Analytical GMM Tests: Asset Pricing with Time-Varying Risk Premiums

Guofu Zhou
Washington University

We propose alternative generalized method of moments (GMM) tests that are analytically solvable in many econometric models, yielding in particular analytical GMM tests for asset pricing models with time-varying risk premiums. We also provide simulation evidence showing that the proposed tests have good finite sample properties and that their asymptotic distribution is reliable for the sample size commonly used. We apply our tests to study the number of latent factors in the predictable variations of the returns on portfolios grouped by industries. Using data from October 1941 to September 1986 and two sets of instrumental variables, we find that the tests reject a one-factor model but not a two-factor one.

A fundamental problem in finance is to characterize the expected return on a security. Sharpe (1964), Lintner (1965), Black (1972), Ross (1976), Merton (1973), and Breeden (1979), among others, develop asset pricing models that imply that the expected return on a security is a linear function of factor risk premiums. Traditional empirical analysis assumes that the factor risk premiums are constant. There has been an enormous amount of research in this direction.

I am grateful to Ronald Allen, Kerry Back, Philip Dybvig, Wayne Ferson, Lars Hansen, Campbell Harvey, James Jones, James Steeley, and Jack Strauss for helpful discussions and comments. I am especially grateful to Robert Stambaugh (the editor) and an anonymous referee for providing many insightful and detailed suggestions that substantially improved the article.
Address correspondence to Guofu Zhou, John M. Olin School of Business, Washington University, St. Louis, MO 63130.

The Review of Financial Studies Winter 1994 Vol. 7, No. 4, pp. 687–709

Gibbons, Ross, and Shanken (1989), Kandel and Stambaugh (1990), Zhou (1991), and Shanken (1992), for example, provide both the estimation strategies and the associated tests for the constant risk premium beta pricing models.

Many recent studies allow the factor risk premiums to be time-varying, generating a class of *latent variables* models. Hansen and Hodrick (1983) and Gibbons and Ferson (1985) are the first to derive such models to examine time-varying factor risk premiums. Subsequently, Campbell (1987), Cumby (1987), Chan (1988), Stambaugh (1988), Campbell and Hamao (1992), Ferson (1990), Chang, Pinegar, and Ravichandran (1991), and Zhou (1993a), among others, use similar models to study stock returns, forward currency premiums, international equity returns, and capital market integration. Ferson and Foerster (1991) provide a brief survey of the literature in addition to their study of finite sample properties of Hansen's (1982) generalized method of moments (GMM) test. This GMM test is the predominant approach for parameter estimation and hypothesis testing in the latent variables models.

This article proposes new GMM tests that are analytically solvable in a wide range of asset pricing models, a special case of which is the latent variables models. One of the primary difficulties in applying the traditional GMM procedure is to solve the GMM optimization problem, which is often done numerically over a large parameter space. For example, in a simple case where there are 12 assets, 2 factors, and 5 instrumental variables, the optimization problem in the latent variables models requires the minimization of a complex nonlinear function in a 30-dimensional space. If there are 24 assets and 4 factors, the dimensionality increases to 100. One of the well-known problems with numerical procedures is that the solution may not converge to the global minimum or even converge at all. Because of this, the task of solving for the traditional GMM test can be very difficult if not impossible to accomplish in many applications. Our tests overcome this difficulty, making the GMM approach applicable to cases where there may be hundreds of parameters. The underlying idea of the tests is simple. Realizing that it is extremely difficult to solve the GMM optimization problem analytically in general, we focus instead on a special case where the model residuals are independent and identically distributed (i.i.d.). Fortunately, explicit solutions in the i.i.d. case are available for many models, and they can be adjusted to yield analytical GMM tests that are valid even in the general case with heteroskedasticity.

The article is organized as follows. In Section 1, we propose our alternative GMM tests based on arbitrary weighting matrices (in contrast, the usual GMM tests are obtained based on the optimal weight-

ing matrix). In Section 2, we apply our method to the latent variables models to obtain analytical GMM tests. The method may also be applied to testing, among others, the arbitrage pricing theory (APT) and the multibeta pricing models. In Section 3, we apply the tests to investigate the number of "priced" latent factors in the U.S. equity market. In Section 4, we provide simulation evidence on the performance of the proposed GMM tests in finite samples. Section 5 concludes the article.

1. Alternative GMM Tests

In this section, we show how to obtain the alternative GMM tests that are based on arbitrary weighting matrices. As will be clear later, the important advantage of the alternative GMM tests is that they can be analytically obtained in many applications where it may be extremely difficult to obtain numerically the conventional GMM test.

Hansen (1982) proposes a GMM for the estimation and testing of a wide class of econometric models. The idea of the GMM approach is to use sample moment conditions to replace those of the model. Then, the parameter estimators are obtained by minimizing a weighted quadratic form of the sample moments. Formally, let $\boldsymbol{\theta}$ be a q-vector of the parameters of an econometric model, $\mathbf{U}_t(\boldsymbol{\theta})$ be an N-vector of the model disturbances, and $\mathbf{Z}_{t-1}$ be an L-vector of the instruments. Then, the model can be written as

$$E[\mathbf{f}_t(\boldsymbol{\theta})] = \mathbf{0}, \qquad \mathbf{f}_t(\boldsymbol{\theta}) \equiv \mathbf{U}_t(\boldsymbol{\theta}) \otimes \mathbf{Z}_{t-1}, \tag{1}$$

where $\otimes$ is the Kronecker product that makes $\mathbf{f}_t$ an NL-vector function of both the disturbances and the instruments. Let $\mathbf{g}_T$ be the sample mean of $\mathbf{f}_t$:

$$\mathbf{g}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t(\boldsymbol{\theta}), \qquad NL \times 1. \tag{2}$$

At the true population parameters, because the population mean of $\mathbf{f}_t$ must be zero [satisfying the above NL moment conditions (1)], the sample mean $\mathbf{g}_T$ should be small and so should a quadratic form of $\mathbf{g}_T$. To estimate the parameters, we use the solution, $\hat{\boldsymbol{\theta}}$, which minimizes the quadratic form:

$$\min Q \equiv \mathbf{g}_T(\boldsymbol{\theta})'\mathbf{W}_T\mathbf{g}_T(\boldsymbol{\theta}), \tag{3}$$

where $\mathbf{W}_T$ is an $NL \times NL$ weighting matrix that is positive definite. The resulting estimator is Hansen's (1982) GMM estimator.

With different choices of the weighting matrix, one obtains different estimators. Although differing in statistical properties, these estimators are nevertheless consistent. Hansen (1982) derives the optimal

estimator, which is obtained by choosing the weighting matrix that minimizes the asymptotic covariance matrix of the estimator. Depending on the type of heteroskedasticity and autocorrelation assumed, the optimal weighting matrix may take different forms. In general, it can be taken as $\mathbf{W}_T = \mathbf{S}_T^{-1}$, where $\mathbf{S}_T$ is a consistent estimator of the covariance matrix of the model's moment conditions [Newey and West (1987) provide a general approach to obtain such an estimator].

Assume the number of moment conditions (NL) is greater than the number of parameters (q), so there are ($NL - q$) overidentification restrictions in the model. These overidentification restrictions are often tested by computing the well-known Hansen test statistic:

$$H_0 = T\mathbf{g}_T(\hat{\boldsymbol{\theta}})\ \mathbf{W}_T\mathbf{g}_T(\hat{\boldsymbol{\theta}}), \tag{4}$$

where $\mathbf{W}_T$ is the optimal weighting matrix, $\mathbf{W}_T = \mathbf{S}_T^{-1}$. Under certain regularity conditions [see Hansen (1982)], H_0 is asymptotically distributed χ^2 with the degree of freedom ($NL - q$).

Notice that, to obtain H_0, we have to solve the GMM optimization problem (3). Usually, the problem cannot be solved analytically and numerical procedures are the only available approach. However, there are three major difficulties in using numerical procedures. First, the success or failure of a numerical optimization algorithm usually depends on how close an initial estimate is to its true solution, and a good initial estimate is often difficult to obtain. Second, it is well known that, as the number of parameters increases, it becomes more difficult to search for the minimum in a space whose dimensionality is the number of the parameters. Finally, as emphasized by Judge et al. (1985, p. 969), numerical algorithms often converge only to the local maximum or minimum. Therefore, if the numerical solution of the GMM optimization problem (3) achieves only a local minimum, then the test statistic obtained will be greater than the global minimum and an erroneous rejection of the tested theory can result.

