Hansen-Jagannathan Distance: Geometry and Exact Distribution

RAYMOND KAN and GUOFU ZHOU*

First draft: April, 2002
This version: July, 2004

*Kan is from the University of Toronto, Zhou is from Washington University in St. Louis. We thank Kerry Back, Devraj Basu, Jeremy Berkowitz, Tim Bollerslev, Douglas Breeden, Nai-fu Chen, Sean Cleary, Phil Dybvig, Heber Farnsworth, Christopher Gadarowski, Yaniv Grinstein, Campbell Harvey, Yongmiao Hong, Pete Kyle, Bruce Lehmann, Haitao Li, Ludan Liu, Kasing Man, Stephen Ross, Jay Shanken, Kevin Wang, Chu Zhang, seminar participants at Atlanta Fed, Cornell University, Duke University, Hong Kong University, McGill University, Penn State University, Queen's University, Simon Fraser University, Syracuse University, University of California at Irvine, University of Toronto, Washington University in St. Louis, York University, participants at the 2002 International Finance Conference at Tsinghua University, 2002 Northern Finance Meetings, 2003 SBFSIF conference, 2004 Western Finance Meetings, and especially Wayne Ferson and Ravi Jagannathan for helpful discussions and comments. We would also like to thank Robert Hodrick and Xiaoyan Zhang for sharing their data set with us, and Robert Chen for his research assistance. Kan gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of Canada.

Hansen-Jagannathan Distance: Geometry and Exact Distribution

ABSTRACT

This paper provides an in-depth analysis of the Hansen-Jagannathan (HJ) distance, a measure that is widely used for the diagnosis of asset pricing models and model selection. We provide a geometric interpretation of the HJ-distance in the mean and standard deviation space of portfolio returns. In relation to the traditional regression approach of testing beta asset pricing models, we show that the sample HJ-distance is a scaled version of Shanken's (1985) cross-sectional regression test (CSRT) statistic, with the major difference being the way the zero-beta rate is estimated. Simulation evidence shows that for a typical length of time series, the asymptotic distribution for the sample HJ-distance is grossly inappropriate when the number of factors or the number of test assets is large. We provide an exact distribution of the sample HJ-distance under both the null and the alternative hypotheses. In addition, we also suggest a simple numerical procedure for computing its distribution function.

Asset pricing models are at best approximations. Therefore, although it is of interest to test whether a particular asset pricing model is literally true or not, a more interesting task for empirical researchers is to find out how wrong a model is and to compare the performance of different asset pricing models. For the latter task, we need to establish a scalar measure of model misspecification. While there are many reasonable measures that can be used, the one recently introduced by Hansen and Jagannathan (1997) has gained tremendous popularity in the empirical asset pricing literature. Their proposed measure, called the HJ-distance, has been used both as a model diagnostic and as a tool for model selection by many researchers. Examples include Jagannathan and Wang (1996), Jagannathan, Kubota, and Takehara (1998), Campbell and Cochrane (2000), Lettau and Ludvigson (2001), Hodrick and Zhang (2001), Farnsworth, Ferson, Jackson, and Todd (2002), and Dittmar (2002), among others.

In this paper, we attempt to provide an improved understanding of the HJ-distance by focusing on the case of linear beta pricing models. In order to gain some intuition on what the HJ-distance is attempting to measure, we provide a geometric interpretation of the HJ-distance in terms of the minimum-variance frontiers of the test assets and the factor mimicking positions. We then provide a comparison of the HJ-distance with the cross-sectional regression test (CSRT) statistic of Shanken (1985) and the GMM over-identification test statistic of Hansen (1982). This comparison allows us to better understand how the HJ-distance is different from the traditional test statistics.

While the HJ-distance has emerged as one of the most dominant measures of model misspecification, the understanding of the statistical behavior of the sample HJ-distance appears to be poor. In most cases, statistical inference on the HJ-distance is based on its asymptotic distribution. Little is known about the finite sample distribution of the sample HJ-distance. However, asymptotic distribution can be grossly misleading. For example, using simulation evidence, Ahn and Gadarowski (2004) find that the asymptotic distribution of the sample HJ-distance rejects the correct model too often. Therefore, it is important for us to obtain the finite sample distribution of the HJ-distance.

Under the normality assumption, we present an exact distribution of the sample HJ-distance under both the null that the model is correctly specified and under the alternatives that the model is misspecified. Our analysis on the exact distribution not only helps us understand what the parameters determining the distribution are, it also provides us a simple numerical method to compute

the distribution in practice. Moreover, this analysis allows us to understand how the asymptotic distribution of sample HJ-distance differs from the exact distribution and the circumstances that lead to under-rejection or over-rejection problems for the asymptotic test. As a by-product of this analysis, we also provide an approximate F-test that has better finite sample properties than the asymptotic test.

We present simulation evidence to verify our finite sample distribution and to determine the size problem of the asymptotic test and the approximate F-test. We also perform a simulation experiment to examine the ability of the sample HJ-distance to tell good models apart from bad ones. We find that the HJ-distance has a tendency to prefer noisy factors and is not always reliable in telling good models apart from bad ones in finite samples.

The rest of the paper is organized as follows. The next section discusses the population measure of the HJ-distance and its justification as a measure of model misspecification as well as contrasts the HJ-distance with the traditional measure of model misspecification, and illustrates why they could generate very different rankings of competing models. Section II provides the sample HJ-distance and its geometric interpretation, together with a comparison with traditional specification test statistics. Section III provides the finite sample distribution of the sample HJ-distance. Section IV presents simulation evidence. The final section concludes our findings and the Appendix contains proofs of all propositions.

I. Population Measures of Model Misspecification

A. HJ-Distance and Traditional Measure of Model Misspecification

When an asset pricing model is misspecified, one is often interested in obtaining a measure of how wrong the model is. Hansen and Jagannathan (1997) suggest that a natural measure of misspecification is the minimum distance between the stochastic discount factor of an asset pricing model and a set of correct stochastic discount factors. Define y as the stochastic discount factor associated with an asset pricing model, and $\mathcal{M}$ as the set of stochastic discount factors that price

all the assets correctly. The HJ-distance is defined as[1]

$$\delta = \min_{m \in \mathcal{M}} \|m - y\|, \tag{1}$$

where $\|X\| = E[X^2]^{\frac{1}{2}}$ is the standard L^2 norm. The HJ-distance can also be interpreted as a measure of the maximum pricing error of a portfolio that has a unit second moment. Define ξ as the random payoff of a portfolio. Hansen and Jagannathan (1997) show that

$$\delta = \max_{\|\xi\|=1} |\pi(\xi) - \pi^y(\xi)|, \tag{2}$$

where $\pi(\xi)$ and $\pi^y(\xi)$ are the prices of ξ assigned by the true and the proposed asset pricing model, respectively.

To provide analytical insights in this paper, we focus on the class of linear stochastic discount factor models. Suppose the stochastic discount factor y is a linear function of K common factors f, given by

$$y(\lambda) = \lambda_0 + f'\lambda_1 = x'\lambda, \tag{3}$$

where $x = [1,\ f']'$ and $\lambda = [\lambda_0,\ \lambda_1']'$.[2] If the stochastic discount factor $y(\lambda)$ prices all the assets, then the price of a vector of test assets, q, must obey

$$E[px'\lambda] = q, \tag{4}$$

where p is the random payoff of the test assets at the end of the period. In particular, if p_i is the gross return on an asset, we have $q_i = 1$, and if p_i is the payoff of a zero cost portfolio, we have $q_i = 0$.

For a given value of λ, the vector of pricing errors of the test assets is given by

$$g(\lambda) = q - E[px'\lambda] = q - D\lambda, \tag{5}$$

where $D = E[px']$ and it is assumed to be of full column rank. It is well known that the squared HJ-distance has an explicit expression

$$\delta^2 = \min_{\lambda} g(\lambda)'U^{-1}g(\lambda) = q'[U^{-1} - U^{-1}D(D'U^{-1}D)^{-1}D'U^{-1}]q, \tag{6}$$

[1]Hansen and Jagannathan (1997) also define another measure of distance by restricting the admissible set of stochastic discount factors to be nonnegative. In this paper, we limit our attention to the HJ-distance as defined in (1).

[2]The linearity assumption here may not be as restrictive as it appears because f can contain power terms of the same common factor. For example, Bansal and Viswanathan (1993) and Dittmar (2002) write y as a polynomial of the market return.

where $U = E[pp']$ is assumed to be nonsingular. Note that while we can define the elements of p as either gross returns or excess returns, they cannot all be excess returns. Otherwise, q is a zero vector and δ will always be equal to zero.[3]

In many empirical studies, p is chosen to be the gross return on N test assets, denoted as R_2. Let $Y = [f',\ R_2']'$ and define its mean and variance as

$$\mu = E[Y] \equiv \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \tag{7}$$

$$V = \text{Var}[Y] \equiv \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}. \tag{8}$$

Using these notations, we can write $U = E[R_2 R_2'] = V_{22} + \mu_2 \mu_2'$ and $D = E[R_2 x'] = [\mu_2,\ V_{21} + \mu_2 \mu_1']$. Since the elements of R_2 are gross returns, we have $q = 1_N$ and it is easy to show that the λ that minimizes δ^2 is given by

$$\lambda^{HJ} = (D'U^{-1}D)^{-1}(D'U^{-1}1_N), \tag{9}$$

and hence the squared HJ-distance of (6) with $q = 1_N$ is

$$\delta^2 = 1_N'[U^{-1} - U^{-1}D(D'U^{-1}D)^{-1}D'U^{-1}]1_N. \tag{10}$$

We will focus our analysis on this expression for the rest of the paper because it is the one most widely used in literature. It may be noted that some researchers, for example, Hodrick and Zhang (2001), combine gross returns of some assets with excess returns of other assets as the payoffs vector p, but the results in this paper are equally applicable. The only modification is to replace the vector 1_N in our analysis by the vector q, where q is the cost of the N test assets. In fact, HJ-distance is invariant to repackaging of the test assets, so if we use $p^* = Ap$ instead of p as the payoffs of the test assets where A is a nonsingular square matrix, it can be easily shown that δ^2 in (6) remains unchanged. Therefore, the results in Hodrick and Zhang (2001) would remain the same if they were to use gross returns on all of their test assets as the payoffs vector p.

Before discussing the traditional measures of model misspecification, we first present an alternative expression for the HJ-distance. As it turns out, this alternative expression provides important insights on the differences between the HJ-distance and the traditional measures of model misspecification, which eventually leads to our geometrical interpretation of the HJ-distance.

[3]For the case where all the test assets are zero cost portfolios, we need to modify the definition of HJ-distance. Details of this modification are available upon request.

Lemma 1 *Let $\beta = V_{21}V_{11}^{-1}$ be the regression slope coefficient of regressing R_2 on a constant term and f, and $\Sigma = V_{22} - V_{21}V_{11}^{-1}V_{12}$ be the covariance matrix of the residuals and it is assumed to be nonsingular. Define the pricing errors as*

$$e_{HJ}(\eta) = 1_N - \mu_2\eta_0 - \beta\eta_1 = 1_N - H\eta, \tag{11}$$

where $H = [\mu_2,\ \beta]$ and $\eta = [\eta_0,\ \eta_1']'$, we have

$$\delta^2 = \min_{\eta} e_{HJ}(\eta)'\Sigma^{-1}e_{HJ}(\eta) = 1_N'[\Sigma^{-1} - \Sigma^{-1}H(H'\Sigma^{-1}H)^{-1}H'\Sigma^{-1}]1_N. \tag{12}$$

The lemma suggests that the squared HJ-distance can be expressed as an aggregate measure of the pricing errors in the generalized least squares (GLS) cross-sectional regression (CSR) of regressing 1_N on μ_2 and β. While the main purpose of presenting (12) is to facilitate comparison with other measures of model misspecification, this alternative expression also has practical value. In the standard way of computing HJ-distance, one needs to take the inverse of U. Some researchers (for example, Cochrane (1996)) find that taking the inverse of U is numerically unstable because all the elements of R_2 are close to one and the matrix U is close to singular. Our alternative way of computing HJ-distance will overcome this numerical problem because only Σ is inverted here, which is numerically more stable than inverting U.

One of the perceived advantages of HJ-distance over other specification tests is that it uses U^{-1} as the weighting matrix, which is model independent. Some readers may feel uncomfortable that we use Σ^{-1} as the weighting matrix in (12), which is model dependent. From the proof of Lemma 1, it is clear that when computing δ^2 in (10), the results are mathematically identical whether we use U^{-1}, V_{22}^{-1}, or Σ^{-1} as the weighting matrix. We choose to present our results using Σ^{-1} because it allows for easier comparisons with traditional asset pricing tests that often use Σ^{-1} as the weighting matrix.

Instead of expressing an asset pricing model in the form of a stochastic discount factor, earlier asset pricing theories, such as those of Sharpe (1964), Lintner (1965), Black (1972), Merton (1973), Ross (1976) and Breeden (1979), relate the expected return on a financial asset to its covariances (or betas) with some systematic risk factors. Under the K-factor beta pricing model, we have

$$\mu_2 = 1_N\gamma_0 + \beta\gamma_1 = G\gamma, \tag{13}$$

where $G = [1_N,\ \beta]$ and $\gamma = [\gamma_0,\ \gamma_1']'$. In the literature on beta pricing models, γ_0 is called the zero-beta rate and γ_1 is called the risk premium associated with the K factors.

