the framework here shows precisely how the economic variables determine the latent factors.

4. Applications to Chen, Roll and Ross's model

In this section, we apply our methodology to address two questions. First, given the five economic variables of Chen et al. (1986), we ask how many linear combinations of the economic variables are needed to explain security returns in the standard K-factor model. Second, given the linear combinations of the economic variables that best forecast the latent factors (plus risk premiums), we examine which of the economic variables contribute most to the linear combinations.

4.1. The data

The security returns data are the returns on industry portfolios grouped by following Sharpe (1964), Breeden et al. (1989), Gibbons et al. (1989) and Ferson and Harvey (1991) with raw data available from the Center for Research in Security Prices (CRSP) at the University of Chicago. There are twelve industries: petroleum, finance/real estate, consumer durables, basic industries, food/tobacco, construction, capital goods, transportation, utilities, textiles/trade, services and leisure. With these industry portfolio returns, the excess asset returns are easily computed as those in excess of the 30-day Treasury bill rate available from Ibbotson Associates. The returns are monthly data from January 1953 to December 1989, a sample size of $T = 444$. Although the sample size of the security returns available is much larger than 444, it is limited to 444 due to the availability of the Chen, Roll and Ross's (CRR) macroeconomic variables.

The five CRR macroeconomic variables are IP = industrial production, UI = unanticipated inflation, DI = change in expected inflation, DF = default risk (measured as the difference between the monthly returns on corporate bonds and long-term government bonds) and MT = term premium (measured as the difference between the monthly returns on 30 day T-bills and long-term government bonds). These are clearly the most widely used and collected economic variables. They appear to contain information on the realization of the latent factors in the APT that are the underlying economic risks affecting security returns. The macroeconomic variables are monthly observations, and are exactly those used by Epps and Kramer (1995) in which a full description of the data is provided.[10]

4.2. Extracted factors

At time t, observation on the five CRR macroeconomic variables consists of $\boldsymbol{X}_t$. As discussed earlier in Section 3, the instrumental variables $\boldsymbol{Z}_t$ may be chosen as $\{\mathbf{1}_T, \boldsymbol{X}_t\}$. Given the data and the instruments, the parameter estimate can be obtained first by using the identity weighting matrix. Then, a new weighting matrix, the inverse of $(\sum_{t=1}^{T} \boldsymbol{U}_t\boldsymbol{U}_t'/T)\otimes(\sum_{t=1}^{T} \boldsymbol{Z}_t\boldsymbol{Z}_t'/T)$ which is the optimal under i.i.d. assumption, is computed and used to obtain a new parameter estimate. Based on this estimate and the associated weighting matrix, the GMM test H_z is straightforward to compute, as described in detail in Section 2.2.

On the question as to the minimum number of linear combinations of the economic variables needed to explain security returns in the factor model, Panel A of Table 1 reports the results. The question is the same as testing for the number of factors in the APT model. Under the null hypothesis that one linear combination is sufficient or there is a one-factor model, the GMM test statistic is 82.3992 and the asymptotic P-value, based on the chi-squared distribution, is 0.0098. This suggests rejection of the null at the usual 5% significance level. However, for a two-factor model specification, the GMM test statistic is 82.3992 and the P-value is 0.4254. We can no longer reject the null, suggesting that, given a universe of the five CRR economic variables, two linear combinations of them are sufficient to explain the returns.

Now, given one or two of the linear combinations of the economic variables that best forecast the latent factors, it is of interest to know which of the economic variables contribute most to the linear combinations. Panel B of Table 1 reports the coefficients, $\boldsymbol{\phi}$, of the linear combinations and the associated standard errors. In $K = 1$ case, there is only one linear combination. The coefficients on the five CRR macroeconomic variables (industrial production, unanticipated inflation, change in expected inflation, default risk and term

[10] The author is grateful to T.W. Epps and C.F. Kramer for permission and forward of their data.