We can avoid the above difficulties if an analytical solution that achieves the global minimum is available. Notice that H_0 is defined with $\mathbf{W}_T$ as the optimal weighting matrix. Given the complexity of the optimal weighting matrix in the general case, it is difficult or impossible to solve (3) analytically. However, simple intuition suggests that an analytical solution to (3) may be possible if the weighting matrix is chosen of some particular form. Indeed, as shown in Section 2, we can obtain analytical solutions for a class of weighting matrices, such as the identity matrix or the inverse of a covariance matrix based on an i.i.d. assumption. Moreover, such analytical solutions are available not only for the latent variables models, but also for many others. Given that analytical solutions are available to (3) for some simplified weighting matrices different from the optimal one, there is no reason

that the resulting quadratic form multiplied by T should have an asymptotic χ^2 distribution. Hence, to make use of the possible analytical solutions, we need a new GMM test that is well defined for the simplified weighting matrices.

Theorem 1. *Let $\mathbf{W}_T$ be any $NL \times NL$ positive definite matrix that may be either stochastic or nonstochastic. The associated GMM estimator $\hat{\boldsymbol{\theta}}$ must be consistent.*[1] *Define*

$$H_z \equiv T(\mathbf{M}_T\mathbf{g}_T)'\mathbf{V}_T(\mathbf{M}_T\mathbf{g}_T), \tag{5}$$

where $\mathbf{g}_T = \mathbf{g}_T(\hat{\boldsymbol{\theta}})$ evaluated from (2); $\mathbf{V}_T$ is a diagonal matrix, $\mathbf{V}_T = \mathrm{Diag}(1/\nu_1, \ldots, 1/\nu_d, 0, \ldots, 0)$, formed by $\nu_1 > \ldots > \nu_d > 0$, the positive eigenvalues of the following $NL \times NL$ semidefinite matrix:

$$\boldsymbol{\Omega}_T \equiv [\mathbf{I} - \mathbf{D}_T(\mathbf{D}_T'\mathbf{W}_T\mathbf{D}_T)^{-1}\mathbf{D}_T'\mathbf{W}_T]\mathbf{S}_T[\mathbf{I} - \mathbf{D}_T(\mathbf{D}_T'\mathbf{W}_T\mathbf{D}_T)^{-1}\mathbf{D}_T'\mathbf{W}_T]', \tag{6}$$

where $\mathbf{S}_T$ is a consistent estimator of the residual covariance matrix; $\mathbf{D}_T$ is an $NL \times q$ matrix of the first-order derivatives of $\mathbf{g}_T$ with respect to $\boldsymbol{\theta}$; and $\mathbf{M}_T$ is an $NL \times NL$ matrix, of which the ith row is the standardized eigenvector corresponding to the ith largest eigenvalue ($i = 1, \ldots, NL$). Then H_z is asymptotically distributed χ^2 with degrees of freedom $(NL - q)$.

Proof. See Appendix A.

Theorem 1 says that, based on an arbitrary weighting matrix or an arbitrary consistent GMM estimator, an asymptotic test can be constructed. If the consistent GMM estimator is analytically available or if (3) can be solved analytically for some specially chosen weighting matrix, then H_z provides an analytical GMM test. As will be clear later, we often impose the i.i.d. assumption on the model residuals to obtain an analytical GMM estimator, but H_z computed from (5) is valid under the most general heteroskedasticity assumptions as spelled out in Hansen (1982).

In comparison with the traditional GMM test, which is often difficult or impossible to obtain, H_z is *analytically* tractable for a number of models, of which the latent variables models is a special case. This is the major advantage of H_z over the traditional GMM test. On the other hand, the motivations of the two tests are closely related. Both H_z and H_0 are based on normalizations of the asymptotic covariance matrix of $\mathbf{S}_T^{-1/2}\mathbf{g}_T$. To examine this linkage, consider $\mathbf{W}_T = \mathbf{S}_T^{-1}$ in the i.i.d. case. Because $\mathbf{W}_T$ is now the optimal weighting matrix, the

[1] Here we suppose for Theorem 2 that the regularity conditions, Assumptions 3.1–3.6 of Hansen (1982), are satisfied.

asymptotic covariance matrix is already idempotent, and thus no normalization is necessary. This implies that the $\mathbf{M}_T$ and the $\mathbf{V}_T$ matrices, which play the role of normalization, can be chosen as $\mathbf{S}_T^{-1/2}$ and the identity matrix, respectively. This procedure gives rise to the traditional GMM test. On the other hand, H_z is obtained with normalization of the asymptotic covariance matrix regardless of whether the weighting matrix is optimal or not. Hence, H_z is a generalization of the traditional GMM test: the former is defined for an *arbitrary* weighting matrix while the latter is defined only for the optimal weighting matrix.

2. Applications

In this section, we discuss first the asset pricing restrictions in the latent variables models and show that the usual latent variables tests are in fact a test of a rank hypothesis. Second, we apply our method in Section 1 to obtain analytical GMM tests to test the rank hypothesis. The method can be useful wherever the GMM approach is relevant. For example, it can also yield analytical and semianalytical tests of such nonlatent variables models as the multibeta pricing models [see, e.g., Kandel and Stambaugh (1990)], the APT [Ross (1976)], and the consumption capital asset pricing model (CCAPM) [Breeden (1979)]. Finally, we examine the relationship between the GMM estimator and the maximum likelihood estimator under the rank restriction.

2.1 Asset pricing restrictions

Consider a general K-factor asset pricing model of the form

$$E(r_{it} \mid \mathbf{Z}_{t-1}) = \lambda_0(\mathbf{Z}_{t-1}) + \beta_{1i}\lambda_1(\mathbf{Z}_{t-1}) + \ldots + \beta_{Ki}\lambda_K(\mathbf{Z}_{t-1}), \quad (7)$$
$$i = 0, 1, \ldots, N, \qquad t = 1, \ldots, T,$$

where r_{it} is the return on asset i between period $t-1$ and t; $\lambda_j(\mathbf{Z}_{t-1})$ is the marketwide expected risk premium on the jth factor; $\mathbf{Z}_{t-1}$, $L \times 1$, is the marketwide information available at t; $\beta_{1i}, \ldots, \beta_{Ki}$ are the conditional betas of asset i; $N+1$ is the number of assets ($N > K$); and T is the number of periods.

In terms of excess returns, the pricing relation (8) can be written

$$E(R_{it} \mid \mathbf{Z}_{t-1}) = b_{1i}\lambda_1(\mathbf{Z}_{t-1}) + \ldots + b_{Ki}\lambda_K(\mathbf{Z}_{t-1}), \quad (8)$$
$$i = 1, \ldots, N, \qquad t = 1, \ldots, T,$$

where $R_{it} = r_{it} - r_{0t}$ is the return on the ith asset in excess of the return on the 0th asset (the 0th asset is arbitrarily ordered), and $b_{ji} = \beta_{ji} - \beta_{j0}$ is the "excess" conditional beta. In matrix form,

$$E(\mathbf{R} \mid \mathbf{Z}) = \lambda(\mathbf{Z})\mathbf{B}, \quad (9)$$

where $\mathbf{R}$ is a $T \times N$ matrix formed by the N excess returns over T periods; $\mathbf{Z}$, $T \times L$, the instrumental variables; $\lambda(\mathbf{Z})$, $T \times K$, the risk premiums on the K factors; and $\mathbf{B}$, $K \times N$, the excess conditional betas. We assume that the number of information variables is greater than the number of factors, that is, $L > K$. In addition, we assume that both $\lambda(\mathbf{Z})$ and $\mathbf{B}$ have full-column rank K [$\lambda(\mathbf{Z})$ is so with probability one]. Otherwise, (8) would be reduced to a pricing model with the number of factors being less than K.

As emphasized by Fama (1991), we cannot test any pricing theory without specifying the law of motion for the asset returns. Given statistical assumptions about the stochastic behavior of the returns, Equation (9) has testable restrictions on the parameters of the statistical model. Following most studies, we assume the returns are governed by the multivariate regression model:

$$R_{it} = \theta_{1i}Z_{t-1,1} + \ldots + \theta_{Li}Z_{t-1,L} + u_{it}, \qquad (10)$$
$$i = 1, \ldots, N, \qquad t = 1, \ldots, T,$$

where u_{it}'s are the disturbances or forecasting errors that have zero means conditional on the instruments. Put differently, the realized returns are decomposed into two parts: the predictable part and unpredictable part. The predictable part is $\theta_{1i}Z_{t-1,1} + \ldots + \theta_{Li}Z_{t-1,L}$, which can be forecasted by the investor's available information at t. The unpredictable part cannot be forecasted with the available information.