When a beta pricing model is misspecified, (13) will not hold regardless of what values of γ are chosen. If we define the model errors on expected return as

$$e_{CS}(\gamma) = \mu_2 - G\gamma, \tag{14}$$

then a classical measure of model misspecification for a beta pricing model is the aggregate expected return errors

$$Q_C = \min_{\gamma} e_{CS}(\gamma)'\Sigma^{-1}e_{CS}(\gamma). \tag{15}$$

Assuming that G is of full column rank, it is easy to show that the γ that attains the minimum is given by

$$\gamma^{CS} = (G'\Sigma^{-1}G)^{-1}(G'\Sigma^{-1}\mu_2), \tag{16}$$

and we have[4]

$$Q_C = \mu_2'[\Sigma^{-1} - \Sigma^{-1}G(G'\Sigma^{-1}G)^{-1}G'\Sigma^{-1}]\mu_2. \tag{17}$$

Traditional specification tests of beta pricing models often rely on some transformation of the sample version of Q_C. These include, for example, the CSRT statistic developed by Shanken (1985) and the likelihood ratio test statistic of Shanken (1986). Comparing δ^2 and Q_C, we see that δ^2 is an aggregate measure of model errors on prices whereas Q_C is an aggregate measure of model errors on expected returns.

B. Geometrical Interpretation

While the interpretation of the HJ-distance as the maximum pricing error is intuitive, it is somewhat difficult to visualize. For the case of linear models, we present an alternative interpretation of the HJ-distance that is easy to visualize. We first define the payoffs of the K factor mimicking positions as $R_1 = WR_2$, where W is a $K \times N$ matrix obtained by projecting f on a constant term and R_2 as

$$f = w_0 + WR_2 + \epsilon_f, \tag{18}$$

[4]Following the proof of Lemma 1, we can replace Σ in the expression of Q_C by V_{22} without affecting the value of Q_C. The equivalence between using Σ and V_{22} in Q_C was first observed by Shanken (1985).

where R_2 and ϵ_f are uncorrelated with each other. It is easy to verify that $W = V_{12}V_{22}^{-1}$ and we have $R_1 = V_{12}V_{22}^{-1}R_2$. Although unnecessary, we assume $W1_N = V_{12}V_{22}^{-1}1_N \neq 0_K$ in our analysis for convenience, i.e., at least one of the mimicking positions is not a zero cost portfolio of the N risky assets. This is equivalent to assuming that the global minimum-variance portfolio of the N test assets has nonzero systematic risk.[5]

It is well known that the minimum-variance frontier of the N test assets is given by

$$\sigma_p^2 = \frac{a_2 - 2b_2\mu_p + c_2\mu_p^2}{a_2c_2 - b_2^2}, \tag{19}$$

where $a_2 = \mu_2'V_{22}^{-1}\mu_2$, $b_2 = \mu_2'V_{22}^{-1}1_N$, and $c_2 = 1_N'V_{22}^{-1}1_N$ are the usual efficiency set constants. The following lemma provides the minimum-variance frontier of the K mimicking positions.

Lemma 2 *Suppose $V_{12}V_{22}^{-1}1_N \neq 0_K$. For $K > 1$, the minimum-variance frontier of unit cost portfolios that are created using the K factor mimicking positions $R_1 = V_{12}V_{22}^{-1}R_2$ is given by*

$$\sigma_p^2 = \frac{a_1 - 2b_1\mu_p + c_1\mu_p^2}{a_1c_1 - b_1^2}, \tag{20}$$

where

$$a_1 = \mu_2'V_{22}^{-1}V_{21}(V_{12}V_{22}^{-1}V_{21})^{-1}V_{12}V_{22}^{-1}\mu_2 = \mu_2'V_{22}^{-1}\beta(\beta'V_{22}^{-1}\beta)^{-1}\beta'V_{22}^{-1}\mu_2, \tag{21}$$

$$b_1 = \mu_2'V_{22}^{-1}V_{21}(V_{12}V_{22}^{-1}V_{21})^{-1}V_{12}V_{22}^{-1}1_N = \mu_2'V_{22}^{-1}\beta(\beta'V_{22}^{-1}\beta)^{-1}\beta'V_{22}^{-1}1_N, \tag{22}$$

$$c_1 = 1_N'V_{22}^{-1}V_{21}(V_{12}V_{22}^{-1}V_{21})^{-1}V_{12}V_{22}^{-1}1_N = 1_N'V_{22}^{-1}\beta(\beta'V_{22}^{-1}\beta)^{-1}\beta'V_{22}^{-1}1_N. \tag{23}$$

For $K = 1$, the unit cost factor mimicking portfolio has a mean of b_1/c_1 and a variance of $1/c_1$.

Our first Proposition expresses the two measures of model misspecification, δ^2 and Q_C, in terms of Sharpe ratios of the two frontiers and also provides a characterization of the implied zero-beta rates chosen by these two measures.[6]

Proposition 1: *Define $\Delta a = a_2 - a_1$, $\Delta b = b_2 - b_1$, and $\Delta c = c_2 - c_1$, the squared HJ-distance (δ^2) and the aggregate expected return errors (Q_C) can be written as*

$$\delta^2 = \min_{\gamma_0} \frac{\theta_2^2(\gamma_0) - \theta_1^2(\gamma_0)}{\gamma_0^2} = \frac{\theta_2^2(\gamma_0^{HJ}) - \theta_1^2(\gamma_0^{HJ})}{(\gamma_0^{HJ})^2}, \tag{24}$$

[5] See Huberman, Kandel, and Stambaugh (1987) for a discussion of this assumption.

[6] Gibbons, Ross, and Shanken (1989) provide a similar geometrical interpretation for the specification test of the CAPM but our results differ from theirs in two important ways. First, the frontier here is in terms of gross returns but not excess returns and the value of the zero-beta rate is not explicitly specified by the model. Second, the factors here are not necessarily portfolio returns.

$$Q_C \quad = \quad \min_{\gamma_0} \theta_2^2(\gamma_0) - \theta_1^2(\gamma_0) = \theta_2^2(\gamma_0^{CS}) - \theta_1^2(\gamma_0^{CS}), \tag{25}$$

where $\gamma_0^{HJ} = \Delta a/\Delta b$, $\gamma_0^{CS} = \Delta b/\Delta c$, *and* $\theta_1(\gamma_0)$ *and* $\theta_2(\gamma_0)$ *are the Sharpe ratios of the tangency portfolio of the* K *mimicking positions and of the* N *test assets, respectively, when* γ_0 *is the y-intercept of the tangent line. If* $\Delta b \geq 0$, *we have* $\gamma_0^{HJ} \geq \gamma_0^{CS}$, *and if* $\Delta b < 0$, *we have* $\gamma_0^{HJ} \leq \gamma_0^{CS}$.

Note that γ_0 is defined as the expected gross return of the zero-beta asset, so when there is limited liability, γ_0^{CS} and γ_0^{HJ} are unlikely to be negative. Since Δa and Δc are positive, the more relevant case is $\Delta b > 0$ and we should expect $\gamma_0^{HJ} \geq \gamma_0^{CS} > 0$.

It is important to note that the HJ-distance does not choose a zero-beta rate to minimize the difference in the squared Sharpe ratios of the two tangency portfolios; instead, the zero-beta rate is chosen to minimize the difference in squared Sharpe ratios of the two tangency portfolios divided by the squared zero-beta rate. One may wonder why δ^2 and Q_C pick different zero-beta rates. It turns out that this difference originates from the difference in the focus between the traditional beta pricing models and the newer stochastic discount factor models. In the traditional beta pricing models, our focus is to try to find a zero-beta rate γ_0 and risk premium γ_1 to minimize the model errors of the expected returns on the N test assets, i.e., to minimize an aggregate of the following expected return errors

$$e_{CS} = \mu_2 - 1_N\gamma_0 - \beta\gamma_1. \tag{26}$$

However, in the stochastic discount factor approach, our focus is to obtain a linear combination of expected return and the betas of the N test assets to come up with a model price that is closest to their actual cost of 1_N, i.e., to minimize an aggregate of the following pricing errors (see Lemma 1 for this interpretation of the HJ-distance)

$$e_{HJ} = 1_N - \mu_2\eta_0 - \beta\eta_1, \tag{27}$$

where $\mu_2\eta_0 + \beta\eta_1$ is the price of the N assets predicted by the model. Using a reparameterization of $\gamma_0 = 1/\eta_0$ and $\gamma_1 = -\eta_1/\eta_0$, we can rewrite the pricing errors as

$$e_{HJ} = -\frac{1}{\gamma_0}(\mu_2 - 1_N\gamma_0 - \beta\gamma_1). \tag{28}$$

Comparing (26) with (28), we can see that the pricing errors differ from the expected return errors by a scale factor of $-1/\gamma_0$. Therefore, the γ_0 that minimizes δ^2 is in general different from the γ_0 that minimizes Q_C.

When the beta pricing model is correctly specified, the two frontiers touch each other at some point, and we have a unique γ_0 such that $\theta_2(\gamma_0) = \theta_1(\gamma_0)$.[7] In this case, we have $\gamma_0^{CS} = \gamma_0^{HJ}$. However, when the asset pricing model does not hold, γ_0^{CS} and γ_0^{HJ} are different. From (28), we can see that as γ_0 appears in the denominator of e_{HJ}, there is a tendency for the HJ-distance to choose a higher absolute value of zero-beta rate as a large value of γ_0 can deflate the pricing errors. Clearly, which choice of the zero-beta rate is more appropriate depends on whether the focus is to minimize errors on expected returns or errors on prices of the test assets.

C. Ranking Models

Although both δ and Q_C can be used to rank asset pricing models, these two measures can often lead to different rankings of competing models. Under the stochastic discount factor framework that the HJ-distance uses, one considers an asset pricing model to be a good one if it can explain the prices of the test assets well. More specifically, if one can find a linear combination of μ_2 and β that is close to 1_N (the actual price of the N assets), then the HJ-distance is small and the model will be considered a good model. However, it is important to note that an asset pricing model that explains prices well does not have to be a model that explains expected returns well. As an extreme case, suppose one finds a factor such that the betas are constant across all the test assets (i.e., $\beta \propto 1_N$). In that case, regardless of the values of the expected returns μ_2, β alone will fully explain 1_N and we will have zero pricing errors and zero HJ-distance.[8] However, the betas of such a factor are totally incapable of explaining expected returns μ_2 and as a result Q_C will be nonzero.

Conversely, a beta pricing model that explains expected returns perfectly may still produce pricing errors. Consider the case where the zero-beta rate is zero and we have a set of factors such that

$$\mu_2 = \beta\gamma_1. \tag{29}$$

This model explains expected returns perfectly. However, there is not a linear combination of μ_2 and β such that it is equal to 1_N and we will still have nonzero pricing errors and nonzero

[7] Here and in our following analysis, we assume that the two frontiers are not identical to each other when the asset pricing model is correctly specified. If this is not the case, we have $\theta_1(r) = \theta_2(r)$ for all r and γ_0 is not uniquely defined. See Cheung, Kwan, and Mountain (2000) for a further discussion of this point and its impact on statistical tests of asset pricing models.

[8] For the general K factor cases, if there exists a K-vector c such that $\beta c = 1_N$, then we will have zero HJ-distance for the model regardless of μ_2. Geometrically, this corresponds to the case that the two minimum-variance frontiers touch each other at the global minimum-variance portfolio.

HJ-distance for this model.

These two examples illustrate that when $\gamma_0 = 0$ or $\gamma_0 = \pm\infty$, the equivalence of a beta pricing model and a linear stochastic discount factor model breaks down. While these two examples are extreme cases, our point is that when one is concerned with minimizing HJ-distance across models with different factors, one can end up locating a factor such that the betas with respect to this factor are roughly constant across assets, without knowing that the betas of such a factor may do a very poor job in explaining expected returns. On the other hand, when one is concerned with minimizing Q_C across models, one may still end up with a model that produces relatively large pricing errors. Therefore, ranking models using δ and Q_C can yield very different conclusions. To make our point more concrete, we present a simple numerical example. Suppose we have four test assets with their returns driven by the following process

$$R_2 = \mu_2 + \beta_1 f_1 + \beta_2 f_2 + \epsilon, \tag{30}$$

where $f_1 \sim N(0, 0.01)$, $f_2 \sim N(0, 0.01)$, $\epsilon \sim N(0_4, \Sigma)$, independent of each other and the parameters are given by

$$\mu_2 = \begin{bmatrix} 1.04 \\ 1.08 \\ 1.12 \\ 1.16 \end{bmatrix}, \quad \beta_1 = \begin{bmatrix} 1.03 \\ 1.08 \\ 1.12 \\ 1.2 \end{bmatrix}, \quad \beta_2 = \begin{bmatrix} 1.05 \\ 1 \\ 1.05 \\ 1 \end{bmatrix}, \quad \Sigma = 0.01 \begin{bmatrix} 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 \end{bmatrix}. \tag{31}$$

In Figure 1, we plot the minimum-variance frontier of the four test assets as well as the mimicking portfolios for each of the two factors. When one calculates δ^2, one will find that the model with just the first factor has $\delta^2 = 0.798$, but a competing model with just the second factor has a smaller $\delta^2 = 0.500$. Therefore, using the HJ-distance, one considers the model with the second factor to be a superior model in explaining prices, despite the fact that its mimicking portfolio is further away from the minimum-variance frontier than the one for the first factor. However, if one chooses to rank the two models using Q_C, then one will find that the model with the first factor has a $Q_C = 0.090$ and is far superior to the model with the second factor, which has a Q_C of 3.033. This example goes to show that ranking models by Q_C and δ^2 can give conflicting conclusions. When that happens, researchers have to be careful in selecting which criterion to rely on. The bottom line is if one is interested in explaining prices, one should use HJ-distance to rank models, but if one is interested in explaining expected returns, then one is better off using Q_C for model selection.