Table 1
Tests of the APT based on the CRR factors

Panel A: *Tests*

Number of factors	Test statistic	P-value
$K = 1$	82.3992	0.0098
$K = 2$	41.0238	0.4254

Panel B: *Factor compositions*

Number of factors	IP	UI	DI	DF	MT
$K = 1$	− 0.1238	− 0.1501	− 8.3789	0.1601	0.0358
	(0.0485)	(0.4239)	(3.1068)	(0.1099)	(0.0506)
$K = 2$	− 0.1137	− 0.0683	− 7.6110	0.2091	0.1032
	(0.0546)	(0.4690)	(3.5059)	(0.1514)	(0.1249)
	− 0.0865	0.1789	− 5.818	0.3542	0.2923
	(0.0556)	(0.5448)	(3.1652)	(0.1616)	(0.1092)

Panel A of the table reports both the test statistic based on the generalized method of moments (GMM) and the associated asympototic P-value, and Panel B reports the coefficients, ϕ, of the linear combinations and the associated standard errors (in the brackets).

premium) are − 0.1238, − 0.1501, − 8.3789, 0.1601 and 0.0358, respectively. The associated standard errors are 0.0485, 0.4239, 3.1068, 0.1099 and 0.0506. Clearly, industrial production and change in expected inflation are important in the linear combination, whereas unanticipated inflation, default risk and term premium have no statistically significant, incremental contribution. In the two-factor model case, there are two linear combinations of the CRR macroeconomic variables. The two sets of coefficients are − 0.1137, − 0.0683, − 7.6110, 0.2091, 0.1032 and − 0.0865, 0.1789, − 5.818, 0.3542, 0.2923, respectively. The associated standard errors are 0.0546, 0.4690, 3.5059, 0.1514, 0.1249 and 0.0556, 0.5448, 3.1652, 0.1616, 0.1092. In analyzing the standard errors, it is interesting to observe that it is still the case that a linear combination of industrial production and change in expected inflation determines the first factor. However, a linear combination of default risk and term premium contributes to the second factor.

As a comparison with the above two-factor model, we also extract two factors by using the standard factor analysis approach (see, e.g, Seber, 1984).[11] Denote by f_1 and f_2 the factors extracted based on the CRR economic variables, and by

[11] Details of the factor analysis results are available upon request.

Table 2
Correlations and risk premiums

	f_1^*	f_2^*
Panel A: Correlations		
f_1	0.1660	0.2665
f_2	− 0.2570	− 0.3329
Q_c	γ_1	γ_2
Panel B: Risk Premiums for f_1 and f_2		
0.0226	0.0110	0.0144
(0.3881)	(0.2169)	(0.1575)
Panel C: Risk Premiums for f_1^* and f_2^*		
0.0786	0.0605	− 0.0676
(0.0002)	(0.0879)	(0.0625)

Panel A of the table reports the correlations of the extracted factors based on the CRR economic variables (f_1 and f_2) and the factors from a standard factor analysis (f_1^* and f_2^*). Panel B reports the test, Q_c, for the asset pricing restrictions and the risk premiums for factors f_1 and f_2, and Panel C does the same for f_1^* and f_2^* (the P-values and standard errors are in the brackets).

f_1^* and f_2^* the factors extracted from the standard factor analysis. Panel A of Table 2 reports the correlations among the factors. For example, the correlation between f_1 and f_1^* is 0.1660. In general, the correlations are not too strong. The reason is that the factors are extracted from two theoretically totally different models. In extracting f_1 and f_2, we assume both (28) and (29) hold exactly, implying that f_1^* and f_2^* are extracted from a mis-specified model. Intuitively, if the true factors are truly a simple function of the given economic variables, there is no guarantee that f_1^* and f_2^* should be close to f_1 and f_2. On the other hand, if the standard factor model is true and if Eq. (29) is incorrect, we cannot expect f_1 and f_2 are good factors. Hence, in comparison f_1 and f_2 with f_1^* and f_2^*, we should not expect close relationships because they are based on different model specifications. Although theoretically it is difficult to determine which of the factors are better ones, one can always examine their empirical performances. Generally, this will depend on the criteria used. As explaining the cross-section of expected stock returns is of great interest in recent studies (see Fama and French, 1993; Jagannathan and Wang, 1996 and references therein), we use the standard OLS two-pass regressions to estimate the risk premiums, γ_1 and γ_2, associated with the factors. The results are reported in Panels B and C of Table 2. It is seen that the numerical values are sizably different, a fact reflecting the differences in the factors.