The K-factor asset pricing model says that the predictable returns are driven by K latent factors, resulting in the pricing restriction (9). The objective is to test whether this restriction is valid given that the returns are governed by (10). By (9) and (10), we have $E(\mathbf{R} \mid \mathbf{Z}) = \mathbf{Z\Theta}$ and $\lambda(\mathbf{Z}) = \mathbf{ZA}$, where $\mathbf{A}$ is an $L \times K$ constant matrix. Hence, we can conclude that $\mathbf{Z\Theta} = \lambda(\mathbf{Z})\mathbf{B} = \mathbf{ZAB}$. Assume throughout that there are no redundancies in the information set, so that the $L \times L$ matrix $\mathbf{Z'Z}$ is invertible (with probability one). Then we must have $\mathbf{\Theta} = \mathbf{AB}$, implying that $\mathbf{\Theta}$ has rank less than or equal to K. By the assumptions on $\lambda(\mathbf{Z})$ and $\mathbf{B}$, $\mathbf{\Theta}$ must have rank K. On the other hand, if $\mathbf{\Theta}$ has rank K, there must exist an $L \times K$ matrix $\mathbf{A}$ and $K \times N$ matrix such that $\mathbf{\Theta} = \mathbf{AB}$ and the pricing restriction follows. Therefore, given the return process (10), the pricing restriction (9) is valid if and only if the multivariate regression coefficient matrix $\mathbf{\Theta}$ has rank K.

2.2 Analytical GMM tests

In latent variables models, the conditional mean of the disturbances is assumed to be zero:

$$E(\mathbf{U}_t \mid \mathbf{Z}_{t-1}, \mathbf{U}_{t-1}, \mathbf{Z}_{t-2}, \ldots) = 0. \qquad (11)$$

Equation (11) says that economic agents' forecast of the return of the unpredictable part is zero conditional on the available information. In this case, as shown in Hansen (1982), a consistent estimator of the covariance matrix of the moment conditions is given by

$$\mathbf{S}_T = \frac{1}{T}\sum_{t=1}^{T} (\mathbf{U}_t\mathbf{U}_t' \otimes \mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'), \tag{12}$$

where $\mathbf{U}_t$ is evaluated at any consistent estimate of the model parameters $\boldsymbol{\Theta}$. If we assume that the conditional covariance matrix of the residuals is constant, then a consistent estimator of the covariance matrix of the moment conditions is

$$\mathbf{S}_T = \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{U}_t\mathbf{U}_t'\right) \otimes \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\right). \tag{13}$$

This latter $\mathbf{S}_T$ is also a consistent estimator when the residuals are i.i.d. In other applications, the heteroskedasticity may be different, and so the general estimator of Newey and West (1987) should be used in place of $\mathbf{S}_T$.

Given the optimal weighting matrix $\mathbf{W}_T = \mathbf{S}_T^{-1}$ where $\mathbf{S}_T$ is given by (12), it is very difficult if not impossible to solve analytically the GMM optimization problem (3). Fortunately, we can obtain an explicit GMM estimator (which achieves the global minimum) for a wide class of weighting matrices including, in particular, the identity matrix and $\mathbf{W}_T = \mathbf{S}_T^{-1}$ where $\mathbf{S}_T$ is given by (13). To make it easier for applications, we summarize the results as Theorem 2. To simplify the presentation, we will use the matrix form of the return process (10):

$$\mathbf{R} = \mathbf{Z}\boldsymbol{\Theta} + \mathbf{U}, \tag{10}$$

where $\mathbf{R}$, $\mathbf{Z}$, and $\boldsymbol{\Theta}$ are defined in the previous subsection.

Theorem 2. *If the weighting matrix is of the form:*

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

then the GMM estimator of $\boldsymbol{\Theta}$ *under the rank K restriction is explicitly given by*

$$\hat{\boldsymbol{\Theta}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}, \qquad \tilde{\boldsymbol{A}}\colon L \times K, \qquad \tilde{\boldsymbol{B}}\colon K \times N, \tag{14}$$

where

$$\tilde{\boldsymbol{A}} = (\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}/T^2)^{-1/2}\boldsymbol{E}, \qquad \boldsymbol{P} \equiv \boldsymbol{Z}\boldsymbol{W}_2\boldsymbol{Z}',\ \boldsymbol{P}\colon T \times T,$$

$$\tilde{\boldsymbol{B}} = (\boldsymbol{Z}^{*\prime}\boldsymbol{P}\boldsymbol{Z}^*)^{-1}\boldsymbol{Z}^{*\prime}\boldsymbol{P}\boldsymbol{R}, \qquad \boldsymbol{Z}^* \equiv \boldsymbol{Z}\tilde{\boldsymbol{A}},\ \boldsymbol{Z}^*\colon T \times K,$$

and $\boldsymbol{E}$ *is the* $L \times K$ *matrix that stacks the "standardized" eigenvectors* ($\boldsymbol{E}'\boldsymbol{E} = \boldsymbol{I}_K$) *corresponding to the K largest eigenvalues of the* $L \times L$

matrix:

$$(\mathbf{Z}'\mathbf{PZ}/T^2)^{-1/2\prime}(\mathbf{Z}'\mathbf{PR}/T^2)\mathbf{W}_1(\mathbf{Z}'\mathbf{PR}/T^2)'(\mathbf{Z}'\mathbf{PZ}/T^2)^{-1/2}. \quad (15)$$

Furthermore, the minimum of Q is given by

$$Q^* = tr\mathbf{W}_1(\mathbf{R}'\mathbf{PR}/T^2) - \gamma_1 - \ldots - \gamma_K, \quad (16)$$

where $\gamma_1, \ldots, \gamma_K$ are the K largest eigenvalues of the $L \times L$ matrix given by (15).

Proof. See Appendix B.

The difficulties in the usual numerical approach can potentially be insurmountable in applications where the number of the parameters is large. This is especially true when either the number of assets or the number of instruments is large. In contrast, there are in principle no difficulties in using the analytical solutions. As long as $\mathbf{Z}'\mathbf{PZ}$ and $\mathbf{Z}^{*\prime}\mathbf{PZ}^*$ are invertible matrices, the analytical estimators can be evaluated for any number of assets or instruments. There is no concern with convergence and the result is guaranteed to be the global minimum.

To efficiently implement the analytical GMM estimator, one can compute $\mathbf{Z}'\mathbf{PZ}$ as $(\mathbf{Z}'\mathbf{Z})'\mathbf{W}_2(\mathbf{Z}'\mathbf{Z})$. $\mathbf{Z}'\mathbf{PR}$ can be treated similarly so that no $T \times T$ matrices are needed for storage in the computer program. Because it is computationally more efficient to obtain the Cholesky decomposition than to obtain the square root, $(\mathbf{Z}'\mathbf{PZ}/T^2)^{-1/2\prime}$ may be replaced by the lower triangular matrix of the Cholesky decomposition of the matrix $(\mathbf{Z}'\mathbf{PZ}/T^2)^{-1}$. To verify the results against possible coding errors in the computer program, the minimized Q may be computed from both (16) and $Q = \mathbf{g}_T'\mathbf{W}_T\mathbf{g}_T$. Additionally, the first-order derivatives of Q with respect to the parameters may also be computed and checked as to whether they are zero.