Figure 1 about here

The main reason why Q_C and δ^2 do not provide the same ranking on models is because the choice of zero-beta rate depends on the criterion that we use in selecting models, and it is also model dependent. If one can *ex ante* fix the zero-beta rate to be the same across models, then we would not have this problem. Some recent empirical studies attempt to address this problem by including a short-term T-bill as a test asset (e.g., Hodrick and Zhang (2001) and Dittmar (2002)). However, in these empirical studies, the T-bill is treated just like any other risky asset and its returns have nonzero variance as well as nonzero covariances with other risky assets. Therefore, the zero-beta rate is still not constant across different models, and the divergence between Q_C and δ^2 can still exist in these studies.

II. Sample Measures of Model Misspecification

A. Sample HJ-Distance and CSRT Statistic

The discussion on model misspecification so far has been conducted using population expectations. In practice, we typically assume that the data is jointly stationary and ergodic; therefore, these expectations can be approximated using sample averages. Suppose we have T observations of $Y_t = [f_t',\ R_{2t}']'$, where f_t and R_{2t} are the realizations of K common factors and gross returns on N risky assets at time t. Define the sample mean and variance of Y_t as

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} Y_t \equiv \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix}, \tag{32}$$

$$\hat{V} = \frac{1}{T}\sum_{t=1}^{T} (Y_t - \hat{\mu})(Y_t - \hat{\mu})' \equiv \begin{bmatrix} \hat{V}_{11} & \hat{V}_{12} \\ \hat{V}_{21} & \hat{V}_{22} \end{bmatrix}, \tag{33}$$

where $\hat{V}$ is assumed to be nonsingular. The squared sample HJ-distance is given by

$$\hat{\delta}^2 = 1_N'[\hat{U}^{-1} - \hat{U}^{-1}\hat{D}(\hat{D}'\hat{U}^{-1}\hat{D})^{-1}\hat{D}'\hat{U}^{-1}]1_N, \tag{34}$$

where $\hat{D} = \frac{1}{T}\sum_{t=1}^{T} R_{2t}[1,\ f_t'] = [\hat{\mu}_2,\ \hat{V}_{21} + \hat{\mu}_2\hat{\mu}_1']$ and $\hat{U} = \frac{1}{T}\sum_{t=1}^{T} R_{2t}R_{2t}' = \hat{V}_{22} + \hat{\mu}_2\hat{\mu}_2'$.

In computing the sample HJ-distance (34), the standard practice is to estimate the linear coefficients of the stochastic discount factor, λ, to minimize the sample HJ-distance. The resulting

estimate of λ is given by

$$\hat{\lambda}^{HJ} \equiv \begin{bmatrix} \hat{\lambda}_0^{HJ} \\ \hat{\lambda}_1^{HJ} \end{bmatrix} = \text{argmin}_\lambda (\hat{D}\lambda - 1_N)'\hat{U}^{-1}(\hat{D}\lambda - 1_N) = (\hat{D}'\hat{U}^{-1}\hat{D})^{-1}(\hat{D}'\hat{U}^{-1}1_N), \tag{35}$$

where $\hat{\lambda}_0^{HJ}$ is a scalar and $\hat{\lambda}_1^{HJ}$ is a K-vector. However, to facilitate our later comparison with traditional specification tests of beta pricing models, we introduce the estimated zero-beta rate and risk premium implied by $\hat{\lambda}^{HJ}$ as

$$\hat{\gamma}^{HJ} \equiv \begin{bmatrix} \hat{\gamma}_0^{HJ} \\ \hat{\gamma}_1^{HJ} \end{bmatrix} = \frac{1}{\hat{\lambda}_0^{HJ} + \hat{\mu}_1'\hat{\lambda}_1^{HJ}} \begin{bmatrix} 1 \\ -\hat{V}_{11}\hat{\lambda}_1^{HJ} \end{bmatrix}. \tag{36}$$

Since there is a one-to-one correspondence between $\hat{\lambda}^{HJ}$ and $\hat{\gamma}^{HJ}$, we can interpret $\hat{\gamma}_0^{HJ}$ and $\hat{\gamma}_1^{HJ}$ as the estimated zero-beta rate and risk premium that minimize the sample HJ-distance.[9]

In the actual calculation of the sample HJ-distance, it is probably better to use the following expression instead of (34)

$$\hat{\delta}^2 = 1_N'[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}\hat{H}(\hat{H}'\hat{\Sigma}^{-1}\hat{H})^{-1}\hat{H}'\hat{\Sigma}^{-1}]1_N, \tag{38}$$

where $\hat{H} = [\hat{\mu}_2,\ \hat{\beta}]$, $\hat{\Sigma} = \hat{V}_{22} - \hat{V}_{21}\hat{V}_{11}^{-1}\hat{V}_{12}$, and $\hat{\beta} = \hat{V}_{21}\hat{V}_{11}^{-1}$. From Lemma 1, we know (34) and (38) are mathematically equivalent, but inverting $\hat{\Sigma}$ in (38) is numerically more stable than inverting $\hat{U}$ in (34).

For the beta pricing models, Shanken (1985) suggests a GLS cross-sectional regression test (CSRT) which is a sample counterpart of the aggregate pricing errors Q_C discussed in the previous section. The CSRT statistic of Shanken (1985) is obtained from running a GLS CSR of $\hat{\mu}_2$ on $\hat{G} = [1_N,\ \hat{\beta}]$. The estimated zero-beta rate γ_0^{CS} and risk premium γ_1^{CS} in this GLS CSR are given by

$$\hat{\gamma}^{CS} \equiv \begin{bmatrix} \hat{\gamma}_0^{CS} \\ \hat{\gamma}_1^{CS} \end{bmatrix} = (\hat{G}'\hat{\Sigma}^{-1}\hat{G})^{-1}(\hat{G}'\hat{\Sigma}^{-1}\hat{\mu}_2). \tag{39}$$

With this estimate of γ, the average return errors from this GLS CSR are given by

$$\hat{e}_{CS} = \hat{\mu}_2 - 1_N\hat{\gamma}_0^{CS} - \hat{\beta}\hat{\gamma}_1^{CS}. \tag{40}$$

[9]For a given value of $\hat{\gamma}^{HJ}$, it is easy to show that

$$\hat{\lambda}^{HJ} = \frac{1}{\hat{\gamma}_0^{HJ}} \begin{bmatrix} 1 + \hat{\mu}_1'\hat{V}_{11}^{-1}\hat{\gamma}_1^{HJ} \\ -\hat{V}_{11}^{-1}\hat{\gamma}_1^{HJ} \end{bmatrix}. \tag{37}$$

Shanken (1985) defines the CSRT statistic as an aggregate of these errors on average returns[10]

$$\hat{Q}_C = \hat{e}'_{CS}\hat{\Sigma}^{-1}\hat{e}_{CS}. \tag{41}$$

Shanken (1985) shows that under the null hypothesis that the model is correctly specified, we have

$$\frac{T\hat{Q}_C}{1+\hat{\gamma}_1^{CS\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{CS}} \overset{A}{\sim} \chi^2_{N-K-1}. \tag{42}$$

In addition, he also suggests the following approximate finite sample distribution under the null hypothesis

$$\frac{\hat{Q}_C}{1+\hat{\gamma}_1^{CS\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{CS}} \sim \left(\frac{N-K-1}{T-N+1}\right) F_{N-K-1,T-N+1}. \tag{43}$$

The term $\hat{\gamma}_1^{CS\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{CS}$ is called the errors-in-variables adjustment by Shanken (1985), which reflects the fact that estimated betas instead of true betas are used in the CSR.

B. The Geometry of Sample HJ-Distance and CSRT Statistic

While it is important to have finite sample distributions of the sample HJ-distance, it is equally important to develop a measure that allows one to examine the economic significance of departures from the true model. Fortunately, we can provide an appealing geometric interpretation of both the sample HJ-distance and the CSRT statistic. To prepare for our presentation of the geometry, we introduce three sample efficiency set constants $\hat{a}_2 = \hat{\mu}_2'\hat{V}_{22}^{-1}\hat{\mu}_2$, $\hat{b}_2 = \hat{\mu}_2'\hat{V}_{22}^{-1}1_N$, $\hat{c}_2 = 1_N'\hat{V}_{22}^{-1}1_N$. Similarly, we define $R_{1t} = \hat{V}_{12}\hat{V}_{22}^{-1}R_{2t}$ as the payoffs on K mimicking positions and the corresponding three sample efficiency set constants as $\hat{a}_1 = \hat{\mu}_2'\hat{V}_{22}^{-1}\hat{V}_{21}(\hat{V}_{12}\hat{V}_{22}^{-1}\hat{V}_{21})^{-1}\hat{V}_{12}\hat{V}_{22}^{-1}\hat{\mu}_2$, $\hat{b}_1 = \hat{\mu}_2'\hat{V}_{22}^{-1}\hat{V}_{21}(\hat{V}_{12}\hat{V}_{22}^{-1}\hat{V}_{21})^{-1}\hat{V}_{12}\hat{V}_{22}^{-1}1_N$, and $\hat{c}_1 = 1_N'\hat{V}_{22}^{-1}\hat{V}_{21}(\hat{V}_{12}\hat{V}_{22}^{-1}\hat{V}_{21})^{-1}\hat{V}_{12}\hat{V}_{22}^{-1}1_N$. Let $\Delta\hat{a} = \hat{a}_2 - \hat{a}_1$, $\Delta\hat{b} = \hat{b}_2 - \hat{b}_1$, and $\Delta\hat{c} = \hat{c}_2 - \hat{c}_1$. The following Proposition is the sample counterpart of Proposition 1. It expresses the two test statistics in terms of sample Sharpe ratios of the two *ex post* frontiers and also provides a characterization of the estimated zero-beta rates of the two test statistics.

Proposition 2: *The sample HJ-distance* $(\hat{\delta}^2)$ *and the CSRT statistic* $(\hat{Q}_C)$ *of a* K*-factor beta pricing model can be written as*

$$\hat{\delta}^2 \;=\; \min_{\gamma_0}\frac{\hat{\theta}_2^2(\gamma_0)-\hat{\theta}_1^2(\gamma_0)}{\gamma_0^2} = \frac{\hat{\theta}_2^2(\hat{\gamma}_0^{HJ})-\hat{\theta}_1^2(\hat{\gamma}_0^{HJ})}{(\hat{\gamma}_0^{HJ})^2}, \tag{44}$$

[10]Shanken's version of $\hat{Q}_C$ actually multiplies the aggregate average return errors by T and uses the unbiased estimate of Σ. We modify his definition here to allow for easier comparison with the sample HJ-distance.

$$\hat{Q}_C \quad = \quad \min_{\gamma_0} \hat{\theta}_2^2(\gamma_0) - \hat{\theta}_1^2(\gamma_0) = \hat{\theta}_2^2(\hat{\gamma}_0^{CS}) - \hat{\theta}_1^2(\hat{\gamma}_0^{CS}), \tag{45}$$

where $\hat{\gamma}_0^{HJ} = \Delta\hat{a}/\Delta\hat{b}$, $\hat{\gamma}_0^{CS} = \Delta\hat{b}/\Delta\hat{c}$, *and* $\hat{\theta}_1(\gamma_0)$ *and* $\hat{\theta}_2(\gamma_0)$ *are the sample Sharpe ratios of the ex post tangency portfolio of the* K *mimicking positions and of the* N *test assets, respectively, when* γ_0 *is treated as the y-intercept of the tangent line. If* $\Delta\hat{b} \geq 0$, *we have* $\hat{\gamma}_0^{HJ} \geq \hat{\gamma}_0^{CS}$, *and if* $\Delta\hat{b} < 0$, *we have* $\hat{\gamma}_0^{HJ} \leq \hat{\gamma}_0^{CS}$.

In Figure 2, we plot the *ex post* minimum-variance frontier of the K mimicking positions and the minimum-variance frontier of the N test assets in the $(\hat{\sigma}, \hat{\mu})$ space. The two lines HA and HB are tangent to the *ex post* minimum-variance frontiers of the K mimicking positions and N test assets, respectively. The x-intercepts of these two tangent lines are points A and B, respectively. Let ψ be the angle $\angle HAO$, then we have $\tan(\psi) = |\hat{\theta}_2(\hat{\gamma}_0^{HJ})|$ and it is easy to see that the length of OA is $\hat{\gamma}_0^{HJ}/|\hat{\theta}_2(\hat{\gamma}_0^{HJ})|$. Similarly, the length of OB is $\hat{\gamma}_0^{HJ}/|\hat{\theta}_1(\hat{\gamma}_0^{HJ})|$. Therefore, we can write

$$\hat{\delta}^2 = \frac{1}{OA^2} - \frac{1}{OB^2}. \tag{46}$$

There is yet another geometric interpretation of $\hat{\delta}^2$. For each of the two tangent lines, we find a point on it that is closest to the origin. For the tangent line HA, the point is C and for the tangent line HB, the point is D. Since OC is perpendicular to HA, the angle $\angle HOC$ is also the same as the angle $\angle HAO$, which is ψ. Therefore, the length of OC is equal to $\hat{\gamma}_0^{HJ}\cos(\psi) = \hat{\gamma}_0^{HJ}/\sqrt{1+\hat{\theta}_2^2(\hat{\gamma}_0^{HJ})}$. Similarly, the length of OD is $\hat{\gamma}_0^{HJ}/\sqrt{1+\hat{\theta}_1^2(\hat{\gamma}_0^{HJ})}$. With these results, we can also write

$$\hat{\delta}^2 = \frac{1}{OC^2} - \frac{1}{OD^2}. \tag{47}$$

Heuristically, if we treat $\hat{\gamma}_0^{HJ}$ as the risk-free rate, we can think of C as the *ex post* minimum second moment portfolio (with unit cost) of the N assets plus the risk-free asset, and this portfolio has a second moment of OC^2. If we scale this portfolio so that its second moment is equal to one, then its cost is $1/OC$ and we can interpret $1/OC$ as the maximum price one is willing to pay for a unit second moment portfolio of the N test assets and the risk-free asset. Similarly, D can be interpreted as the *ex post* minimum second moment portfolio (with unit cost) of the K mimicking positions plus the risk-free asset, and it has a second moment of OD^2. If we scale portfolio D so that it has unit second moment, then its cost is $1/OD$. Therefore, $\hat{\delta}^2$ can be thought of as the estimated squared price difference of the two portfolios C and D, when both are scaled to have unit second moment. This is exactly what HJ-distance is trying to measure — the maximum pricing

error of a model. From both of these geometrical interpretations of $\hat{\delta}^2$, we can see that HJ-distance is a measure of how close the two tangency portfolios are when the y-intercept of the tangent lines is chosen to be $\hat{\gamma}_0^{HJ}$.