Because of differences in the factors, there is not much basis for model choice by using information on the magnitude of the risk premiums. However, it is interesting and meaningful to examine the cross-sectional pricing errors associated with the factors. Shanken's (1985) Q_c statistic (which extends straightforwardly from his one-factor case to the present multi-factor one) provides both a measure of the cross-sectional pricing errors and a test for the asset pricing restrictions. The smaller the Q_c, the smaller the pricing errors. As Q_c are 0.0226 and 0.0786, respectively, we know that f_1 and f_2 perform better in explaining the cross-section of expected stock returns than f_1^* and f_2^*. This is also confirmed by the P-values, which are 38.81% and 0.02%, respectively, for the two sets of extracted factors.

Our results show that there are essentially two factors that are sufficient to explain the asset returns out of Chen et al.'s (1986) original five. However, the number of factors itself seems an illusive concept because it depends on the context and on the particular application. For example, theoretically, Hansen and Jagannathan (1997) show a single projection portfolio can price all of the assets under consideration. Similarly, as pointed out by Roll (1977), any efficient portfolio can also price a given set of assets in the single beta model. However, neither the projection portfolio nor the efficient portfolio is observable directly from the market data. Hence, empirically, one may have to use more than one factors to estimate the parameters and price the assets. For example, in a truly two-factor APT, it is known that certain transformations imply that only one-factor is priced (see, e.g., Ingersoll, 1987, p. 175), which seems also be supported by the empirical results of Geweke and Zhou (1996), but both of the factors have to be used *jointly* to estimate the parameters of the two-factor APT model. If only one factor were used, the model would be mis-specified, and hence the parameter estimates might be biased and inaccurate. On the other hand, as shown by Jagannathan and Wang (1996), a conditional one-factor model can imply an unconditional model with three factors, and all of which are necessary in explaining the asset returns used there. In comparison, we show here that two factors are sufficient. The differences lie in the number of assets used, the observable economic variables under consideration, and the particular asset pricing model analyzed. Hence, empirical findings on the number of factors and factor compositions may vary, depending on the specific set-up and data.

5. Applications to Fama and French's model

Fama and French (1993) identify five common risk factors to explain the average returns on 25 stock portfolios formed on size and book-to-market (M/B). The factors are an overall market factor, *MKT*, as represented by the excess returns on the weighted market portfolio; a size factor, *SMB*, as represented by the return on small stocks minus those on large stocks; a book-to-market factor, *HML*, as represented by the return on high M/B stocks minus those on low B/M stocks; a term premium factor, *TERM*, as represented by the return on

Table 3
Optimal linear combinations of Fama–French factors

Number of factors	Test statistic	P-value
$K = 1$	198.704800	0.000011
$K = 2$	157.722445	0.000044
$K = 3$	108.042303	0.001880
$K = 4$	77.749437	0.002370
$K = 5$	56.580360	0.000304

The table reports both the test statistic based on the generalized method of moments (GMM) and the associated asympototic P-value for the number of optimal linear combinations in the Fama–French factor model.

long-term government bond minus those on 30-day T-bill; and a default risk factor, *DEF*, as represented by the return on corporate bond minus those on long-term government bond.[12] To explain the average returns on the 25 stock portfolios, a regression of the portfolio returns in excess of the 30-day T-bill rate is run on the factors,

$$r_i = \alpha_i + \beta_{i1} f_{MKT} + \beta_{i2} f_{SMB} + \beta_{i3} f_{HML} + \beta_{i4} f_{TERM} + \beta_{i5} f_{DEF} + \varepsilon_{it}. \quad (33)$$

As it is not clear whether a four- or three-factor model is already appropriate, we test a K-factor model,

$$r_i = \alpha_i + b_{i1} f_1^* + \cdots + b_{iK} f_K^* + \varepsilon_{it}, \quad (34)$$

for $K = 1, \ldots, 5$, where $f_1^*, \ldots, f_K^*$ are linear combinations of the original five Fama–French factors that best explain the returns. Given our earlier discussion in Section 3, the analytical GMM tests are straightforward to carry out.