The estimates of $\mathbf{A}$ and $\mathbf{B}$ are not unique, since for any given estimate of $\mathbf{A}$ and $\mathbf{B}$ linear transformations of them, $\mathbf{AC}$ and $\mathbf{C}^{-1}\mathbf{B}$, give rise to the same estimate of $\boldsymbol{\Theta}$, where $\mathbf{C}$ is any $K \times K$ invertible matrix. Fortunately, the estimate of $\boldsymbol{\Theta}$ is unique, and so the estimates of both $\mathbf{A}$ and $\mathbf{B}$ are determined up to a linear transformation. In particular, they are unique under the following normalization:

$$\tilde{\mathbf{A}}'\mathbf{Z}'\mathbf{PZ}\tilde{\mathbf{A}} = T^2\mathbf{I}_K. \quad (17)$$

The unique estimator provided by Theorem 2 uses this normalization. In the latent-variables literature, the most widely used normalization partitions the parameter matrix $\mathbf{B}$ into $(\mathbf{I}_K, \mathbf{B}_2)$, thus requiring the first $K \times K$ submatrix of $\hat{\mathbf{B}}$ to be the identity matrix, that is, $\hat{\mathbf{B}} = (\mathbf{I}_K, \hat{\mathbf{B}}_2)$ [see, e.g., Ferson and Foerster (1991), p. 9]. The estimators under this conventional normalization are easily obtained from $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ as $\hat{\mathbf{A}} =$

$\tilde{\mathbf{A}}\mathbf{C}^{-1}$ and $\hat{\mathbf{B}} = \mathbf{C}\tilde{\mathbf{B}}$, where $\mathbf{C}$ is a $K \times K$ nonsingular matrix such that the upper $K \times K$ submatrix of $\mathbf{C}\tilde{\mathbf{B}}$ is an identity matrix of order K. Such a matrix $\mathbf{C}$ is in fact uniquely given by the inverse of the first $K \times K$ submatrix of $\tilde{\mathbf{B}}$, which is nonsingular (with probability one).

Based on Theorem 2, a consistent estimate of $\boldsymbol{\theta}$ is first analytically obtained by choosing the weighting matrix as the identity matrix. Then, $\mathbf{S}_T$ can be computed from (13) and the GMM estimator can be analytically evaluated for $\mathbf{W}_T = \mathbf{S}_T^{-1}$. This estimator will be the optimal GMM estimator in the case where the model residuals satisfy (11) and where the conditional covariance matrix is constant, as occurs with i.i.d. residuals. In general, however, the consistent estimator provided by Theorem 2 is not the optimal estimator in the presence of heteroskedasticity. Nevertheless, Theorem 1 allows us to construct the analytical GMM test H_z as given by (5).

After the usual normalization of the population parameters, we let $\boldsymbol{\theta} = \text{vec}(\mathbf{A}, \mathbf{B}_2)$ be a vector of all the free parameters, which has in total $q = KL + K(N - K) = K(N - K + L)$ elements. At the above analytical estimator provided by Theorem 2, we can evaluate $\mathbf{g}_T$, $\mathbf{D}_T$, and the $\mathbf{S}_T$ matrix as given by (12) and hence obtain H_z by (5). [$\mathbf{S}_T$ can be another consisitent estimator such as Newey and West's (1987).] Based on Theorem 1, H_z is asymptotically distributed χ^2 with degrees of freedom $(L - K)(N - K)$ under the very general heteroskedasticity conditions of Hansen (1982). This offers an analytical test[2] of the overidentification restrictions imposed on the model by the K-factor pricing theory.

The above analytical test H_z will be applied later to the U.S. equity market (Section 3), but it can also be applied to study bond returns, term structure models, forward currency premiums, international equity returns, and capital market integration, in the same manner as Campbell (1987), Cumby (1987), Chan (1988), Stambaugh (1988), Campbell and Hamao (1992), Ferson (1990), Chang, Pinegar, and Ravichandran (1991), Ferson and Foerster (1991), and others. H_z will be especially helpful in situations where it is difficult to obtain the traditional GMM test by numerical methods. For example, in their study of international capital markets, Harvey, Solnik, and Zhou (1992) use 44 asset returns and 8 instruments, giving rise to a GMM optimization problem with 100 parameters in a two-factor case. Because there are so many parameters, the authors cannot solve the GMM optimization problem numerically, but our analytical test can be

[2] Rob Stambaugh pointed out to the author a much simpler analytical two-step estimator: obtain first the fitted latent factors by using K unconstrained regressions and then run $N - K$ regressions on the fitted latent variables. However, this is not a GMM estimator. Nevertheless, it seems possible to derive a test based on this two-step estimator by using techniques advanced in Gallant and White (1988).

applied in a straightforward manner. In addition, Zhou (1993a) shows that latent variables models provide a convenient and powerful framework for testing asset pricing models with constant risk premiums because many of such models imply similar rank restrictions. Therefore, the results of this study may also be used to examine the validity of constant risk premium models. Examples include the APT and the multibeta pricing models.

2.3 Relationship with the maximum likelihood approach

If the model residuals are not only i.i.d., but also normally distributed with mean zero and a constant nonsingular covariance matrix, a maximum likehood (ML) approach may be used. Gibbons and Ferson (1985) are the first to use such an approach to test asset pricing theory in latent variables models. However, their inference is based on an asymptotic test and the estimation is done by nonlinear numerical maximization. As an extension of their work, Zhou (1993a) provides an exact Wald test, which is computed analytically from the ordinary least square (OLS) estimator. The Wald test is also shown to be equivalent to the likelihood ratio test in small sample, making it straightforward to use the ML approach in latent variables models. Although it seems difficult to derive the exact distribution of the constrained ML estimator, the asymptotic distribution can be obtained from the present GMM framework. By using Theorem 2 and its proof, we know that the constrained ML estimator is a GMM estimator with the past regressors as instruments, and so it is asymptotically normal with the asymptotic covariance matrix computed as usual.

3. Empirical Results

In this section, we apply our testing method to examine the number of "priced" latent factors in the U.S. equity market. The asset returns used in the K-factor model are 46 portfolios of monthly stock returns that are consistently available from October 1941 to September 1986 and are from the CRSP data base (the Center for Research in Security Prices at the University of Chicago). The portfolios are value-weighted, grouped by the stock's first two-digit standard industry classification (SIC) code. By arbitrarily taking the first group (in terms of the SIC code) as the first asset, there are 45 industry returns in excess of the first group and $T = 540$ observations.

There are two sets of instrumental variables that are used in our test. The first is a *small instrument set,* which contains three variables: a constant, the lagged return on the equal-weighted index in excess of the 30-day Treasury bill rate (market premium), and the lagged monthly return on a 90-day bill in excess of the 30-day bill rate (term

premium). The second is a *large instrument set* that combines the small instrument set and three other variables: a dummy variable for January; the lagged yield on Moody's BAA-rated bonds minus the yield on Moody's AAA-rated bonds (junk bond premium); and the lagged dividend yield on the Standard and Poor's Composite Stock Price Index minus the return on a 30-day bill (dividend yield spread). Fama (1984) and Campbell (1987) find that measures of the interest rate term premium can predict monthly stock returns. Fama and French (1988, 1989) and Campbell and Shiller (1988), among others, show that the dividend yield spread has power in forecasting returns. Keim and Stambaugh (1986) demonstrate the usefulness of junk bond premiums in predicting stock returns.

To test the K-factor theory, we first need to estimate the parameters of the return generating process (10) under the null hypothesis that the K-factor theory is valid. The estimation is done in three steps. First, an estimator is computed by using our analytical solution (Theorem 2) with the weighting matrix $\mathbf{W}_T$ being the identity matrix. Second, the residuals computed with this estimator are used to compute $\mathbf{S}_T$ using (13). Finally, the weighting matrix $\mathbf{W}_T = \mathbf{S}_T^{-1}$ is used to obtain a second-round analytical estimator. One can indeed verify the result that the derivatives of the objective function with respect to all of the normalized parameters are zero. Based on this second-round analytical estimator, the GMM test H_z is evaluated from (5). For comparison, we also compute the conventional GMM test, H_0. As noted earlier, H_0 has to be solved numerically for general heteroskedastic model residuals.[3] However, as we emphasized previously, the numerical optimization is not always possible. Indeed, we often fail to find convergent solutions for larger values of N, so H_0 is reported only when it is available.

In our testing, we analyze not only the full set of the $N = 45$ excess returns, but also some subsets. In deciding on the subsets, we rank the excess returns by their means and choose the most disperse group possible. For example, assets whose ranks are 1, 5, 10, 15, 20, 25, 30, 35, 40, and 45 are chosen as a subset with $N = 10$ assets. For $N = 20$ ($N = 30$), we choose the assets in the $N = 10$ ($N = 20$) group, plus 10 new assets. This way, we obtain four subsets out of the 45 assets with $N = 10$, 20, 30, and 40 being the number of assets in each of the subsets.