It is well known that the beta asset pricing model holds if and only if the two frontiers touch each other, i.e., there exists a γ_0 such that we have $\theta_2(\gamma_0) = \theta_1(\gamma_0)$ for the two *ex ante* minimum-variance frontiers.[11] Therefore, if the beta asset pricing model is correctly specified, we should expect the two *ex post* frontiers to be very close to each other at some point and hence the length of OA should not be significantly different from the length of OB. If instead we observe a large value of $\hat{\delta}$, then it is an indication that the two *ex ante* frontiers do not touch each other and we reject the model as a result.

Figure 2 about here

In Figure 2, we also plot two tangent lines emanating from point G (which is the point $(0, \hat{\gamma}_0^{CS})$) to the two *ex post* frontiers. The slope of the line GE is equal to $\hat{\theta}_2(\hat{\gamma}_0^{CS})$ and since point E has a standard deviation of one, the length of GE is given by $\sqrt{1 + \hat{\theta}_2^2(\hat{\gamma}_0^{CS})}$. Similarly, the length of GF is given by $\sqrt{1 + \hat{\theta}_1^2(\hat{\gamma}_0^{CS})}$. Therefore, we can write the CSRT statistic as

$$\hat{Q}_C = (GE)^2 - (GF)^2. \tag{48}$$

From this geometric interpretation of $\hat{Q}_C$, we can see that the CSRT statistic is also a measure of how close the two tangency portfolios are except that the y-intercept of the tangent lines is chosen to be $\hat{\gamma}_0^{CS}$.[12]

Under the null hypothesis that the asset pricing model is correctly specified, the two approaches are asymptotically equivalent because both $\hat{\gamma}_0^{CS}$ and $\hat{\gamma}_0^{HJ}$ converge to the same limit as $T \to \infty$. However, when the asset pricing model does not hold, $\hat{\gamma}_0^{CS}$ and $\hat{\gamma}_0^{HJ}$ converge to different limits. As discussed earlier, the sample HJ-distance tends to choose a higher absolute value of zero-beta rate than the CSRT statistic because a large value of γ_0 can deflate the pricing errors. The effect of choosing a higher absolute value of γ_0 by the sample HJ-distance is that one often finds that the HJ-distance focuses on the difference of the two frontiers at the inefficient side.

[11] See, for example, Grinblatt and Titman (1987) and Huberman and Kandel (1987).

[12] For the special case of a one-factor model and the factor is the return on a portfolio, Roll (1985) provides a geometric interpretation of the CSRT statistic, except that his is given in the $(\hat{\sigma}^2, \hat{\mu})$ space, not in the $(\hat{\sigma}, \hat{\mu})$ space.

C. A Comparison with GMM Over-identification Tests

Another popular specification test of linear stochastic discount factor models is the GMM over-identification test of Hansen (1982). Denote $g_t(\lambda) = R_{2t}x_t'\lambda - 1_N$ and

$$\bar{g}(\lambda) = \frac{1}{T}\sum_{t=1}^{T} g_t(\lambda) = \hat{D}\lambda - 1_N. \tag{49}$$

Suppose $\hat{S}$ is a consistent estimator of S, where S is the asymptotic variance of $\bar{g}(\lambda)$. The optimal GMM estimator of λ is given by

$$\hat{\lambda}_{GMM} = (\hat{D}'\hat{S}^{-1}\hat{D})^{-1}(\hat{D}'\hat{S}^{-1}1_N), \tag{50}$$

and the popular GMM over-identification test of the asset pricing model is given by

$$J = T1_N'[\hat{S}^{-1} - \hat{S}^{-1}\hat{D}(\hat{D}'\hat{S}^{-1}\hat{D})^{-1}\hat{D}'\hat{S}^{-1}]1_N. \tag{51}$$

When the model is correctly specified, we have $J \overset{A}{\sim} \chi^2_{N-K-1}$.

The expression of S depends on the distribution of $Y_t = [f_t',\ R_{2t}']'$. Assuming the returns on the N test assets follow a K-factor model

$$R_{2t} = \alpha + \beta f_t + \epsilon_t, \tag{52}$$

where $E[\epsilon_t] = 0_N$ and $E[\epsilon_t|f_t] = 0_N$, the following lemma gives the expression of S for two important cases.

Lemma 3 *Suppose Y_t is a stationary ergodic sequence and $g_t(\lambda)$ is a martingale difference sequence. If $\mathrm{Var}[\epsilon_t|f_t] = \Sigma$, where Σ is a positive definite matrix independent of f_t (i.e., conditional homoskedasticity), we have*

$$S = E[(x_t'\lambda)^2]\Sigma + BCB', \tag{53}$$

where $B = [\alpha,\ \beta]$ and C is a $(K+1)\times(K+1)$ matrix. If Y_t is i.i.d. and follows a multivariate elliptical distribution with a kurtosis parameter κ,[13] *we have*

$$S = \left(E[(x_t'\lambda)^2] + \kappa\lambda_1'V_{11}\lambda_1\right)\Sigma + BCB'. \tag{55}$$

[13]The multivariate kurtosis parameter is defined as

$$\kappa = \frac{E[((Y_t-\mu)'V^{-1}(Y_t-\mu))^2]}{(N+K)(N+K+2)} - 1. \tag{54}$$

For elliptical distribution, this is the same as the univariate kurtosis parameter $\mu_4/(3\sigma^4) - 1$ for any of its marginal distribution.

Note that under both assumptions, S takes the form of $a\Sigma + BCB'$ for some positive scalar a and a matrix C. It turns out that if we choose $\hat{S}$ also to be of this form, the optimal GMM estimate of λ is numerically identical to the HJ-distance estimate of λ, and the GMM over-identification test statistic is closely related to the squared sample HJ-distance.

Proposition 3: *Suppose $S = a\Sigma + BCB'$, where a is a positive scalar and C is a $(K+1)\times(K+1)$ matrix. Define $\hat{B} = [\hat{\alpha},\ \hat{\beta}]$ as the usual OLS estimator of B and $\hat{a}$ is a consistent estimator of a. If we use $\hat{S}^{-1}$ as the optimal GMM weighting matrix where $\hat{S} = \hat{a}\hat{\Sigma} + \hat{B}\hat{C}\hat{B}'$ and $\hat{C}$ is an arbitrary matrix, the GMM estimate of λ is numerically identical to the HJ-distance estimate of λ,*

$$\hat{\lambda}^{GMM} = (\hat{D}'\hat{S}^{-1}\hat{D})^{-1}(\hat{D}'\hat{S}^{-1}1_N) = (\hat{D}'\hat{U}^{-1}\hat{D})^{-1}(\hat{D}'\hat{U}^{-1}1_N) = \hat{\lambda}^{HJ}, \tag{56}$$

and the GMM over-identification test statistic under this choice of weighting matrix is equal to

$$J = T\bar{g}(\hat{\lambda}^{GMM})'\hat{S}^{-1}\bar{g}(\hat{\lambda}^{GMM}) = T1_N'[\hat{S}^{-1} - \hat{S}^{-1}\hat{D}(\hat{D}'\hat{S}^{-1}\hat{D})^{-1}\hat{D}'\hat{S}^{-1}]1_N = \frac{T\hat{\delta}^2}{\hat{a}}, \tag{57}$$

which implies $T\hat{\delta}^2 \overset{A}{\sim} a\chi^2_{N-K-1}$ when the model is correctly specified.

Proposition 3 suggests that under some popular assumptions on the distribution of Y_t, the sample squared HJ-distance is a rescaled version of the GMM over-identification test statistic. In addition, the p-values that we obtain for both tests are identical.[14] In practice, one often does not impose these restrictions in computing $\hat{S}$ for GMM estimation and testing. In that case, even though S is actually of the form $a\Sigma + BCB'$, λ^{GMM} and J will only be asymptotically equivalent, but not numerically identical, to λ^{HJ} and $T\hat{\delta}^2/\hat{a}$, respectively.

How is the GMM over-identification test statistic related to the CSRT statistic? Under the conditional homoskedasticity assumption, we have

$$a = E[(x_t'\lambda)^2] = \lambda' E[x_t x_t']\lambda = \lambda' \begin{bmatrix} 1 & \mu_1 \\ \mu_1 & V_{11} + \mu_1\mu_1' \end{bmatrix} \lambda = \frac{1 + \gamma_1' V_{11}^{-1}\gamma_1}{\gamma_0^2}, \tag{58}$$

where the last equality follows from the reparameterization

$$\gamma \equiv \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \frac{1}{\lambda_0 + \mu_1'\lambda_1} \begin{bmatrix} 1 \\ -V_{11}\lambda_1 \end{bmatrix}. \tag{59}$$

[14] The GMM over-identification test compares $T\hat{\delta}^2/\hat{a}$ with χ^2_{N-K-1}. The test of $H_0 : \delta = 0$ compares $T\hat{\delta}^2$ with $\hat{a}\chi^2_{N-K-1}$. Therefore, the p-values for both tests are $P[\chi^2_{N-K-1} > T\hat{\delta}^2/\hat{a}]$.

When $\hat{S} = \hat{a}\hat{\Sigma} + \hat{B}\hat{C}\hat{B}'$, we have $\hat{\lambda}^{GMM} = \hat{\lambda}^{HJ}$, so a consistent estimator of a is

$$\hat{a} = \frac{1 + \hat{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{HJ}}{(\hat{\gamma}_0^{HJ})^2}, \tag{60}$$

and an asymptotically equivalent version of the optimal GMM over-identification test is

$$J = \frac{T\hat{\delta}^2}{\hat{a}} = \frac{T[\hat{\theta}_2^2(\hat{\gamma}_0^{HJ}) - \hat{\theta}_1^2(\hat{\gamma}_0^{HJ})]}{1 + \hat{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{HJ}}. \tag{61}$$

Comparing with (42), one can think of the CSRT statistic as a GMM over-identification test statistic, with the conditional homoskedasticity assumption imposed and with the use of a different estimate of γ.

When the conditional homoskedasticity assumption is inappropriate, the CSRT statistic is no longer equivalent to the GMM over-identification test. However, under the multivariate elliptical distribution assumption on Y_t, we can make a simple modification to restore the equivalence. Since

$$a = \lambda' E[x_t x_t']\lambda + \kappa\lambda_1' V_{11}\lambda_1 = \lambda' \begin{bmatrix} 1 & \mu_1' \\ \mu_1 & (1+\kappa)V_{11} + \mu_1\mu_1' \end{bmatrix} \lambda = \frac{1 + (1+\kappa)\gamma_1' V_{11}^{-1}\gamma_1}{\gamma_0^2} \tag{62}$$

under the multivariate elliptical distribution assumption, a modified version of the CSRT statistic is given by

$$\frac{T\hat{Q}^C}{1 + (1+\kappa)\hat{\gamma}_1^{CS\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{CS}} = \frac{T[\hat{\theta}_2^2(\hat{\gamma}_0^{CS}) - \hat{\theta}_1^2(\hat{\gamma}_0^{CS})]}{1 + (1+\kappa)\hat{\gamma}_1^{CS\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{CS}} \overset{A}{\sim} \chi^2_{N-K-1}. \tag{63}$$

In addition, an asymptotic equivalent version of the GMM over-identification test can also be obtained by replacing $\hat{\gamma}^{CS}$ with $\hat{\gamma}^{HJ}$. Comparing (63) with (42), we note that the only difference here is that the errors-in-variables adjustment in the denominator of the CSRT statistic needs to be modified to reflect the fact that there are more estimation errors in $\hat{B}$ when the elliptical distribution has fat-tails ($\kappa > 0$). This also suggests that the power of the sample HJ-distance, the CSRT, and the GMM J-test to detect model misspecification is a decreasing function of the kurtosis parameter κ.

III. Finite Sample Distribution of Sample HJ-Distance

A. Simplification of the Problem

After obtaining an understanding of the similarities and differences between the sample HJ-distance and other specification tests, we now turn our attention to the exact distribution of the sample

HJ-distance. Obtaining the exact distribution of the sample HJ-distance is a formidable task even under the normality assumption. In our approach to this problem, we take three different steps to simplify it.