Table 3 reports the GMM test statistics and the associated P-values. When only one linear combination is considered ($K = 1$), the test statistic is 198.704800 with a rather small P-value of 0.000011. The highest P-value is achieved at 0.002370 when there are three factors ($K = 3$). At the conventional 5% significance level, all of the factor models, $K = 1$ through 5, are rejected. Notice that the P-value associated with the five-factor model is 0.000304, smaller than 0.002370, the P-value associated with the four-factor model. Intuitively, the more the number of factors, the better the model should fit the data. However, this seemingly contradictory result can be easily explained. It is true that as measured by the GMM statistic, the five-factor model fits the data better

[12] See Fama and French (1993), Harvey and Kirby (1995), He and Ng (1994) and He et al. (1996) for a detailed description of the data. The author is grateful to C. Harvey and R. Kan for permission and forward of their data.

because it has a statistic of 56.580360, smaller than 77.749437 of the four-factor model. But the five-factor model has 25 additional parameters, and this is penalized by the test which has now 25 fewer degrees of freedom. In other words, by adding 25 additional parameters, the gain in model fitting cannot offset the loss in the degrees of freedom of the test, and hence it produces a P-value for the five-factor model lower than a four-factor one.

As the number of factors increases from one to five, it is of interest to examine how much contributions each pre-specified factor makes to the linear combination. While the factor loadings provide an answer to this question, they vary greatly as the number of factors changes due to the factor normalization. Hence, a simple approach is to examine the regression coefficients or weights on each of the pre-specified five factors. For example, in a one-factor model, the magnitude of the weight offers information on how the original factor contributes to the explanatory power of the model. In a five-factor model, there will be no restrictions on the weights and hence they must be the same as the OLS regression coefficients from regressing the asset returns on all of the Fama and French five factors. Table 4 provides the results. This is done only for the 5jth asset ($j = 1, \dots, 5$) as results on all of the 25 assets would take too much space. Interestingly, the weights change little as the number of factors varies. For example, imposing a one-factor model, the weight or contribution of F_{MKT} is 1.1864, which is not much different from 1.0711, the weight in the five-factor model. Econometrically, this closeness is likely due to that fact that the one-factor model is designed to have the maximum explanatory power given the restriction on the number of factors. Intuitively, as there is not much change in the explanatory power from the one-factor model to the five-factor one, it is not surprising that the weights are not too much different.

By the testing results alone, it is difficult to distinguish among the various factor models. Table 5 provides a few model diagnostics that may help shed some light on the problem. The second column reports the residual errors averaged over time and across the assets. The average error tells how the model fits the data over time and across the assets. The third column reports the absolute residual errors averaged over time and across the assets. This average absolute error detects possibly large and offsetting residual errors. The fourth column reports the average of the R^2 across the assets. The in-sample results are computed from estimation over the entire sample period. As expected, the more the number of factors, the better the model fits the data. From a one-factor model to a two-factor one, the average error is reduced by about 50%. However, the average errors are almost the same from two- to five-factor models. Similarly, there are not much differences in the average absolute errors. In term of the R^2, it is generally higher with more factors.

As an additional diagnostic, we estimate the models by excluding the last 3 years data, and use the newly estimated parameters to compute the R^2s for the last 3 years. This gives rise to the out-of-sample results. In contrast to its

Table 4
Weights on the FF factors in the restricted factor model

Number of factors	MKT	SMB	HML	TERM	DEF
$K=1$	1.1864	0.4439	0.3119	− 0.0246	− 0.0596
	1.1884	0.4446	0.3124	− 0.0246	− 0.0597
	1.0972	0.4105	0.2885	− 0.0228	− 0.0551
	1.1135	0.4167	0.2928	− 0.0231	− 0.0559
	0.8545	0.3197	0.2246	− 0.0177	− 0.0429
$K=2$	1.0817	0.7428	0.7534	0.0066	− 0.0676
	1.1186	0.6444	0.6011	− 0.0058	− 0.0670
	1.0482	0.5512	0.4875	− 0.0108	− 0.0615
	1.0872	0.4934	0.3930	− 0.0192	− 0.0619
	0.8428	0.3520	0.2608	− 0.0180	− 0.0473
$K=3$	1.0551	0.9110	0.6075	− 0.0174	− 0.1242
	1.1162	0.6588	0.5884	− 0.0068	− 0.0714
	1.0573	0.4928	0.5378	− 0.0010	− 0.0413
	1.1104	0.3461	0.5202	0.0037	− 0.0117
	0.8906	0.0474	0.5236	0.0303	0.0571
$K=4$	1.0754	0.8979	0.6222	− 0.0643	− 0.2059
	1.1234	0.6536	0.5930	− 0.0195	− 0.0999
	1.0426	0.5015	0.5264	0.0383	0.0180
	1.0981	0.3529	0.5105	0.0394	0.0380
	0.9165	0.0294	0.5402	− 0.0182	− 0.0465
$K=5$	1.0711	0.8860	0.6107	0.0374	− 0.1796
	1.1233	0.6535	0.5928	− 0.0179	− 0.0994
	1.0478	0.5154	0.5399	− 0.0806	− 0.0128
	1.1024	0.3645	0.5217	− 0.0596	0.0124
	0.9199	0.0387	0.5492	− 0.0979	− 0.0671