Table 1 provides the results. In the case where the small instrument set is used, panel A reports H_z, H_0, and H_{iid}, where H_{iid} is the con-

[3] The GMM minimization problem is solved by using the optimization program of Shanno and Phua (1980). As with any such program, an initial value of the GMM estimator must be supplied. The second-round analytical estimator is used as the starting point.

ventional GMM test statistic computed under an i.i.d. assumption. Note that H_{iid} is provided by Theorem 2 under the i.i.d. assumption. The associated p-values based on the asymptotic χ^2 distributions are in parentheses. When the null hypothesis that there is one "priced" latent factor is tested against the alternative that there are more than one factor, the p-values from these three tests are 0, 9, and 9.5 percent for the $N = 10$ assets case. Notice that the difference between H_{iid} and H_0 lies in the i.i.d. assumption imposed on H_{iid}. If the i.i.d. assumption is true, the numerical values of H_{iid} and H_z are close, as will be demonstrated below in Section 4. Thus, it seems that the low p-value from H_{iid} is due to the presence of heteroskedasticity, and hence the test of the null hypothesis should be based on H_z or H_0. Because the p-values from H_z and H_0 are 9 and 9.5 percent, we cannot reject the one-factor hypothesis for the $N = 10$ assets case at the usual 5 percent significance level. When the null hypothesis of $K = 2$ is examined, all of the p-values are greater than 40 percent, and we cannot reject the null at all. As the number of assets increases from 10 to 20, the p-values become smaller. However, when N increases from 20 to 30 and from 30 to 40, they become slightly larger. Overall, a one-factor model gets rejected at the 5 percent level using 20, 30, 40, and 45 assets, but a two-factor model cannot be rejected in any of the cases.

To study the sensitivity to instruments, we repeat the tests by using the large instrument set. The results are reported in panel B. The p-values are generally smaller than those in the small instrument case, but the conclusion is basically the same. That is, the one-factor hypothesis is rejected at the usual 5 percent level for the 20, 30, and 40 assets cases, but the two-factor model is not rejected. Of course, the failure to reject a two-factor model may be due to the tests having low power, but this issue must await further research.

4. Finite Sample Properties

In this section, we study the finite sample properties of H_0 and H_z in the context of the latent variable models. In contrast to numerical procedures, the analytical solution makes it feasible to study the finite sample properties of H_z in both the i.i.d. case and the heteroskedasticity case. Recall that Theorem 2 gives rise to the optimal estimator in the i.i.d. case, and hence we also obtain the conventional GMM test H_{iid}, which is H_0 when the i.i.d. assumption is imposed. Therefore, in the i.i.d. case, we can easily perform simulations for both H_z and H_{iid}. In the heteroskedasticity case, however, H_z is obtained analytically, but H_0 is not . Without the analytical solution, there are at least two difficulties in performing simulations for H_0. First, it is difficult

Table 1
The number of latent factors

		H_{nd}	H_z	H_0
		A: Small instrument set ($L = 3$)[1]		
$N = 10$	$K = 1$	47.986	26.435	26.206
		(0.000)	(0.090)	(0.095)
	$K = 2$	8.201	8.325	8.022
		(0.414)	(0.402)	(0.431)
$N = 20$	$K = 1$	78.719	61.433	69.693
		(0.000)	(0.009)	(0.001)
	$K = 2$	26.137	27.261	27.707
		(0.097)	(0.074)	(0.067)
$N = 30$	$K = 1$	94.401	83.322	83.336
		(0.002)	(0.016)	(0.016)
	$K = 2$	32.258	32.823	32.228
		(0.264)	(0.242)	(0.265)
$N = 40$	$K = 1$	118.476	104.340	*[2]
		(0.002)	(0.025)	*
	$K = 2$	42.989	44.516	*
		(0.266)	(0.217)	*
$N = 45$	$K = 1$	123.844	112.141	*
		(0.007)	(0.042)	*
	$K = 2$	46.469	48.723	*
		(0.331)	(0.254)	*
		B: Large instrument set ($L = 6$)		
$N = 10$	$K = 1$	87.738	58.202	57.791
		(0.000)	(0.090)	(0.096)
	$K = 2$	44.447	35.897	35.290
		(0.071)	(0.291)	(0.315)
$N = 20$	$K = 1$	151.468	121.787	13.567
		(0.000)	(0.033)	(0.850)
	$K = 2$	94.162	84.601	84.770
		(0.041)	(0.147)	(0.144)
$N = 30$	$K = 1$	224.940	179.706	180.351
		(0.000)	(0.027)	(0.025)
	$K = 2$	146.971	130.796	*
		(0.015)	(0.108)	*
$N = 40$	$K = 1$	287.369	228.847	*
		(0.000)	(0.049)	*
	$K = 2$	194.670	175.553	*
		(0.011)	(0.093)	*
$N = 45$	$K = 1$	310.652	251.769	*
		(0.000)	(0.070)	*
	$K = 2$	214.588	197.742	*
		(0.015)	(0.087)	*

If the K-factor pricing theory is valid, the expected excess asset returns satisfy

$$E(R_{it}|\mathbf{Z}_{t-1}) = b_{1i}\lambda_1(\mathbf{Z}_{t-1}) + \cdots + b_{Ki}\lambda_K(\mathbf{Z}_{t-1}),$$

where $\lambda_j(\mathbf{Z}_{t-1})$ is the market wide expected risk premium on the jth factor; $\mathbf{Z}_{t-1}$, $L \times 1$, are the instruments representing marketwide information available at t; and $b_{1i}, \ldots, b_{Ki}$ are the conditional excess betas. The above pricing relationship is equivalent to a rank K restriction on $\boldsymbol{\Theta}$, an $L \times N$ regression coefficients matrix of the excess return-generating process:

$$R_{it} = \theta_{1i}Z_{t-1,1} + \cdots + \theta_{Li}Z_{t-1,L} + u_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T,$$

where u_{it}'s are the disturbances. The data are the monthly industry returns, and there are 45 excess asset returns with $T = 540$ observations.

to automate the process of obtaining hundreds and thousands of numerical solutions. Second, it is difficult to interpret those solutions that are the possible local minimums or the nonconvergent draws. As a result, we perform simulations only for H_z in the heteroskedasticity case.

Consider first the i.i.d. case where the residuals $\mathbf{U}_t$ are assumed to have a multivariate normal distribution with zero means and covariance matrix $\mathbf{\Sigma}$. Given the parameters $\boldsymbol{\theta}$ and $\mathbf{\Sigma}$, it is straightforward to generate the residuals and hence the returns. In other words, once the parameters $\boldsymbol{\theta}$ and $\mathbf{\Sigma}$ are prespecified, we can generate hundreds of sets of residuals and returns. With these artificial data, H_z and H_{iid} are easily computed and compared with their asymptotic χ^2 distribution.

It is, however, not an easy matter to generate the residuals in the heteroskedasticity case because a specific form of heteroskedasticity has to be specified. There are many possible specifications, but we use a simple one: let $\mathbf{Z}^*_{t-1}$ be a subset of $\mathbf{Z}_{t-1}$ such that we can assume $\mathbf{Z}^*_{t-1}$ and $\mathbf{U}_t$ are jointly multivariate t distributed with the degree of freedom ν and the nonsingular covariance matrix $\mathbf{V}$. Partition $\mathbf{V}$ as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix},$$

where $\mathbf{V}_{22} = \mathbf{\Sigma}$. To be consistent with both (10) and (11), $\mathbf{Z}^*_{t-1}$ and $\mathbf{U}_t$ must be uncorrelated, that is, $\mathbf{V}_{12} = \mathbf{V}_{21} = \mathbf{0}$. However, $\mathbf{Z}^*_{t-1}$ and $\mathbf{U}_t$ will not be independent. Indeed,

$$\text{Var}(\mathbf{U}_t \mid \mathbf{Z}^*_{t-1}) = c[1 + (\mathbf{Z}^*_{t-1} - \mathbf{Z}^*_m)'\mathbf{V}_{11}^{-1}(\mathbf{Z}^*_{t-1} - \mathbf{Z}^*_m)/(\nu - 2)]\mathbf{V}_{22}, \qquad (18)$$

where $\mathbf{Z}^*_m$ is the population mean of $\mathbf{Z}^*_{t-1}$ and $c = (\nu - 2)/(\nu - 1)$. Thus, the covariance matrix of the residuals are heteroskedastic or time-varying in the particular fashion of (18). As ν increases, the multivariate t distribution approaches the multivariate normal distribution and $\text{Var}(\mathbf{U}_t \mid \mathbf{Z}^*_{t-1})$ approaches the constant matrix $\mathbf{V}_{22}$, making the heteroskedasticity less important. In the extreme case of $\nu = +\infty$, it collapses to the i.i.d. case.