For notational brevity, we use the matrix form of model (52) in what follows. Suppose we have T observations of f_t and R_{2t}, we write

$$R_2 = XB' + E, \tag{64}$$

where R_2 is a $T \times N$ matrix with its typical row equal to R_{2t}', X is a $T \times (K+1)$ matrix with its typical row as $[1,\ f_t']$, $B = [\,\alpha,\ \beta\,]$ is the matrix of regression coefficients, and E is a $T \times N$ matrix with ϵ_t' as its typical row. As usual, we assume that $T \geq N + K + 1$, $X'X$ is nonsingular, and β is of full column rank. For the purpose of obtaining an exact distribution of the sample HJ-distance, we assume that, conditional on f_t, the disturbances ϵ_t are independent and identically distributed as multivariate normal with mean zero and variance Σ.[15]

The maximum likelihood estimators of B and Σ are the usual ones:

$$\hat{B} \equiv [\,\hat{\alpha},\ \hat{\beta}\,] = (R_2'X)(X'X)^{-1}, \tag{65}$$

$$\hat{\Sigma} = \frac{1}{T}(R_2 - X\hat{B}')'(R_2 - X\hat{B}'). \tag{66}$$

Under the normality assumption, we have $\hat{B}$ and $\hat{\Sigma}$ independent of each other and their distributions are given by

$$\mathrm{vec}(\hat{B}) \sim N(\mathrm{vec}(B), (X'X)^{-1} \otimes \Sigma), \tag{67}$$

$$T\hat{\Sigma} \sim W_N(T - K - 1, \Sigma), \tag{68}$$

where $W_N(T-K-1, \Sigma)$ is the N-dimensional central Wishart distribution with $T-K-1$ degrees of freedom and covariance matrix Σ.

One of the problems with obtaining the exact distribution of the sample HJ-distance is that $\hat{\delta}^2$ is usually written as a function of $\hat{D}$ and $\hat{U}$, whose distributions are rather difficult to obtain. Our first simplification is to write $\hat{\delta}^2$ as a function of $\hat{B}$ and $\hat{\Sigma}$, so we can use the well established distribution results (67) and (68) above. Using Lemma 1 and noting that

$$\hat{H} = [\hat{\mu}_2,\ \hat{\beta}] = [\hat{\alpha},\ \hat{\beta}] \begin{bmatrix} 1 & 0_K' \\ \hat{\mu}_1 & I_K \end{bmatrix}, \tag{69}$$

[15]Note that we do not require R_{2t} to be multivariate normally distributed; the distribution of f_t can be time-varying and arbitrary. We only need to assume that conditional on f_t, R_{2t} is normally distributed.

we can write

$$\hat{\delta}^2 = 1_N'[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}\hat{B}(\hat{B}'\hat{\Sigma}^{-1}\hat{B})^{-1}\hat{B}'\hat{\Sigma}^{-1}]1_N. \tag{70}$$

Still, it is a daunting task to get an exact distribution of (70). Our second simplification of the problem relies on the following lemma which helps us to get rid of the influence of $\hat{\Sigma}$.

Lemma 4 *Defining*

$$\tilde{\delta}^2 = 1_N'[\Sigma^{-1} - \Sigma^{-1}\hat{B}(\hat{B}'\Sigma^{-1}\hat{B})^{-1}\hat{B}'\Sigma^{-1}]1_N, \tag{71}$$

we have

$$V = T\tilde{\delta}^2/\hat{\delta}^2 \sim \chi^2_{T-N+1} \tag{72}$$

and it is independent of $\tilde{\delta}^2$.

Note that $\tilde{\delta}^2$ is similar to $\hat{\delta}^2$ except that $\tilde{\delta}^2$ has the true Σ instead of the estimated $\hat{\Sigma}$ in its expression. Lemma 4 is extremely useful because it allows us to focus our efforts on obtaining just the distribution of $\tilde{\delta}^2$. Once this is obtained, we can get the distribution of $\hat{\delta}^2$ using the fact that

$$\hat{\delta}^2 = \frac{T\tilde{\delta}^2}{V}, \tag{73}$$

where $\tilde{\delta}^2$ and $V \sim \chi^2_{T-N+1}$ are independent of each other.

Our third simplification is to normalize $\hat{B}$ using a transformation

$$\tilde{B}_n = \Sigma^{-\frac{1}{2}}\hat{B}(X'X)^{\frac{1}{2}}, \tag{74}$$

so $\text{vec}(\tilde{B}_n) \sim N(\text{vec}(B_n), I_{K+1} \otimes I_N)$, where $B_n = \Sigma^{-\frac{1}{2}}B(X'X)^{\frac{1}{2}}$ is the normalized version of B, and all the elements of $\tilde{B}_n$ are independent normal random variables with unit variance. With this normalization and defining $\nu = \Sigma^{-\frac{1}{2}}1_N$, we can write

$$\delta^2 = \nu'[I_N - B_n(B_n'B_n)^{-1}B_n']\nu, \tag{75}$$

$$\tilde{\delta}^2 = \nu'[I_N - \tilde{B}_n(\tilde{B}_n'\tilde{B}_n)^{-1}\tilde{B}_n']\nu. \tag{76}$$

B. Exact Distribution

With all these simplifications, we are now ready to present the distribution of $\hat{\delta}^2$. Let $Q\Lambda Q'$ be the eigenvalue decomposition of $B_n'[I_N - \nu(\nu'\nu)^{-1}\nu']B_n$ where Λ is a diagonal matrix with its diagonal

elements $\lambda_1 \geq \cdots \geq \lambda_{K+1} \geq 0$ equal to the eigenvalues, and Q is an orthonormal matrix of the corresponding eigenvectors. The following proposition expresses the HJ-distance in terms of these quantities.

Proposition 4: *Defining* $\xi = Q'B_n'\nu/(\nu'\nu)^{\frac{1}{2}}$, *we have*

$$\delta^2 = \frac{\nu'\nu}{1+\xi'\Lambda^{-1}\xi}, \tag{77}$$

$$\tilde{\delta}^2 = \frac{\nu'\nu}{1+U_1'W^{-1}U_1}, \tag{78}$$

where $U_1 \sim N(\xi, I_{K+1})$ *is normal,* $W \sim W_{K+1}(N-1, I_{K+1}, \Lambda)$ *is a* $K+1$ *dimensional noncentral Wishart with* $N-1$ *degrees of freedom, covariance matrix* I_{K+1}, *and noncentrality parameter* Λ, *with* U_1 *and* W *independent of each other.*

Although (78) does not admit an explicit expression of the cumulative density function, it allows us to use a Monte Carlo integration approach to obtain the distribution of $\hat{\delta}^2$, which can be easily performed as follows:

1. Simulate $U_1 \sim N(\xi, I_{K+1})$, $W \sim W_{K+1}(N-1, I_{K+1}, \Lambda)$, independent of each other.
2. Compute $\tilde{\delta}^2 = \frac{1_N'\Sigma^{-1}1_N}{1+U_1'W^{-1}U_1}$.
3. Since $\hat{\delta}^2 = T\tilde{\delta}^2/V$ where $V \sim \chi^2_{T-N+1}$ is independent of $\tilde{\delta}^2$, the cumulative distribution function for $\hat{\delta}^2$ can be approximated by

$$P[\hat{\delta}^2 > c] = E[P[V < T\tilde{\delta}^2/c|\tilde{\delta}^2]] \approx \frac{1}{n}\sum_{i=1}^{n} F_{\chi^2_{T-N+1}}(T\tilde{\delta}_i^2/c), \tag{79}$$

   where $F_{\chi^2_{T-N+1}}(x) = P[\chi^2_{T-N+1} \leq x]$, $\tilde{\delta}_i^2$ is the value of $\tilde{\delta}^2$ in the ith simulation, and n is the total number of simulations.

All that is required in this Monte Carlo integration approach is to simulate a $(K+1)$-dimensional normal and a $(K+1)$-dimensional noncentral Wishart random variables. In general, the number of factors (K) is a small number, so this procedure is very efficient. Note that the exact distribution of $\hat{\delta}^2$ depends on $2K+3$ nuisance parameters: Λ, ξ, and $\nu'\nu = 1_N'\Sigma^{-1}1_N$. Since the Monte Carlo integration approach depends on only a few nuisance parameters of the model, the effect of varying these nuisance parameters can be easily studied given the efficiency of this procedure.

When the asset pricing model is correctly specified, 1_N is in the span of the column space of B, or ν is in the span of the column space of B_n, and the matrix $B_n'[I_N - \nu(\nu'\nu)^{-1}\nu']B_n$ is only of rank K, so the last diagonal element of Λ is zero, i.e., $\lambda_{K+1} = 0$. Therefore, the distribution of $\hat{\delta}^2$ under the null depends on only $2K+2$ parameters. Since the only parameter that distinguishes the exact distribution of $\hat{\delta}^2$ under the null and under the alternative is the smallest eigenvalue λ_{K+1}, the power of the test of $H_0 : \delta = 0$ is crucially determined by λ_{K+1}. If λ_{K+1} is close to zero, then the distribution of $\hat{\delta}^2$ cannot be easily distinguished from that under the null hypothesis. If λ_{K+1} is very different from zero, then we will be able to detect the departure from the rank restriction with higher probability.

One may wonder why it is λ_{K+1} but not δ^2 that determines the power of the test. The reason is under the alternative hypothesis that $\delta \neq 0$, some sampling fluctuations in $\hat{\delta}^2$ are still perfectly consistent with the null hypothesis of $H_0 : \delta = 0$. This is so because when the model is correctly specified, we still have $\hat{\delta}^2 \neq 0$ due to sampling fluctuations in $\hat{B}$. As the risk premium of a K-factor asset pricing model is unspecified, the test will attempt to find the best K linear combinations of B such that they can provide the maximum explanatory power of the sampling fluctuations of $\hat{\delta}^2$ under the null hypothesis. It is only the unexplained portion of the sampling fluctuations of $\hat{\delta}^2$ that will allow us to determine whether δ is zero or not. However, the reliability of this inference crucially depends on the the volatility of the remaining $(K+1)$-th linear combination of $\hat{B}$ (i.e., the pricing errors). If λ_{K+1} is small, it implies either that the aggregate pricing errors are very small, or that they are very volatile. As a result, a misspecified model with small λ_{K+1} is very difficult to detect.

When $\delta^2 \neq 0$, we have $\lambda_{K+1} = O_p(T)$ and the null hypothesis of $H_0 : \delta = 0$ will be rejected with probability of one when $T \to \infty$. As a result, all wrong models will eventually be detected if we have a long enough time series of data. However, in a finite sample, it is not true that a worse model (in the sense that it has a higher δ^2) will be rejected with a higher probability. As an illustration, we return to our earlier example in (30). As we showed earlier, the model with just the first factor has $\delta^2 = 0.798$ and the model with just the second factor has $\delta^2 = 0.500$. If we assume in the sample, we have $\hat{\mu}_1 = \mu_1 = 0$ and $\hat{V}_{11} = V_{11} = 0.01$, then we can verify that $\lambda_2 = 0.0017T$ for the first model and $\lambda_2 = 0.0100T$ for the second model. Note that although the first model has a greater HJ-distance than the second model, the second model is rejected more often because

its smallest eigenvalue is 5.8 times larger than that of the first model. In Figure 3, we plot the probabilities of rejection of the null hypothesis $H_0 : \delta = 0$ as a function of T for these two different models. The size of the test is 5% and the distribution of $\hat{\delta}^2$ under the null and the alternatives are computed using the exact distribution, assuming $\hat{\mu}_1 = \mu_1 = 0$ and $\hat{V}_{11} = V_{11} = 0.01$. As we can see in Figure 3, when T is small, both models cannot be easily rejected. As T increases, the model with just the first factor still cannot be easily rejected. Even with $T = 1000$, we get a probability of rejection of only 27.3% for this model. However, the model with just the second factor, despite having a smaller δ^2 than the first model, can be rejected with very high probability when T is large. When $T = 1000$, we can reject $H_0 : \delta = 0$ for the second model with probability 87.9%. This example serves to illustrate that one should be extremely cautious in using the p-value of the sample HJ-distance to rank models. Models with higher δ^2 are not necessarily rejected with higher probability than models with lower δ^2.

Intuitively, the reason why we can reject the model with the second factor with relative ease is because the expected return of the mimicking portfolio of the second factor is close to that of the global minimum-variance portfolio of the test assets. As the variance of the global minimum-variance portfolio depends on only the variance-covariance matrix of the returns on the test assets, its location can be estimated rather accurately from the data. Therefore, we can reject the null hypothesis even when the mimicking portfolio has only a small departure from the global minimum-variance portfolio. On the contrary, for a frontier portfolio that is far away from the global minimum-variance portfolio, its variance depends on both the mean and the variance-covariance matrix of the returns on the test assets. As estimation of expected return is relatively noisy, the location of such a frontier portfolio is harder to determine. Therefore, when a mimicking portfolio has an expected return that is very different from that of the global minimum-variance portfolio, it is difficult to reject the null unless the mimicking portfolio is very far away from the frontier.

Figure 3 about here

C. Approximate Distribution

In using the finite sample distribution for specification test, one encounters a practical problem. It is that the finite sample distribution depends on some nuisance parameters (Λ, ξ and $\nu'\nu$) even

under the null hypothesis.[16] Therefore, one needs to estimate Λ, ξ and $\nu'\nu$ in order to compute the finite sample distribution. For wide applications, we suggest the following procedure to easily compute an approximate exact distribution which is accurate for most practical purposes.