The table reports the regression coefficients or weights on each of the pre-specified five factors in the K-factor model for the $5j$th asset ($j = 1, \ldots, 5$).

in-sample performance, a five-factor performs worse than a two-, three- or four-factor model as measured by the R^2. Combining both the testing and the out-of-sample results, it appears that a three-factor model is better than a five-factor one. In other words, there does seem some gain in using a few linear combinations of the given factors rather than in using all of them.

Our methodology picks up a three-factor model where the factors are linear combinations of Fama and French's (1993) original five. This model performs statistically better than the five-factor model, and also does better (by design) than the three-factor model advocated by Fama and French which simply drops

Table 5
Model diagnostics

Number of factors	Average errors	Average absolute errors	Average R^2
In-sample			
$K = 1$	− 0.000234	0.016741	0.843338
$K = 2$	− 0.000115	0.015270	0.872199
$K = 3$	− 0.000104	0.014044	0.894909
$K = 4$	− 0.000104	0.014004	0.895476
$K = 5$	− 0.000106	0.013981	0.895440
Out-of-sample			
$K = 1$	− 0.001458	0.017829	0.776520
$K = 2$	− 0.001387	0.016923	0.802872
$K = 3$	− 0.001416	0.016997	0.809803
$K = 4$	− 0.001438	0.016946	0.804791
$K = 5$	− 0.001437	0.016967	0.798259

The table reports the average residual errors, the average of the absolute value of the residual errors, and the average (across assets) of the R^2. The in-sample results are for the entire sample period, and the out-of-sample results are for the last three years.

the TERM and DEF factors out of the five. However, it remains unknown whether or not this model can perform better in explaining the cross-section of expected stock returns than the Fama and French three-factor model. To answer this question, we run the standard OLS two-pass regressions by using separately both the extracted factors and those of the Fama and French's chosen three. The results are summarized in Table 6. The risk premium estimates for the extracted factors are all positive. In contrast, the risk premiums for the three factors chosen by Fama and French (1993) do not have the same signs. For example, Fama and French's estimated market risk premium is − 0.0767, slightly negative. As the factors are different, it is difficult to assess which three-factor models is preferred on the basis of the risk premiums (although a negative market risk premium does not seem plausible from the perspective of standard asset pricing theories). Fortunately, the Q_c test statistic is both a measure of the cross-sectional pricing errors and a test of the asset pricing restrictions. As Q_c are 0.1226 and 0.1318 respectively, the model with the extracted factors performs better in explaining the cross-section of expected stock returns than Fama and French's three-factor one. This is also supported by the associated p-values, which are 0.01419 and 0.0066, respectively.

Brennan et al. (1998) examine the explanatory power in the cross-section of stock returns by using 14 economic variables. It seems that the methodology of

Table 6
A comparison with Fama and French risk premiums

Q_c	γ_1	γ_2	γ_3
Panel A: Risk Premiums for f_{MKT}, f_{SMB} and f_{HML}			
0.1318	– 0.0767	0.1018	0.1797
(0.0066)	(0.0906)	(0.0767)	(0.0822)
Panel B: Risk Premiums for f_1, f_2 and f_3			
0.1226	0.1199	0.1463	0.1561
(0.0142)	(0.2611)	(0.2586)	(0.2550)

The table reports both the test, Q_c, for the asset pricing restrictions (its p-value in the bracket) and the risk premium estimates (their asymptotic standard errors in brackets) for both the Fama and French factors (f_{MKT}, f_{SMB} and f_{HML}) and the extracted factors (f_1, f_2 and f_3).

this paper can be useful in reducing the number of variables both cross-sectionally and time-series wise. Furthermore, as the variables are from diverse sources, it may be of interest to use the extended methodology of Section 2.3 to divide the variables into several groups and to extract the factors from each group accordingly. As the methodology is of the primary interest in this paper, additional empirical applications appear to go beyond the scope of this paper.