←

Panels A and B report the testing results by using a small and a large instrument set, respectively. Both H_{iid} and H_0 are the conventional GMM test statistics with and without the i.i.d. assumption, and H_z is the analytical GMM test statistic (Theorem 1). Under the null hypothesis that the K-factor model is true, all the tests are asymptotically χ^2 distributed with $(L - K)(N - K)$ degrees of freedom. The p-values are reported underneath the test statistics.

[1] The *small instrument set* is $\{Z_C, Z_{\text{EW}}, Z_{\text{TP}}\}$, and the *large instrument set* contains all the instrumental variables.

[2] The traditional GMM test statistics and the associated p-values are not reported in these cases because of the failure of convergence of the numerical procedures used.

Table 2
Finite sample properties of GMM tests

	$T = 60$	$T = 120$	$T = 240$	$T = 480$
	A: Rejection rate of H_{iid} (size = 5 percent) under i.i.d.[1]			
$N = 10$	0.035	0.044	0.047	0.050
$N = 20$	0.019	0.033	0.041	0.044
$N = 30$	0.006	0.029	0.041	0.044
$N = 40$	0.001	0.021	0.041	0.047
	B: Rejection rate of H_z (size = 5 percent) under i.i.d.[2]			
$N = 10$	0.032	0.045	0.050	0.049
$N = 20$	0.019	0.034	0.042	0.046
	C: Rejection rate of H_z (size = 5 percent) under heteroskedasticity[2]			
$N = 10$	0.085	0.065	0.054	0.046
$N = 20$	0.058	0.074	0.042	0.034

The table provides simulation evidence of the finite sample properties of both the conventional GMM test (H_{iid}) with the i.i.d. assumption and the proposed analytical GMM test (H_z). At the fixed test size of 5 percent determined from their asymptotic χ^2 distributions, the rejection rates of the tests are computed for an array of the number of assets (N) and the sample size (T).

[1] Based on 10,000 simulated data sets.

[2] Based on 5,000 simulated data sets.

In the simulation study that follows, we fix, for simplicity, the size of the tests at 5 percent.[4] In addition, we let $L = 3$, $K = 1$, and $\nu = 8$ throughout.[5] But we allow both the number of assets (N) and the sample size (T) to vary over a number of plausible values. All parameter values are set equal to those estimated in Section 3. Specifically, $\boldsymbol{\theta}$ is set equal to the second-round analytical estimates, and $\boldsymbol{\Sigma}$ is taken as the sample covariance matrix of the fitted residuals. In addition, $\mathbf{Z}^*_{t-1}$ is taken as the small instrument set excluding the constant, and its sample mean and covariance matrix are taken as $\mathbf{Z}^*_m$ and $\mathbf{V}_{11}$. With these specifications, it is straightforward to carry out our Monte Carlo study.

Table 2 provides the results. Panel A reports the rejection rate of H_{iid} based on 10,000 simulated data sets.[6] When there are only 10 assets and the sample size equals 60, the rejection rate from the 10,000 runs is 3.5 percent as compared with the true size of 5 percent. As the sample size increases to 120, 240, and 480, the rejection rates rise to 4.4, 4.7, and 5 percent. So, for the H_{iid} test in the 10-asset case, the

[4] A size of 10 percent gives rise to similar results. For example, at the 10 percent size, the first row of Table 2 would read as: 0.0846, 0.0937, 0.0967, 0.1005.

[5] The first entry of Table 2 would be 0.032 for an L value of 6. A value of $\nu = 8$ is shown by Zhou (1993b) to model residual nonnormality reasonably well, and the simulation results will not change substantially if a value of ν other than 8 (say $\nu = 6$) is chosen.

[6] For a year after the publication of this article, a Fortran program of the simulation and other applications of the article will be available from the author through e-mail (Zhou@Zhoufin.wustl.edu) upon request.

asymptotic size and the finite-sample size are remarkably similar; they are close even with a sample size as small as 60. When there are 20 assets, the rejection rates are 1.9, 3.3, 4.1, and 4.4 percent for sample sizes of 60, 120, 240, and 480. In comparison with the 10-asset case, a greater sample size is needed to get as accurate an inference. When there are 40 assets, the rejection rates are 0.1, 2.1, 4.1, and 4.7 percent for the four sample sizes. At $T = 240$, the rejection rate is 4.1 percent, suggesting that a sample size of 240 is sufficient to obtain accurate p-values for a system of as many as 40 assets. In their studies of the finite sample properties of H_{iid}, Ferson and Foerster (1991) provide simulation results for N only up to 14 assets, and hence our results complement theirs, showing that H_{iid} produces reliable statistical inference for a sample size of 240 and for as many as 40 assets.

In contrast to H_{iid}, the computation of H_z is far more time consuming because it requires the evaluation of the complex $\mathbf{\Omega}_T$ matrix and its eigenvalues and eigenvectors. Thus, we consider only up to 20 assets and use only 5,000 data sets to obtain the simulation evidence. In the i.i.d. case, the results are reported in panel B. When there are only 10 assets, the rejection rates are 3.2, 4.5, 5.0, and 4.9 percent for sample sizes of 60, 120, 240, and 480. In comparison with those from the H_{iid} test, the differences in the numerical values of the rejection rates are too small to make any substantial differences in inference. This is also true when there are 20 assets. Indeed, for any given value of N and T in Table 2, the numerical value of H_z is virtually identical to that of H_{iid} in each of the 5,000 simulations, where the data are the first 5,000 of the 10,000 data sets used earlier.[7] As a result, we find that H_z produces inferences similar to those from H_{iid}, at least for those cases provided in the table.

In the heteroskedasticity case, the simulation results for H_z are provided in panel C. When there are only 10 assets, the rejection rates are 8.5, 6.5, 5.4, and 4.6 percent for sample sizes of 60, 120, 240, and 480. In contrast to the i.i.d. case, these results show some tendency of slight overrejection, especially when the sample size is small, say $T = 60$. However, as the sample size increases, the rejection rates get much closer to the 5 percent level. When there are 20 assets, we obtain similar conclusions to the 10-asset case. Thus, despite the heteroskedasticity, H_z still produces fairly accurate rejection rates.

In summary, our simulations show that the p-values based on the asymptotic distributions of H_{iid} and H_z are close to their finite sample p-values, suggesting that both H_{iid} and H_z may be reliable for many

[7] The same result between H_z and H_0 also seems true, but we have not established it because simulations for H_0 under heteroskedasticity cannot be easily done without analytical solutions. However, we do find almost identical values between the two from our applications in Section 3.

empirical applications.[8] However, there are certain limitations of our results. First, due to the computational time required, the simulations for H_z are done only up to 20 assets. Thus, one still needs to be careful when applying the tests to a model of a large number of assets. Second, our simulations provide evidence about the finite sample performance of the tests only under the null hypothesis. The properties of the tests under the alternative hypothesis remain unknown. Nevertheless, with faster computers, it is possible to carry out similar simulations for up to 100 or more assets and for up to 10 or more factors, and it may also be possible to study the power of the tests. Both of the issues seem to be interesting topics for future research.

5. Conclusions

We propose alternative GMM tests that are analytically solvable in many econometric models, yielding in particular analytical GMM tests for latent variables models. Because such models arise from many studies of stock returns, term structure theories, forward currency premiums, international equity returns, capital market integration, and the reduction of factors, it is likely that there are wide applications of our results. In addition, we provide simulation evidence showing that the proposed tests have good finite sample properties and that their asymptotic distribution is reliable for the sample size commonly used. In contrast to the conventional GMM test, our tests can be obtained analytically. This overcomes the difficulty in applying the traditional GMM test for which the iterated solution to the nonlinear GMM minimization problem may not converge to the global minimum or even converge at all. Many previously difficult estimation and testing problems, such as tests of the APT, the CCAPM, and the beta pricing models, could become more tractable by using the new GMM tests. The method of this article seems useful wherever the GMM approach is relevant.