Let $\hat{\nu} = \hat{\Sigma}^{-\frac{1}{2}} 1_N$, $\hat{B}_n = \hat{\Sigma}^{-\frac{1}{2}} \hat{B}(X'X)^{\frac{1}{2}}$ be the sample estimates of ν and B_n. Let $\hat{Q}\hat{\Lambda}\hat{Q}'$ be the eigenvalue decomposition of $\hat{B}_n'[I_N - \hat{\nu}(\hat{\nu}'\hat{\nu})^{-1}\hat{\nu}']\hat{B}_n$ and $\hat{\xi} = \hat{Q}'\hat{B}_n'\hat{\nu}/(\hat{\nu}'\hat{\nu})^{\frac{1}{2}}$ be the sample estimates of ξ. Under the null hypothesis, we have $\lambda_{K+1} = 0$ and in addition, as shown in the Appendix, we must have

$$\xi_{K+1}^2 = \frac{\nu'\nu}{\nu' B_n (B_n'B_n)^{-2} B_n'\nu}. \tag{80}$$

Therefore, for the purpose of testing $H_0 : \delta = 0$, we set the last diagonal element of $\hat{\Lambda}$ to zero and the last element of $\hat{\xi}$ to[17]

$$\hat{\xi}_{K+1} = \frac{(\hat{\nu}'\hat{\nu})^{\frac{1}{2}}}{(\hat{\nu}'\hat{B}_n(\hat{B}_n'\hat{B}_n)^{-2}\hat{B}_n'\hat{\nu})^{\frac{1}{2}}}. \tag{81}$$

Using these sample estimates $\hat{\Lambda}$, $\hat{\xi}$, and $\hat{\nu}'\hat{\nu}$ to replace the true ones in (78), we can obtain a finite sample distribution of $\hat{\delta}^2$. Since the sample estimates of the nuisance parameters are used here, the finite sample distribution is only approximate but not exact. However, our simulation evidence shows that this procedure is quite effective in approximating the true finite sample distribution when T is reasonably large.

If one is concerned with the effect of using estimated instead of true nuisance parameters, one can perturb the estimated parameters (say increasing them or decreasing them by 20%) to find out if the computed p-value is robust to the choice of nuisance parameters. Another way is to use a first order approximation of the finite sample distribution. The following Proposition uses the same argument as in Shanken (1985) and provides an approximate finite sample distribution for the sample HJ-distance.

Proposition 5: *Conditional on f_t, the squared sample HJ-distance has the following approximate finite sample distribution*

$$\hat{\delta}^2 \sim \left(\frac{1 + \hat{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{HJ}}{\left(\hat{\gamma}_0^{HJ}\right)^2}\right)\left(\frac{N-K-1}{T-N+1}\right) F_{N-K-1,T-N+1}(d), \tag{82}$$

[16] It is common that the finite sample distributions of test statistics of asset pricing models depend on some nuisance parameters. See, for example, Zhou (1995) and Velu and Zhou (1999).

[17] While the unconstrained $\hat{\xi}_{K+1}$ is still a consistent estimate of ξ_{K+1}, we find that the constrained version of $\hat{\xi}_{K+1}$ is much less volatile than the unconstrained one.

where $F_{N-K-1,T-N+1}(d)$ is a noncentral F-distribution with $N-K-1$ and $T-N+1$ degrees of freedom and noncentrality parameter

$$d = \frac{T\delta^2}{(1+\bar{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\bar{\gamma}_1^{HJ})/(\gamma_0^{HJ})^2} = \frac{T[\theta_2^2(\gamma_0^{HJ}) - \theta_1^2(\gamma_0^{HJ})]}{1+\bar{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\bar{\gamma}_1^{HJ}}, \tag{83}$$

and $\bar{\gamma}_1^{HJ} = \gamma_1^{HJ} + \hat{\mu}_1 - \mu_1$ is the ex post risk premium of the K factors.

Under the null hypothesis, we have $d = 0$ and the noncentral F-distribution becomes a central F-distribution, so an approximate F-test of $H_0 : \delta = 0$ is to compare

$$F_1 = \left(\frac{T-N+1}{N-K-1}\right) \frac{\hat{\delta}^2}{(1+\hat{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{HJ})/(\hat{\gamma}_0^{HJ})^2} \tag{84}$$

with a central F-distribution with $N-K-1$ and $T-N+1$ degrees of freedom.[18]

Under this approximate F-test, the power of the sample HJ-distance in rejecting the null hypothesis is positively related to the magnitude of the noncentrality parameter d. From (83), we can see that this noncentrality parameter depends on not just δ^2 or how far apart the two frontiers are, but also on the term $\bar{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\bar{\gamma}_1^{HJ}$. This term is similar to the errors-in-variables adjustment in Shanken (1985), which is due to the fact that we need to use the estimated betas instead of the true betas in the calculation of the sample HJ-distance. If the estimated betas are noisy, we cannot reliably reject the null hypothesis even though the true δ is nonzero. This observation suggests that besides preferring factors that generate high γ_0^{HJ}, sample HJ-distance also heavily favors models with noisy factors. This is so because if we add pure measurement errors to a factor, it will not change the true δ, but the term $\bar{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\bar{\gamma}_1^{HJ}$ will most likely go up to reduce the power of the test.

To further illustrate this point, we return to our earlier example in Figure 3 where we found that the model with just the second factor was rejected with fairly high probability. Our question here is that whether we can add a noise to the factor and make the model acceptable. Suppose we construct a noisy version of the second factor

$$f_t^* = f_{2t} + n_t, \tag{85}$$

where n_t is a noise with mean zero and variance σ_n^2, and it is independent of the factors and the returns. What would happen to the power of the test if we use f_t^* instead of f_{2t} in the model?

[18] Unlike $\hat{\delta}^2$, the exact distribution of F_1 only depends on ξ and Λ but not $\nu'\nu$. Details on the exact distribution of F_1 are available upon request.

One of the perceived advantages of the HJ-distance is that it does not favor noisy factors because the population HJ-distance stays the same whether we use f_{2t} or f_t^*.[19] However, it turns out that if one uses the sample HJ-distance to test the null hypothesis $H_0 : \delta = 0$, the noisy factors will still be favored. This is because with a noisy factor, there are more estimation errors in $\hat{D}$. As a result, even though the population HJ-distance stays the same with the noisy factor, one can be less sure about whether an observed large sample HJ-distance is due to genuine nonzero population HJ-distance or to sampling errors. In Figure 4, we plot the power function for five different models using f_t^*, differing in terms of their noise to signal ratio σ_n/σ_f, where σ_f is the standard deviation of the second factor. Similar to Figure 3, the size of the test is 5% and the distribution of $\hat{\delta}^2$ under the null and the alternative are computed using the exact distribution, assuming the sample mean and the variance of the factor are equal to their population counterparts. The case of $\sigma_n/\sigma_f = 0$ is the same as the model with just the second factor in Figure 3, and we can see that it is rejected with very high probability. As σ_n increases, the probability of rejection goes down. When $\sigma_n/\sigma_f = 5$, we see that the model with this very noisy factor is hardly rejected, with a probability of rejection of only 8.38% even for $T = 1000$. From an economics point of view, there is no reason to believe the model with this very noisy factor is any better than the model with the original factor, but yet statistically the sample HJ-distance suggests otherwise.

Figure 4 about here

D. Understanding the Biases in Asymptotic Distribution

Our small sample results are not only useful in providing insights on what determine the power of the test of $H_0 : \delta = 0$, but also illuminating for understanding the biases of the asymptotic test that is often used in the literature for testing the null hypothesis. Jagannathan and Wang (1996) show that under the null hypothesis, we have

$$T\hat{\delta}^2 \overset{A}{\sim} \sum_{i=1}^{N-K-1} a_i \chi_1^2, \tag{86}$$

[19] For example, some model diagnostic tools like the HJ-bound suggested by Hansen and Jagannathan (1991) can favor noisy factors. This is because for any arbitrary factor model, it is always possible to satisfy the HJ-bound by adding irrelevant variance to the stochastic discount factor.

which is a linear combination of $N-K-1$ independent χ^2_1 random variables, with the weights a_i equal to the nonzero eigenvalues of

$$S^{\frac{1}{2}}U^{-\frac{1}{2}}[I_N - U^{-\frac{1}{2}}D(D'U^{-1}D)^{-1}D'U^{-\frac{1}{2}}]U^{-\frac{1}{2}}S^{\frac{1}{2}}, \tag{87}$$

or equivalently the eigenvalues of

$$P'U^{-\frac{1}{2}}SU^{-\frac{1}{2}}P, \tag{88}$$

where P is an $N\times(N-K-1)$ orthonormal matrix with its columns orthogonal to $U^{-\frac{1}{2}}D$. Under the conditional homoskedasticity assumption, we can use Lemma 3 to verify that $a_i = (1+\gamma_1'V_{11}^{-1}\gamma_1)/\gamma_0^2$ for $i = 1,\ldots,N-K-1$, and the asymptotic distribution can be simplified to

$$T\hat{\delta}^2 \overset{A}{\sim} \left(\frac{1+\gamma_1'V_{11}^{-1}\gamma_1}{\gamma_0^2}\right)\chi^2_{N-K-1}, \tag{89}$$

which is consistent with the results in Proposition 3. Similar to the exact finite sample distribution, the asymptotic tests also involve nuisance parameters, so we need to obtain estimates of these parameters in order to carry out the asymptotic tests. In practice, researchers replace D, U, and S in (87) with their sample estimates to obtain the estimated eigenvalues $\hat{a}_i$. Similarly, we can replace γ_0, γ_1 and V_{11} in (89) with their sample estimates $\hat{\gamma}_0^{HJ}$, $\hat{\gamma}_1^{HJ}$ and $\hat{V}_{11}$. We refer to asymptotic tests that are based on estimated parameters as the approximate asymptotic tests.

Recall from (73) and Proposition 4, the exact distribution of $T\hat{\delta}^2$ is given by

$$T\hat{\delta}^2 = \left(\frac{T}{V}\right)\left(\frac{T\nu'\nu}{1+U_1'W^{-1}U_1}\right), \tag{90}$$

where $U_1 \sim N(\xi, I_{K+1})$, $W \sim W_{K+1}(N-1, I_{K+1}, \Lambda)$, $V \sim \chi^2_{T-N+1}$, and they are independent of each other. The first term T/V is due to the estimation error from using $\hat{\Sigma}$, which the asymptotic test ignores. The expectation of T/V is given by

$$E\left[\frac{T}{V}\right] = \frac{T}{T-N-1} > 1, \tag{91}$$

so the effect of ignoring this first term is that the finite sample distribution of $T\hat{\delta}^2$ is larger than the finite sample distribution of the second term, which is $T\tilde{\delta}^2$. As a result, ignoring the estimation error of $\hat{\Sigma}$ can lead to over-rejection problem for the asymptotic test and this problem is particularly severe when N is large relative to T. When N is small relative to T, the first term is negligible and the finite sample distribution of $T\hat{\delta}^2$ is almost identical to the finite sample distribution of $T\tilde{\delta}^2$. By

comparing the finite sample distribution of $T\tilde{\delta}^2$ with the asymptotic distribution, we can obtain an understanding of why the asymptotic test can also under-reject the null hypothesis. To facilitate this comparison, we present the following lemma.

Lemma 5 *Let $U_{1,K+1}$ be the last element of U_1 and $W^{K+1,K+1}$ be the last diagonal element of W^{-1}. Under the null hypothesis, we have $1/W^{K+1,K+1} \sim \chi^2_{N-K-1}$ and*

$$1 + U_1' W^{-1} U_1 = U_{1,K+1}^2 W^{K+1,K+1} + O_p(T^{\frac{1}{2}}), \tag{92}$$

with the expected value of the $O_p(T^{\frac{1}{2}})$ term positive. In addition, we have

$$E[U_{1,K+1}^2] = 1 + \xi_{K+1}^2 > \xi_{K+1}^2 = \frac{T\gamma_0^2(\nu'\nu)}{1 + \bar{\gamma}_1' \hat{V}_{11}^{-1} \bar{\gamma}_1}, \tag{93}$$

where $\bar{\gamma}_1 = \gamma_1 + \hat{\mu}_1 - \mu_1$.

With this lemma, we can think of the asymptotic distribution of $T\tilde{\delta}^2$ is obtained by using the following approximation

$$T\tilde{\delta}^2 = \frac{T\nu'\nu}{1 + U_1' W^{-1} U_1} \approx \frac{T\nu'\nu}{\xi_{K+1}^2 W^{K+1,K+1}} \sim \left(\frac{1 + \bar{\gamma}_1' \hat{V}_{11}^{-1} \bar{\gamma}_1}{\gamma_0^2} \right) \chi^2_{N-K-1}. \tag{94}$$

Such an approximation involves dropping all but one of the $(K+1)^2+1$ terms in $1+U_1'W^{-1}U_1$, and replacing $U_{1,K+1}^2$ by ξ_{K+1}^2. The terms that we drop in this approximation have positive expected value, so ignoring them in the denominator of $T\tilde{\delta}^2$ leads to an under-rejection problem for the asymptotic test. The greater the number of factors (K) is, the more terms are dropped in this approximation and the more severe the under-rejection problem. This analysis suggests that, given the number of test assets and length of time series, the asymptotic test tends to favor models with a large number of factors.

IV. Simulation Evidence

A. Design of Experiment

We perform simulation experiments in this section to assess the performance of the asymptotic test, the F-test, and the approximate finite sample test of sample HJ-distance. We study various combinations of number of factors (K), number of test assets (N) and number of time series

observations (T). For the number of factors, we consider three cases: $K = 1$, 3, and 5. For the number of test assets, we also consider three cases: $N = 10$, 25, and 100. For the length of time series observations, we consider five cases: $T = 120$, 240, 360, 480, and 600. Altogether, there are 45 different combinations of K, N, and T in our simulation experiment.