6. Conclusions

This paper provides a new framework to extract factors from either a latent factor model or a factor model with pre-specified factors. In particular, it presents a useful technique to find linear combinations of known economic variables that best forecast latent factors, such as those in Ross's (1976) arbitrage pricing theory and those in models of the term structure of interest rates. Based on the proposed model, we provide a test for the APT based on the generalized method of moments (GMM). This test is potentially more robust than almost all of the existing ones because it does not suffer from the usual errors-in-variables problem, nor does it require the strong assumption that the security returns are normally distributed and temporarily independent.

By using monthly industry returns and the five economic variables of Chen et al. (1986), we find that a two-factor APT model cannot be rejected: a linear combination of industrial production and change in expected inflation determines the first factor and a linear combination of default risk and term premium determines the second. The methodology of the paper appears generally useful in identifying from commonly used or collected economic variables the underlying factors that affect security returns and the term structure of interest rates. As an application of extracting factors as linear combinations of pre-specified factors, we extract and test whether fewer factors out of the five of Fama and French (1993) are sufficient to explain the average returns on their 25 stock portfolios formed on size and book-to-market. While inconclusive in sample, a three-factor model appears to perform better out-of-sample than both four- and five-factor models. It seems interesting future research to apply the methodology of this paper to extract factors for international security returns and for the term structure of interest rates where parsimony in the number of factors is very important for pricing bonds and other interest rate derivative securities.

Appendix A. Derivatives of the objective function

For the reader's convenience, this appendix provides an explicit expression for $\boldsymbol{D}_T$ that is useful for implementing computations of the analytical GMM test as well as for checking possible coding errors by examining whether $\boldsymbol{D}_T' \boldsymbol{W}_T \boldsymbol{g}_T$ is zero.

Consider first the case where $\boldsymbol{\alpha}$ is constrained to zero. In this case, we can order the normalized $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (C_{11}, \dots, C_{M1}, \dots, C_{1K}, \dots, C_{MK}, A_{(K+1)1}, \dots, A_{(K+1)K}, \dots, A_{N1}, \dots, A_{NK})$. Writing out $\boldsymbol{U}_t$ in terms of the parameters and observables, we obtain the expression

$$\boldsymbol{D}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \boldsymbol{U}_t}{\partial \boldsymbol{\theta}} \otimes \boldsymbol{Z}_t \equiv \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} -\boldsymbol{U}_1 & \boldsymbol{0} \\ -\boldsymbol{U}_3 & -\boldsymbol{U}_4 \end{bmatrix} \otimes \boldsymbol{Z}_t, \tag{A.1}$$

where $\boldsymbol{U}_1$ is a $K \times KM$ submatrix, $\boldsymbol{U}_3$ is a $(N-K) \times KM$ submatrix, and $\boldsymbol{U}_4$ is $(N-K) \times (q-KM)$ submatrix of the $N \times q$ matrix of partials of $\boldsymbol{U}_t$ with respect to the parameters. The submatrices can be written

$$\boldsymbol{U}_1 = \begin{bmatrix} X_{t1} & \cdots & X_{tM} & \cdot & \cdots & \cdot & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & X_{t1} & \cdots & X_{tM} \end{bmatrix}, \tag{A.2}$$

$$\boldsymbol{U}_3 = \begin{bmatrix} A_{(K+1)1}X_{t1} & \cdots & A_{(K+1)1}X_{tM} & \cdot & \cdots & \cdot & A_{(K+1)K}X_{t1} & \cdots & A_{(K+1)K}X_{tM} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ A_{N1}X_{t1} & \cdots & A_{N1}X_{tM} & \cdot & \cdots & \cdot & A_{NK}X_{t1} & \cdots & A_{NK}X_{tM} \end{bmatrix},$$

and

$$\boldsymbol{U}_4 = \begin{bmatrix} \sum C_{j1}X_{tj} & \cdots & \sum C_{jK}X_{tj} & \cdot & \cdots & \cdot & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & \cdot & \cdots & \cdot & \sum C_{j1}X_{tj} & \cdots & \sum C_{jK}X_{tj} \end{bmatrix}.$$

Consider now the case where $\boldsymbol{\alpha}$ is unconstrained. There are N more parameters and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', C_{11}, \ldots, C_{M1}, \ldots, C_{1K}, \ldots, C_{MK}, A_{(K+1)1}, \ldots, A_{(K+1)K}, \ldots, A_{N1}, \ldots, A_{NK})$. It is clear that we need only to add $-\boldsymbol{I}_N$ into the previous $\partial \boldsymbol{U}_t / \partial \boldsymbol{\theta}$ matrix to compute $\boldsymbol{D}_T$.