Appendix A: Proof of Theorem 1

This theorem is a result based upon Hansen's (1982) Lemma 4.1. To obtain a χ^2 test, we need to diagonalize the asymptotic covariance matrix of $\sqrt{T}\mathbf{g}_T$, which is given by

$$\mathbf{\Omega}_0 \equiv \mathbf{W}_0^{-1/2}\mathbf{N}_0(\mathbf{W}_0^{1/2})\mathbf{N}_0\mathbf{W}_0^{-1/2}, \tag{A1}$$

[8] An additional analytical GMM test can be constructed following Newey (1985), but this test seems to reject the null hypothesis too often in finite samples. For example, if this test were used, the first entry of Table 2 would be 56 percent. Even when $T = 480$, the rejection rate from this test is still 43.47 percent, well away from the 5 percent level.

with

$$\mathbf{N}_0 = \mathbf{I} - \mathbf{W}_0^{1/2}\mathbf{D}_0(\mathbf{D}_0'\mathbf{W}_0\mathbf{D}_0)^{-1}\mathbf{D}_0'\mathbf{W}_0^{1/2}, \tag{A2}$$

where variables with subscript 0 indicate that they are evaluated at the true population parameters. It is clear that $\mathbf{N}_0$ is idempotent and so it has rank d. This implies that $\mathbf{\Omega}_0$ has rank less than or equal to d. By the eigenvalue decomposition of $\mathbf{N}_0$, we know the rank of $\mathbf{\Omega}_0$ is exactly d. Because of the singularity of $\mathbf{\Omega}_0$, we use the following technique to diagonalize it. Let $\mu_1 \geq \ldots \geq \mu_d$ be the nonzero eigenvalues of $\mathbf{\Omega}_0$. Then there is a unique $\mathbf{M}_0$ such that

$$\mathbf{\Omega}_0 = \mathbf{M}_0'\mathrm{Diag}(\mu_1, \ldots, \mu_d, 0, \ldots, 0)\mathbf{M}_0, \tag{A3}$$

where $\mathbf{M}_0\mathbf{M}_0' = \mathbf{M}_0'\mathbf{M}_0 = \boldsymbol{I}$. In fact, the ith row of $\mathbf{M}_0$ is the standardized eigenvector corresponding to the ith largest eigenvalue μ_i, for $i = 1, \ldots, NL$. Therefore, the covariance matrix of $\sqrt{T}\mathbf{V}_0^{-1/2}\mathbf{M}_0\mathbf{g}_T$ has asymptotic covariance matrix $\mathrm{Diag}(1, \ldots, 1, 0, \ldots, 0)$. Finally, notice that $\mathbf{M}_0$ and $\mathbf{V}_0$ are continuous functions of the elements of $\mathbf{\Omega}_0$ and hence can be consistently estimated by their sample analogue. Thus, the theorem follows. Q.E.D.

Appendix B: Proof of Theorem 2

We derive in this Appendix the analytical solution to the GMM estimator for the following multivariate regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{U}, \tag{B1}$$

where $\mathbf{Y}$, $T \times N$, are the dependent variables; $\mathbf{X}$, $T \times M$, are the regressors; $\boldsymbol{\Theta}$, $M \times N$, is the regression coefficient matrix; and $\mathbf{U}$ is the disturbance matrix. Notice that $\mathbf{X}$ may contain contemporaneous variables as well as some or all of the information variables. Equation (10) is a special case of (B1) with $\mathbf{Y} = \mathbf{R}$, $\mathbf{X} = \mathbf{Z}$, and $M = L$. Following Hansen (1982), we obtain the GMM estimator by minimizing:

$$\min Q = \mathbf{g}_T'\mathbf{W}_T\mathbf{g}_T, \qquad \mathbf{g}_T \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t, \qquad NL \times 1 \tag{B2}$$

where $\mathbf{f}_t = \mathbf{U}_t \otimes \mathbf{Z}_{t-1}$; $\mathbf{W}_T$, $NL \times NL$, is the weighting matrix; $\mathbf{U}_t$, $N \times 1$, is the model residuals at time t; and $\mathbf{Z}_{t-1}$, $L \times 1$, the instruments. Let $\mathbf{Z}$ be a $T \times L$ matrix of the instruments ($L \geq M$). It is seen that Theorem 2 is a special case of the following:

Theorem 2*. *If the weighting matrix is of the form*

$$\boldsymbol{W}_T \equiv \boldsymbol{W}_1 \otimes \boldsymbol{W}_2, \qquad \boldsymbol{W}_1\text{: } N \times N, \qquad \boldsymbol{W}_2\text{: } L \times L,$$

then the GMM estimator of $\boldsymbol{\Theta}$ *under the rank K restriction, that the rank of* $\boldsymbol{\Theta}$ *is K, is explicitly given by*

$$\hat{\boldsymbol{\Theta}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}, \qquad \tilde{\boldsymbol{A}}: M \times K, \qquad \tilde{\boldsymbol{B}}: K \times N, \tag{B3}$$

where

$$\tilde{\boldsymbol{A}} = (\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}/T^2)^{-1/2}\boldsymbol{E}, \qquad \tilde{\boldsymbol{B}} = (\boldsymbol{X}^{*\prime}\boldsymbol{P}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{P}\boldsymbol{Y},$$

$\boldsymbol{P} = \boldsymbol{Z}\boldsymbol{W}_2\boldsymbol{Z}'$, $\boldsymbol{X}^* = \boldsymbol{X}\tilde{\boldsymbol{A}}$ *and* $\boldsymbol{E}$ *is the* $M \times K$ *matrix stacked by the "standardized" eigenvectors* ($\boldsymbol{E}'\boldsymbol{E} = \boldsymbol{I}_K$) *corresponding to the K largest eigenvalues of the* $M \times M$ *matrix:*

$$(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}/T^2)^{-1/2}(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{Y}/T^2)\boldsymbol{W}_1(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{Y}/T^2)'(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}/T^2)^{-1/2}. \tag{B4}$$

Furthermore, the minimum of Q *is* $Q^* = tr\boldsymbol{W}_1(\boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y}/T^2) - \gamma_1 - \ldots - \gamma_K$, *where* $\gamma_1, \ldots, \gamma_K$ *are the K largest eigenvalues of the matrix given in (B4).*

Proof. Let $\mathbf{G}_T = \mathbf{Z}'\mathbf{U}/T$. Then $\mathbf{g}_T = \mathrm{vec}(\mathbf{G}_T)$, and the objective function can be written

$$Q = [\mathrm{vec}(\mathbf{G}_T)]'(\mathbf{W}_1 \otimes \mathbf{W}_2)[\mathrm{vec}(\mathbf{G}_T)] = \mathrm{tr}(\mathbf{W}_1\mathbf{G}_T'\mathbf{W}_2\mathbf{G}_T)$$

$$= \frac{1}{T^2}\mathrm{tr}(\mathbf{W}_1\mathbf{U}'\mathbf{Z}\mathbf{W}_2\mathbf{Z}'\mathbf{U}) = \frac{1}{T^2}\mathrm{tr}(\mathbf{W}_1\mathbf{U}'\mathbf{P}\mathbf{U}),$$

where $\mathbf{P} \equiv \mathbf{Z}\mathbf{W}_2\mathbf{Z}'$. Under the null, we can write $\boldsymbol{\Theta}$ as

$$\boldsymbol{\Theta} = \mathbf{A}\mathbf{B}, \qquad \mathbf{A}: M \times K, \qquad \mathbf{B}: K \times N,$$

for suitable $\mathbf{A}$ and $\mathbf{B}$. Now, it is easy to verify that

$$\mathbf{U}'\mathbf{P}\mathbf{U} = (\mathbf{Y} - \mathbf{X}^*\mathbf{B})'\mathbf{P}(\mathbf{Y} - \mathbf{X}^*\mathbf{B})$$

$$= (\mathbf{Y} - \mathbf{X}^*\mathbf{B})'\mathbf{P}(\mathbf{Y} - \mathbf{X}^*\tilde{\mathbf{B}}) + (\mathbf{B} - \tilde{\mathbf{B}})'\mathbf{X}^{*\prime}\mathbf{P}\mathbf{X}^*(\mathbf{B} - \tilde{\mathbf{B}}),$$

where $\mathbf{X}^* \equiv \mathbf{X}\mathbf{A}$. Therefore, conditional on $\mathbf{A}$, the estimator of $\mathbf{B}$ is given by $\tilde{\mathbf{B}} = (\mathbf{X}^{*\prime}\mathbf{P}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{P}\mathbf{Y}$. Replacing $\mathbf{B}$ by $\tilde{\mathbf{B}}$, we get

$$\mathbf{U}'\mathbf{P}\mathbf{U} = (\mathbf{Y} - \mathbf{X}^*\tilde{\mathbf{B}})'\mathbf{P}(\mathbf{Y} - \mathbf{X}^*\tilde{\mathbf{B}})$$

$$= \mathbf{Y}'[\mathbf{P} - \mathbf{P}\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{P}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{P}]\mathbf{Y}.$$