For $K = 1$, the factor that we consider is the excess return of the value-weighted market index from the Center for Research in Security Prices (CRSP). For $K = 3$, the factors are the three factors of Jagannathan and Wang (1996): return on the value-weighted market, growth of per capita labor income, and default premium which is the difference between the yields of Baa and Aaa corporate bonds. For $K = 5$, the factors are the five factors of Fama and French (1993). They include excess return on the value-weighted market index, return difference between portfolios of small and large stocks, return difference between portfolios of high and low book-to-market ratio stocks, a term structure factor which is measured by the difference of the yields of long-term Treasury bond and short-term Treasury bill, and a default premium which is measured as the difference between the yields of Baa and Aaa corporate bonds. Monthly data of these factors are kindly provided to us by Hodrick and Zhang, and the data are the same as the ones that are used in Hodrick and Zhang (2001). As the finite sample distribution that we derive is conditional on the realizations of the factors, we consider the case where the matrix $(X'X)/T$ is set to

$$\frac{X'X}{T} = \begin{bmatrix} 1 & \hat{\mu}_1' \\ \hat{\mu}_1 & \hat{V}_{11} + \hat{\mu}_1\hat{\mu}_1' \end{bmatrix}, \tag{95}$$

where $\hat{\mu}_1$ is the sample mean of the factors and $\hat{V}_{11}$ is the sample variance of the factors, estimated from the monthly data over the sample period of January 1951 to December 1997. Under this setting, we retain the same matrix $(X'X)/T$ for different values of T.

The test assets that we consider for $N = 10$ are the ten size-ranked portfolios of common stocks in the New York Stock Exchange (NYSE). For $N = 25$, the test assets are 25 size and book-to-market ranked portfolios of Fama and French (1993). For $N = 100$, the test assets are 100 size and beta ranked portfolios of common stocks in the NYSE. Data for the 25 size and book-to-market ranked portfolios are provided to us by Hodrick and Zhang, whereas the data for other test assets are constructed from the CRSP monthly files. To obtain the nuisance parameters for our simulation, we need $B = [\alpha,\ \beta]$ and Σ. Both β and Σ are set equal to their sample estimates obtained from

the regressions using monthly data from January 1951 to December 1997. As for α, we set

$$\alpha = 1_N\gamma_0 + \hat{\beta}(\gamma_1 - \hat{\mu}_1), \tag{96}$$

where γ_0 and γ_1 are obtained using a GLS CSR of $\hat{\mu}_2$ on 1_N and $\hat{\beta}$. Under this choice of α, the parameters are set such that the null hypothesis is true (i.e., $\delta = 0$). Finally, we derive the nuisance parameters $1_N'\Sigma^{-1}1_N$, Λ and ξ for each case based on our chosen values of $(X'X)/T$, B and Σ, and the exact distribution of $\hat{\delta}^2$ can then be computed. The exact distribution of $\hat{\delta}^2$ is evaluated using the Monte Carlo integration approach based on 100,000 simulations. For most practical purposes, the errors due to simulations are negligible.

B. Actual Sizes of Various Tests

In Table I, we present the actual probabilities of rejection using the asymptotic test in (89) for three different levels of significance, under the assumption that the null hypothesis of $H_0 : \delta = 0$ is true. Table I shows that when $K = 1$ and $N = 10$, the probabilities of rejection of the asymptotic test are very close to the size of the test. However, when N increases, we find that there is a significant over-rejection problem from using the asymptotic test. For example, for the case of $K = 1$ and $N = 100$, we find that even for T as large as 600, the asymptotic test rejects the null hypothesis 38.6% of the time, when the asymptotic size of the test is supposed to be 5%.

For a fixed N, we find that the rejection rate of the asymptotic test goes down as we increase K. For the case of $N = 10$, we find that the asymptotic test significantly under-rejects the null hypothesis when K is large. For example, when $K = 5$ and $N = 10$, the asymptotic test rejects the null hypothesis only 1.6% of the time, when the asymptotic size of the test is supposed to be 5%. The fact that the probability of rejection of the asymptotic test is a decreasing function of K has two implications. One implication is that even though additional factor does not help explaining expected returns, they will be favored by the asymptotic test. Another implication is that even though a model with a large number of factors is wrong, the asymptotic test can have low power to detect it.

Overall, there are two countering effects. When N is relatively large to T, the asymptotic test tends to over-reject. When K is relatively large to N, the asymptotic test tends to under-reject. Which of the two effects dominates depends on the values of K, N, T and the nuisance parameters.

Table I about here

With an understanding of the performance of the asymptotic test, we now turn our attention to the approximate F-test. In Table II, we present the actual probabilities of rejection using the approximate F-test in (84) for three different levels of significance, under the assumption that the null hypothesis of $H_0 : \delta = 0$ is true. Overall, the approximate F-test is much better behaved than the asymptotic test. When $K = 1$, the actual probability of rejection is almost identical to the size of the test. In addition, the over-rejection problem associated with large number of test assets for the asymptotic test is basically gone. This is expected because the over-rejection problem in the asymptotic test is due to the estimation error in $\hat{\Sigma}$, which the approximate F-test fully takes into account in its denominator using the χ^2_{T-N+1} random variable. When K is large relative to N, there is still some under-rejection problem with the approximate F-test, especially when T is small.

Table II about here

As the exact distribution of $\hat{\delta}^2$ involves some unknown nuisance parameters, practical use of it requires the estimation of these nuisance parameters. In Section III.C, we describe a procedure for estimating these nuisance parameters. It is of interest to investigate to what extent that the use of estimated nuisance parameters distorts the size of the test. In Table III, we present the actual probabilities of rejection using this approximate finite sample test based on the estimated nuisance parameters for three different levels of significance, under the assumption that the null hypothesis of $H_0 : \delta = 0$ is true. When $K = 1$, the actual probability of rejection is almost identical to the size of the test, indicating that the effect of using the estimated nuisance parameters is negligible. When $K = 3$ or 5, we find that the approximate finite sample test has some under-rejection and over-rejection problems, especially when N is large. Comparing the approximate finite sample test with the approximate F-test, it is not clear which test is better behaved. However, when T is small, the approximate finite sample test appears to have an advantage because it results in less of an under-rejection problem than the approximate F-test. Nevertheless, both of them are decisively better than the asymptotic test.

Table III about here

C. Nonnormal Reisduals

While the small sample distribution of $\hat{\delta}^2$ assumes that the residuals are multivariate normally distributed, we have good reasons to believe that it works fairly well even though the residuals are not normally distributed. For tests of mean-variance efficiency of a given portfolio, the work of MacKinlay (1985) and Zhou (1993) shows that although the F-test of Gibbons, Ross, and Shanken (1989) relies on multivariate normality assumption of the residuals, it is rather robust to departure from normality of the residuals. To examine if this is still the case for the finite sample distribution of $\hat{\delta}^2$, we repeat the same simulation experiment as before except that the residuals are now generated using a multivariate t-distribution with five degrees of freedom and with the same variance-covariance matrix as in the normal case. Under this multivariate t-distribution assumption, the residuals and hence the returns have fat tails, which is what we often find in the data. In each simulation, we generate data under the null hypothesis and test $H_0 : \delta = 0$ based on the exact distribution under the normality assumption. In Table IV, we present the rejection rates for various combinations of K, N and T. As we can see from Table IV, we find that the rejection rates based on our exact distribution are extremely close to the size of the test, despite the residuals are very different from normal. This robustness result is not surprising because while $\hat{B}$ is not exactly normal and $\hat{\Sigma}$ is not exactly Wishart when the residuals are not multivariate normally distributed, such approximations are in fact quite good even for moderate size of T. As a result, we can still reasonably apply our small sample test even for cases that the residuals are not multivariate normally distributed.[20]

Table IV about here

V. Conclusion

In this paper, we have conducted a comprehensive analysis of the HJ-distance for the case of linear asset pricing models. We have also provided a geometric interpretation of the HJ-distance and showed that it is a measure of how close the minimum-variance frontier of the test assets is to the minimum-variance frontier of the factor mimicking positions, but the distance is normalized by the

[20]Although not reported, we have repeated our simulation experiment for a few other nonnormal distributions of residuals, and the results are largely the same.

zero-beta rate. A comparison of the sample HJ-distance with Shanken's CSRT statistic revealed that the fundamental difference between the regression approach and the stochastic discount factor approach to tests of asset pricing models is in the choice of the estimated zero-beta rate. Under normality assumption, we have provided an analysis of the exact distribution of the sample HJ-distance. In addition, a simple and efficient numerical method to obtain the finite sample distribution of the sample HJ-distance was presented. Simulation evidence has shown that asymptotic distribution for sample HJ-distance is grossly inappropriate when the number of test assets or the number of factors is large. For finite sample inference, one is better off using the exact distribution presented in this paper.

Despite the theoretical appeal of the population HJ-distance, researchers should be cautious in using the sample HJ-distance for model evaluation and selection. We have shown that models with small HJ-distance are good in explaining prices of the test assets but not necessary good in explaining their expected returns. In addition, we have found that the sample HJ-distance is not much different from many traditional specification tests. As a result, the sample HJ-distance shares the same problems that plagued those specification tests. Specifically, our analysis and simulation have shown that the sample HJ-distance tends to favor asset pricing models that have noisy factors and is not very reliable in telling apart good models from bad models.

Appendix

We first present two matrix identities that will be used repeatedly in the Appendix.[21]

Claim: Suppose $Q = P + BCB'$, where P and Q are $m \times m$ nonsingular matrices, B is an $m \times p$ matrix with full column rank, and C is a $p \times p$ matrix. Then we have

$$(B'Q^{-1}B)^{-1}B'Q^{-1} = (B'P^{-1}B)^{-1}B'P^{-1}, \tag{A1}$$

$$Q^{-1} - Q^{-1}B(B'Q^{-1}B)^{-1}B'Q^{-1} = P^{-1} - P^{-1}B(B'P^{-1}B)^{-1}B'P^{-1}. \tag{A2}$$

Proof: Since

$$Q^{-1} = P^{-1} - P^{-1}BC(I_p + B'P^{-1}BC)^{-1}B'P^{-1}, \tag{A3}$$

we have

$$B'Q^{-1} = (I_p + B'P^{-1}BC)^{-1}B'P^{-1} \tag{A4}$$

and

$$(B'Q^{-1}B)^{-1} = (B'P^{-1}B)^{-1}(I_p + B'P^{-1}BC). \tag{A5}$$

Multiplying (A4) with (A5), we have the first identity

$$(B'Q^{-1}B)^{-1}B'Q^{-1} = (B'P^{-1}B)^{-1}B'P^{-1}. \tag{A6}$$

For the second identity, we have

$$\begin{aligned} & Q^{-1} - Q^{-1}B(B'Q^{-1}B)^{-1}B'Q^{-1} \\ = \; & Q^{-1}[I_m - B(B'Q^{-1}B)^{-1}B'Q^{-1}] \\ = \; & [P^{-1} - P^{-1}BC(I_p + B'P^{-1}BC)^{-1}B'P^{-1}][I_m - B(B'P^{-1}B)^{-1}B'P^{-1}] \\ = \; & P^{-1} - P^{-1}B(B'P^{-1}B)^{-1}B'P^{-1}, \end{aligned} \tag{A7}$$

with the second last equality following from (A3) and (A6). This completes the proof. *Q.E.D.*

Proof of Lemma 1: Observe that we can write

$$U = \Sigma + D \begin{bmatrix} 1 + \mu_1' V_{11}^{-1} \mu_1 & -\mu_1' V_{11}^{-1} \\ -V_{11}^{-1}\mu_1 & V_{11}^{-1} \end{bmatrix} D' \tag{A8}$$

[21] A single-factor version of these two identities were presented in Shanken (1982).

and $D = [\mu_2,\ V_{21} + \mu_2\mu_1'] = HA$, where A is a nonsingular matrix given by

$$A = \begin{bmatrix} 1 & \mu_1' \\ 0_K & V_{11} \end{bmatrix}. \tag{A9}$$

Letting $P = \Sigma$, $Q = U$, and $B = D$, we can invoke (A2) and have

$$\begin{aligned} U^{-1} - U^{-1}D(D'U^{-1}D)^{-1}D'U^{-1} &= \Sigma^{-1} - \Sigma^{-1}D(D'\Sigma^{-1}D)^{-1}D'\Sigma^{-1} \\ &= \Sigma^{-1} - \Sigma^{-1}HA(A'H'\Sigma^{-1}HA)^{-1}A'H'\Sigma^{-1} \\ &= \Sigma^{-1} - \Sigma^{-1}H(H'\Sigma^{-1}H)^{-1}H'\Sigma^{-1}. \end{aligned} \tag{A10}$$

Putting this expression in (10), we obtain (12). This completes the proof. *Q.E.D.*

Proof of Lemma 2: Suppose μ_m and V_m are the mean and variance of R_1, and q_m is a vector of the cost of these K factor mimicking positions. When $K > 1$, a minimum-variance portfolio (with unit cost) of the K factor mimicking positions is obtained by solving the following problem:

$$\begin{aligned} \min_w \sigma_p^2 &= w'V_m w \\ \text{s.t.} \quad w'\mu_m &= \mu_p, \qquad \text{(A11)} \\ w'q_m &= 1. \qquad \text{(A12)} \end{aligned}$$

Except using q_m instead of 1_K, it is the same as the standard portfolio optimization problem. Standard derivation then gives (20) with $a_1 = \mu_m'V_m^{-1}\mu_m$, $b_1 = \mu_m'V_m^{-1}q_m$ and $c_1 = q_m'V_m^{-1}q_m$. Using $\mu_m = V_{12}V_{22}^{-1}\mu_2$, $V_m = \text{Var}[R_1] = V_{12}V_{22}^{-1}V_{21}$ and $q_m = V_{12}V_{22}^{-1}1_N$, we obtain the expressions for a_1, b_1 and c_1. When $K = 1$, we must have $w = 1/q_m$ and hence $\mu_p = \mu_m/q_m = b_1/c_1$ and $\sigma_p^2 = V_m/q_m^2 = 1/c_1$. This completes the proof. *Q.E.D.*