Appendix B. Analytical iterations

To show (15), we re-write the model as

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + \boldsymbol{A}_1 \boldsymbol{C}_1 \boldsymbol{X}_{t1} + \cdots + \boldsymbol{A}_p \boldsymbol{C}_p \boldsymbol{X}_{tp} + \boldsymbol{U}_t \tag{B.1}$$

$$= [\boldsymbol{\alpha}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_p] \begin{pmatrix} 1 \\ \boldsymbol{C}_1 \boldsymbol{X}_{t1} \\ \vdots \\ \boldsymbol{C}_p \boldsymbol{X}_{tp} \end{pmatrix} + \boldsymbol{U}_t. \tag{B.2}$$

Denote the first term of Eq. (B.2) as $\boldsymbol{A}\boldsymbol{X}_t$. Using the matrix formula, $\mathrm{vec}(PAQ) = (Q' \otimes P)\mathrm{vec}\, A$, on the vec operator sucessively, we obtain

$$\begin{aligned} \boldsymbol{A}\boldsymbol{X}_t \otimes \boldsymbol{Z}_t &= \mathrm{vec}[\boldsymbol{Z}_t \cdot 1 \cdot (\boldsymbol{A}\boldsymbol{X}_t)'] \\ &= \mathrm{vec}(\boldsymbol{Z}_t \boldsymbol{X}_t' \boldsymbol{A}' \boldsymbol{I}_N) = (\boldsymbol{I}_N \otimes \boldsymbol{Z}_t \boldsymbol{X}_t')\mathrm{vec}\boldsymbol{A}'. \end{aligned} \tag{B.3}$$

Hence, the GMM sample moments function has the following form:

$$g_T = \frac{1}{T} \sum_{t=1}^{T} (\boldsymbol{R}_t - \boldsymbol{A}\boldsymbol{X}_t) \otimes \boldsymbol{Z}_t = \boldsymbol{y} - \boldsymbol{X}^{**}\boldsymbol{a}, \tag{B.4}$$

where $\boldsymbol{y}$, $\boldsymbol{X}^{**}$ and $\boldsymbol{a}$ are the same as those defined in the text. It then follows from (B.4) that (15) holds as the solution to the GMM minimization problem.

To show (16), we re-write the model as

$$\boldsymbol{R}_t = \boldsymbol{\alpha} + (\boldsymbol{X}_1' \otimes \boldsymbol{A}_1)\mathrm{vec}\boldsymbol{C}_1 + \cdots + (\boldsymbol{X}_p' \otimes \boldsymbol{A}_p)\mathrm{vec}\boldsymbol{C}_p + \boldsymbol{U}_t \tag{B.5}$$

$$= \boldsymbol{\alpha} + [\boldsymbol{X}_1' \otimes \boldsymbol{A}_1, \ \ldots, \boldsymbol{X}_p' \otimes \boldsymbol{A}_p] \begin{pmatrix} \mathrm{vec}\boldsymbol{C}_1 \\ \vdots \\ \mathrm{vec}\boldsymbol{C}_p \end{pmatrix} + \boldsymbol{U}_t. \tag{B.6}$$

Denote the first two terms of Eq. (B.6) as $\boldsymbol{\alpha} + \tilde{\boldsymbol{X}}_t \boldsymbol{C}$. Now, it is easy to show that $\tilde{\boldsymbol{X}}_t \boldsymbol{C} \otimes Z_t = (\tilde{\boldsymbol{X}}_t \otimes Z_t)\mathrm{vec}\, \boldsymbol{C}'$ and hence we have (16).

References

Bakshi, G.S., Chen, Z., 1996. Asset pricing without consumption or market portfolio data. Working paper, Ohio State University.

Breeden, D.T., Gibbons, M.R., Litzenberger, R.H., 1989. Empirical tests of the consumption based CAPM.. Journal of Finance 44, 231–262.

Brennan, M.J., Chordia, T., Subrahmanyam, A., 1998. Cross-sectional determinants of expected returns. Journal of Financial Economics 49, 345–373.

Brown, S.J., Goetzmann, W.N., Grinblatt, M., 1997. Positive portfolio factors. Working paper, Yale University.