Notice that Q is now a function of $\mathbf{A}$ alone. To minimize it, we normalize $\mathbf{A}$ such that $(\mathbf{X}^{*\prime}\mathbf{P}\mathbf{X}^*)/T^2 = \mathbf{I}_K$, or $\mathbf{A}'(\mathbf{X}'\mathbf{P}\mathbf{X})\mathbf{A} = T^2\mathbf{I}_K$. Then, we have

$$T^2Q = \mathrm{tr}(\mathbf{Y}'[\mathbf{P} - \mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}'\mathbf{P}/T^2]\,\mathbf{Y}\mathbf{W}_1)$$

$$= \mathrm{tr}(\mathbf{Y}'\mathbf{P}\mathbf{Y}\mathbf{W}_1) - \mathrm{tr}(\mathbf{Y}'\mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{P}\mathbf{Y}\mathbf{W}_1)/T^2,$$

so we need only to maximize

$$Q^{**} \equiv \mathrm{tr}(\mathbf{Y}'\mathbf{P}\mathbf{X}\mathbf{A}\mathbf{A}'\mathbf{X}'\mathbf{P}\mathbf{Y}\mathbf{W}_1) = \mathrm{tr}(\mathbf{A}'\mathbf{X}'\mathbf{P}\mathbf{Y}\mathbf{W}_1\mathbf{Y}'\mathbf{P}\mathbf{X}\mathbf{A})$$

$$= \mathrm{tr}[(\mathbf{S}^{1/2}\mathbf{A})'\mathbf{S}^{-1/2}\mathbf{X}'\mathbf{P}\mathbf{Y}\mathbf{W}_1\mathbf{Y}'\mathbf{P}\mathbf{X}\mathbf{S}^{-1/2}(\mathbf{S}^{1/2}\mathbf{A})],$$

where $\mathbf{S} \equiv (\mathbf{X}'\mathbf{P}\mathbf{X})/T^2$, $M \times M$, is positive definite if $L \geq M$, and $\mathbf{X}$ has rank M. Hence, applying the Poincare Separation Theorem the same way as in Zhou (1993a), the trace is maximized if $\mathbf{S}^{1/2}\mathbf{A} = \mathbf{E}$, and the maximum is given by the K largest eigenvalues of the $M \times M$ matrix $\mathbf{S}^{1/2}\mathbf{X}'\mathbf{P}\mathbf{Y}\mathbf{W}_1\mathbf{Y}'\mathbf{P}\mathbf{X}\mathbf{S}^{-1/2}$. Q.E.D.

References

Back, K., and D. Brown, 1993, "Implied Probabilities in GMM Estimators," *Econometrica,* 61, 971–975.

Black, F., 1972, "Capital Market Equilibrium with Restricted Borrowing," *Journal of Business,* 45, 444–454.

Breeden, D. T., 1979, "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics,* 7, 265–296.

Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger, 1989, "Empirical Tests of the Consumption Based CAPM," *Journal of Finance,* 44, 231–262.

Campbell, J. Y., 1987, "Stock Returns and the Term Structure," *Journal of Financial Economics,* 18, 373–399.

Campbell, J. Y., and Y. Hamao, 1992, "Predictable Bond and Stock Returns in the United States and Japan: A Study of Long-term Capital Market Integration," *Journal of Finance,* 47, 43–69.

Campbell, J. Y., and R. J. Shiller, 1988, "The Dividend Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies,* 1, 195–228.

Chan, K. C., 1988, "Autocorrelation of Stock Returns and Market Efficiency," working paper, Ohio State University.

Chang, E. C., J. M. Pinegar, and R. Ravichandran, 1991, "Latent Variables Tests of the Integration of European Equity Markets," working paper, University of Maryland.

Cumby, R. E., 1987, "Is It Risk? Explaining Deviations from Uncovered Interest Parity," *Journal of Monetary Economics,* 22, 279–299.

Fama, E. F., 1984, "The Information in the Term Structure," *Journal of Financial Economics,* 13, 509–528.

Fama, E. F., 1991, "Efficient Capital Markets: II," *Journal of Finance,* 46, 1575–1617.

Fama, E. F., and K. R. French, 1988, "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics,* 22, 3–25.

Fama, E. F., and K. R. French, 1989, "Business Conditions and Expected Stock Returns," *Journal of Financial Economics,* 25, 23–50.

Ferson, W. E., 1990, "Are the Latent Variables in Time-varying Expected Returns Compensation for Consumption Risk?" *Journal of Finance,* 45, 397–430.

Ferson, W. E., and S. R. Foerster, 1994, "Finite Sample Properties of Methods of Moments in Latent Variable Tests of Asset Pricing Models," *Journal of Financial Economics,* 36, 29–55.

Ferson, W. E., S. R. Foerster, and D. Keim, 1993, "General Tests of Latent Variable Models and Mean Variance Spanning," *Journal of Finance,* 48, 131–156.

Ferson, W. E., and C. R. Harvey, 1991, "The Variation of Economic Risk Premiums," *Journal of Political Economy,* 99, 385–415.

Gallant, A. R., and H. White, 1988, *A United Theory of Estimation and Inference for Nonlinear Dynamic Models,* Basil Blackwell, New York.

Gibbons, M. R., and W. Ferson, 1985, "Testing Asset Pricing Models with Changing Expectations and an Unobservable Market Portfolio," *Journal of Financial Economics,* 14, 217–236.

Gibbons, M. R., S. A. Ross, and J. Shanken, 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., and R. J. Hodrick, 1983, "Risk Averse Speculation in the Forward Foreign Exchange Market: An Econometric Analysis of Linear Models," in J. A. Frenkel (ed.), *Exchange Rates and International Macroeconomics,* University of Chicago Press, Chicago.

Harvey, C., 1989, "Time-varying Conditional Covariances in Tests of Asset Pricing Models," *Journal of Financial Economics,* 24, 289–317.

Harvey, C., 1991, "World Price of Covariance Risk," *Journal of Finance,* 46, 111–157.

Harvey, C., B. Solnik, and G. Zhou, 1992, "What Determines International Asset Returns?" working paper, Duke University.

Judge, G. G., et al., 1985, *The Theory and Practice of Econometrics* (2d ed.), Wiley, New York.

Kandel, S., and R. Stambaugh, 1990, "A Mean-Variance Framework for Tests of Asset Pricing Models," *Review of Financial Studies,* 2, 125–156.

Keim, D. B., and R. F. Stambaugh, 1986, "Predicting Returns in the Bond and Stock Market," *Journal of Financial Economics,* 17, 357–390.

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.

Markowitz, H. M., 1952, "Portfolio Selection," *Journal of Finance,* 7, 77–91.

Merton, R. C., 1973, "An Intertemporal Capital Asset Pricing Model," *Econometrica,* 41, 867–887.

Newey, W. K., 1985, "Generalized Method of Moments Specifications Testing," *Journal of Econometrics,* 29, 229–256.

Newey, W. K., and K. D. West, 1987, "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica,* 55, 703–708.

Ross, S. A., 1976, "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory,* 13, 341–360.

Shanken, J., 1985, "Multivariate Tests of the Zero-beta CAPM," *Journal of Financial Economics,* 14, 327–348.

Shanken, J., 1986, "Testing Portfolio Efficiency When the Zero-beta Rate Is Unknown: A Note," *Journal of Finance,* 41, 269–276.

Shanken, J., 1990, "Intertemporal Asset Pricing: An Empirical Investigation," *Journal of Econometrics,* 45, 99–120.

Shanken, J., 1992, "On the Estimation of Beta-pricing Models," *Review of Financial Studies,* 5, 1–33.

Shanno, D. F., and K. H. Phua, 1980, "Minimization of Unconstrained Multivariate Functions," *ACM Transactions on Mathematical Software,* 6, 618–622.

Sharpe, W., 1964, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance,* 19, 425–442.

Stambaugh, R., 1988, "The Information in Forward Rates: Implications for Models of the Term Structure," *Journal of Financial Economics,* 21, 41–70.

Zhou, G., 1991, "Small Sample Tests of Portfolio Efficiency," *Journal of Financial Economics,* 30, 165–191.

Zhou, G., 1993a, "Small Sample Rank Tests with Applications to Asset Pricing," working paper, Washington University in St. Louis.

Zhou, G., 1993b, "Asset Pricing Tests under Alternative Distributions," *Journal of Finance,* 48, 1925–1942.