Chen, N.-f., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. Journal of Business 59, 383–403.

Connor, G., 1984. A united beta pricing theory. Journal of Economic Theory 34, 13–31.

Connor, G., Korajczyk, R.A., 1986. Performance measurement with the arbitrage pricing theory: A new framework for analysis. Journal of Financial Economics 15, 373–394.

Connor, G., Korajczyk, R.A., 1995. The arbitrage pricing theory and multifactor models of asset returns. in: Jarrow, R.A. (Ed.), Handbooks in Operations Research and Management Science: Finance, Vol. 9. North-Holland, Amsterdam, pp. 87–144.

Epps, T.W., Kramer, C.F., 1995. A variance test of the linear factor representation of returns and equilibrium pricing restrictions. Working paper, University of Virginia.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Ferson, W.E., Harvey, C.R., 1991. The variation of economic risk premiums. Journal of Political Economy 99, 385–415.

Geweke, J., Zhou, G., 1996. Measuring the pricing error of the arbitrage pricing theory. Review of Financial Studies 9, 553–583.

Gibbons, M.R., Ferson, W., 1985. Testing asset pricing models with changing expectations and an unobservable market portfolio. Journal of Financial Economics 14, 217–236.

Gibbons, M.R., Ross, S.A., Shanken, J., 1989. A test of the efficiency of a given portfolio. Econometrica 57, 1121–1152.

Hansen, L.P., 1982. Large sample properties of the generalized method of moments estimators. Econometrica 50, 1029–1054.

Hansen, L.P., Jagannathan, R., 1997. Assessing specification errors in stochastic discount factor model. Journal of Finance 52, 557–590.

Harvey, C.R., Kirby, C., 1995. Analytic tests of factor pricing models. Working paper, Duke University, Durham, NC.

He, J., Ng, L.K., 1994. Economic forces, fundamental variables, and equity returns. Journal of Business 67, 599–609.

He, J., Kan, R., Ng, L.K., Zhang, C., 1996. Tests of the relationships among economic forces, fundamental variables, and equity returns using multi-factor models. Journal of Finance 51, 1891–1908.

Ingersoll, J.E., 1987. Theory of Financial Decision Making. Rowman and Littlefield, New York.

Jagannathan, R., Wang, Z., 1996. The conditional CAPM and the cross-section of expected returns. Journal of Finance 51, 3–53.

Jöreskog, K.G., Goldberger, A.S., 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of American Statistical Association 70, 631–639.

Lehmann, B.N., Modest, D.M., 1988. The empirical foundations of the arbitrage pricing theory. Journal of Financial Economics 21, 213–254.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics 47, 13–37.

MacKinlay, A.C., Richardson, M.P., 1991. Using generalized method of moments to test mean-variance efficiency. Journal of Finance 46, 511–527.

Nardari, F., Zhou, G., 1999. What is the decomposition of the market wealth portfolio in the CAPM? Working paper, Washington University.

Newey, W.K., 1985. Generalized method of moments specification testing. Journal of Econometrics 29, 229–256.
Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.
Roll, R., 1977. A critique of the asset pricing theory's tests: Part I. Journal of Financial Economics 4, 129–176.
Ross, S.A., 1976. The arbitrage theory of capital asset pricing. Journal of Economic Theory 13, 341–360.
Seber, G.A.F., 1984. Multivariate Observations. Wiley, New York.
Shanken, J., 1985. Multivariate tests of the zero-beta CAPM. Journal of Financial Economics 14, 327–348.
Shanken, J., 1992. The current state of the arbitrage pricing theory. Journal of Finance 47, 1569–1574.
Shanken, J., 1996. Statistical methods in tests of portfolio efficiency: a synthesis. in: Maddala, G.S., Rao, C.R. (Eds.), Handbooks of Statistics: Statistical Methods in Finance, Vol. 14. North-Holland, Amsterdam, pp. 693–711.
Sharpe, W.F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance 19, 425–442.
Stambaugh, R., 1982. On the exclusion of assets from tests of the two-parameter model: a sensitivity analysis. Journal of Financial Economics 10, 237–268.
Velu, R., Zhou, G., 1999. Testing multi-beta pricing models. Journal of Empirical Finance, forthcoming.
Zhou, G., 1994. Analytical GMM tests: asset pricing with time-varying risk premiums. Review of Financial Studies 7, 687–709.