Proof of Proposition 1: One way to prove (24) is to express D and U in terms of μ and V. For brevity, we present a more intuitive proof here. Writing $\lambda = [\lambda_0,\ \lambda_1']'$, where λ_0 is a scalar and λ_1 is a K-vector, the squared HJ-distance is given by

$$\delta^2 = \min_\lambda (D\lambda - 1_N)'U^{-1}(D\lambda - 1_N). \tag{A13}$$

Since $D = E[R_2x'] = [\mu_2,\ V_{21} + \mu_2\mu_1']$ and

$$U = E[R_2R_2'] = V_{22} + \mu_2\mu_2' = V_{22} + D\begin{bmatrix} 1 & 0_K' \\ 0_K & \text{O}_{K\times K} \end{bmatrix} D', \tag{A14}$$

we can invoke (A1) and (A2) and write

$$\begin{aligned}\delta^2 &= \min_{\lambda}(D\lambda - 1_N)'V_{22}^{-1}(D\lambda - 1_N) \\ &= \min_{\lambda}(\mu_2\lambda_0 + V_{21}\lambda_1 + \mu_2\mu_1'\lambda_1 - 1_N)'V_{22}^{-1}(\mu_2\lambda_0 + V_{21}\lambda_1 + \mu_2\mu_1'\lambda_1 - 1_N). \qquad \text{(A15)}\end{aligned}$$

Using a reparameterization of λ to γ where

$$\gamma \equiv \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \frac{1}{\lambda_0 + \mu_1'\lambda_1} \begin{bmatrix} 1 \\ -V_{11}\lambda_1 \end{bmatrix}, \qquad \text{(A16)}$$

we can then write

$$\delta^2 = \min_{\gamma_0,\gamma_1} \frac{(\mu_2 - 1_N\gamma_0 - \beta\gamma_1)'V_{22}^{-1}(\mu_2 - 1_N\gamma_0 - \beta\gamma_1)}{\gamma_0^2}. \qquad \text{(A17)}$$

Conditional on a given choice of γ_0, one only needs to choose γ_1 to minimize the numerator. It is easy to show that

$$\gamma_1^* = (\beta'V_{22}^{-1}\beta)^{-1}\beta'V_{22}^{-1}(\mu_2 - 1_N\gamma_0). \qquad \text{(A18)}$$

With this choice of γ_1, we can minimize the objective function with respect to γ_0 alone and have

$$\delta^2 = \min_{\gamma_0} \frac{(\mu_2 - 1_N\gamma_0)'[V_{22}^{-1} - V_{22}^{-1}\beta(\beta'V_{22}^{-1}\beta)^{-1}\beta'V_{22}^{-1}](\mu_2 - 1_N\gamma_0)}{\gamma_0^2} = \min_{\gamma_0} \frac{\theta_2^2(\gamma_0) - \theta_1^2(\gamma_0)}{\gamma_0^2}. \qquad \text{(A19)}$$

Using

$$\theta_2^2(\gamma_0) - \theta_1^2(\gamma_0) = a - 2b\gamma_0 + c\gamma_0^2 - (a_1 - 2b_1\gamma_0 + c_1\gamma_0^2) = \Delta a - 2\Delta b\gamma_0 + \Delta c\gamma_0^2, \qquad \text{(A20)}$$

we have

$$\frac{\theta_2^2(\gamma_0) - \theta_1^2(\gamma_0)}{\gamma_0^2} = \Delta a \left(\frac{1}{\gamma_0}\right)^2 - 2\Delta b \left(\frac{1}{\gamma_0}\right) + \Delta c, \qquad \text{(A21)}$$

which is a quadratic function in $1/\gamma_0$. The minimum is obtained at $\gamma_0^{HJ} = \Delta a/\Delta b$ and hence $\delta^2 = \left(\theta_2^2(\gamma_0^{HJ}) - \theta_1^2(\gamma_0^{HJ})\right)/(\gamma_0^{HJ})^2$.

As for Q_C, we have conditional on a given value of γ_0, the expected return errors are $e_{CS}(\gamma_1) = (\mu_2 - \gamma_0 1_N) - \beta\gamma_1$. It is easy to see that

$$\min_{\gamma_1} e_{CS}(\gamma_1)'\Sigma^{-1}e_{CS}(\gamma_1) = (\mu_2 - 1_N\gamma_0)'[\Sigma^{-1} - \Sigma^{-1}\beta(\beta'\Sigma^{-1}\beta)^{-1}\beta'\Sigma^{-1}](\mu_2 - 1_N\gamma_0). \qquad \text{(A22)}$$

Since $\Sigma = V_{22} - \beta V_{11}\beta'$, invoking the identity (A2), we have

$$\begin{aligned} & (\mu_2 - 1_N\gamma_0)'[\Sigma^{-1} - \Sigma^{-1}\beta(\beta'\Sigma^{-1}\beta)^{-1}\beta'\Sigma^{-1}](\mu_2 - 1_N\gamma_0) \\ = \; & (\mu_2 - 1_N\gamma_0)'[V_{22}^{-1} - V_{22}^{-1}\beta(\beta'V_{22}^{-1}\beta)^{-1}\beta'V_{22}^{-1}](\mu_2 - 1_N\gamma_0) \\ = \; & \theta_2^2(\gamma_0) - \theta_1^2(\gamma_0) \\ = \; & \Delta a - 2\Delta b\gamma_0 + \Delta c\gamma_0^2. \qquad \text{(A23)}\end{aligned}$$

The γ_0 that minimizes this expression is $\gamma_0^{CS} = \Delta b/\Delta c$, and hence Q_C is given by $\theta_2^2(\gamma_0^{CS}) - \theta_1^2(\gamma_0^{CS})$. Finally, since $\Delta a - 2\Delta b\gamma_0 + \Delta c\gamma_0^2 \geq 0$ for any γ_0, the determinant of the quadratic equation must be nonpositive and we have $(\Delta b)^2 \leq \Delta a \Delta c$. Since $\Delta a > 0$ and $\Delta c > 0$, we have $\Delta a/\Delta b \geq \Delta b/\Delta c$ if $\Delta b \geq 0$, and $\Delta a/\Delta b \leq \Delta b/\Delta c$ if $\Delta b < 0$. This completes the proof. *Q.E.D.*

Proof of Proposition 2: The proof of Proposition 2 is identical to the proof of Proposition 1. All we need is to replace all the population moments in the proof of Proposition 1 with their sample counterparts.

Proof of Lemma 3: When $g_t(\lambda)$ is a martingale difference sequence, we have

$$S = \text{Var}[R_{2t}x_t'\lambda - 1_N] = \text{Var}[R_{2t}x_t'\lambda] = \text{Var}[(Bx_t + \epsilon_t)x_t'\lambda] = \text{Var}[\epsilon_t x_t'\lambda] + B\text{Var}[x_t x_t'\lambda]B'. \quad \text{(A24)}$$

Under the conditional homoskedasticity assumption, we have

$$\text{Var}[\epsilon_t x_t'\lambda] = E[\text{Var}[\epsilon_t x_t'\lambda|x_t]] + \text{Var}[E[\epsilon_t x_t'\lambda|x_t]] = E[(x_t'\lambda)^2\Sigma] + \text{O}_{N\times N} = E[(x_t'\lambda)^2]\Sigma \quad \text{(A25)}$$

and hence

$$S = E[(x_t'\lambda)^2]\Sigma + B\text{Var}[x_t x_t'\lambda]B'. \quad \text{(A26)}$$

When Y_t follows a multivariate elliptical distribution, we have the following results from the Proposition 2 of Kan and Zhou (2002)

$$\text{Var}[x_t \otimes \epsilon_t] = E[x_t x_t' \otimes \epsilon_t\epsilon_t'] = E[x_t x_t'] \otimes \Sigma + \begin{bmatrix} 0 & 0_K' \\ 0_K & \kappa V_{11} \end{bmatrix} \otimes \Sigma. \quad \text{(A27)}$$

Using this result, we can write S as

$$\begin{aligned} S &= \text{Var}[(\lambda' \otimes I_N)(x_t \otimes \epsilon_t)] + B\text{Var}[x_t x_t'\lambda]B' \\ &= (\lambda' \otimes I_N)\left(E[x_t x_t'] \otimes \Sigma + \begin{bmatrix} 0 & 0_K' \\ 0_K & \kappa V_{11} \end{bmatrix} \otimes \Sigma\right)(\lambda \otimes I_N) + B\text{Var}[x_t x_t'\lambda]B' \\ &= \left(E[(x_t'\lambda)^2] + \kappa\lambda_1' V_{11}\lambda_1\right)\Sigma + B\text{Var}[x_t x_t'\lambda]B'. \end{aligned} \quad \text{(A28)}$$

This completes the proof. *Q.E.D.*

Proof of Proposition 3: From the proof of Lemma 1, it is easy to see that

$$\hat{\lambda}^{HJ} = (\hat{D}'\hat{\Sigma}^{-1}\hat{D})^{-1}(\hat{D}'\hat{\Sigma}^{-1}1_N), \quad \text{(A29)}$$

$$\hat{\delta}^2 = 1_N'[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}\hat{D}(\hat{D}'\hat{\Sigma}^{-1}\hat{D})^{-1}\hat{D}'\hat{\Sigma}^{-1}]1_N. \quad \text{(A30)}$$

Let $P = \hat{a}\hat{\Sigma}$, $Q = \hat{S}$. Note that we have $\hat{B} = \hat{D}A$, where

$$A = \begin{bmatrix} 1 + \hat{\mu}_1 \hat{V}_{11}^{-1} \hat{\mu}_1 & -\hat{\mu}_1' \hat{V}_{11}^{-1} \\ -\hat{V}_{11}^{-1} \hat{\mu}_1 & \hat{V}_{11}^{-1} \end{bmatrix}, \tag{A31}$$

so we can write $Q = P + \hat{D}A\hat{C}A'\hat{D}'$ and invoke (A1) to obtain

$$\hat{\lambda}^{GMM} = (\hat{D}'(\hat{a}\hat{\Sigma})^{-1}\hat{D})^{-1}(\hat{D}'(\hat{a}\hat{\Sigma})^{-1}1_N) = (\hat{D}'\hat{\Sigma}^{-1}\hat{D})^{-1}(\hat{D}'\hat{\Sigma}^{-1}1_N) = \hat{\lambda}^{HJ}. \tag{A32}$$

Similarly, we can invoke (A2) to obtain

$$\begin{aligned} J &= T1_N'[(\hat{a}\hat{\Sigma})^{-1} - (\hat{a}\hat{\Sigma})^{-1}\hat{D}(\hat{D}'(\hat{a}\hat{\Sigma})^{-1}\hat{D})^{-1}\hat{D}'(\hat{a}\hat{\Sigma})^{-1}]1_N \\ &= \frac{T1_N'[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}\hat{D}(\hat{D}'\hat{\Sigma}^{-1}\hat{D})^{-1}\hat{D}'\hat{\Sigma}^{-1}]1_N}{\hat{a}} = \frac{T\hat{\delta}^2}{\hat{a}}. \end{aligned} \tag{A33}$$

This completes the proof. *Q.E.D.*

Proof of Lemma 4: Consider the following matrix

$$\hat{A} = [1_N,\ \hat{B}(X'X)^{\frac{1}{2}}]'\hat{\Sigma}^{-1}[1_N,\ \hat{B}(X'X)^{\frac{1}{2}}]. \tag{A34}$$

Using Theorem 3.2.11 of Muirhead (1982), we have conditional on $\hat{B}$,

$$\hat{A}^{-1} \sim W_{K+2}(T - N + 1, \tilde{A}^{-1}/T), \tag{A35}$$

where

$$\tilde{A} = [1_N,\ \hat{B}(X'X)^{\frac{1}{2}}]'\Sigma^{-1}[1_N,\ \hat{B}(X'X)^{\frac{1}{2}}]. \tag{A36}$$

Now, using Corollary 3.2.6 of Muirhead (1982) and noting that the $(1,1)$ element of $\hat{A}^{-1}$ is $1/\hat{\delta}^2$ whereas the $(1,1)$ element of $\tilde{A}^{-1}$ is $1/\tilde{\delta}^2$, conditional on $\hat{B}$, we have

$$\frac{1}{\hat{\delta}^2} \sim W_1(T - N + 1, \frac{1}{T\tilde{\delta}^2}), \tag{A37}$$

and therefore

$$\frac{T\tilde{\delta}^2}{\hat{\delta}^2} \sim \chi^2_{T-N+1}. \tag{A38}$$

Finally, since this conditional distribution does not depend on $\hat{B}$, this is also the unconditional distribution and in addition the ratio is also independent of $\tilde{\delta}^2$ (which is a function of $\hat{B}$). This completes the proof. *Q.E.D.*

Proof of Proposition 4: Define $P = [P_1,\ P_2]$ as an $N \times N$ orthonormal matrix with its first column equal to

$$P_1 = \frac{\nu}{(\nu'\nu)^{\frac{1}{2}}}. \tag{A39}$$

Since the columns of P_2 form an orthonormal basis for the space orthogonal to P_1, this implies

$$P_2 P_2' = I_N - \nu(\nu'\nu)^{-1}\nu' \tag{A40}$$

and

$$Q'B_n'P_2P_2'B_nQ = Q'B_n'[I_N - \nu(\nu'\nu)^{-1}\nu']B_nQ = Q'Q\Lambda Q'Q = \Lambda. \tag{A41}$$

Let $U \equiv [U_1,\ U_2] = [Q'\tilde{B}_n'P_1,\ Q'\tilde{B}_n'P_2]$, we have $\text{vec}(U) \sim N(\text{vec}(Q'B_n'P), I_N \otimes I_{K+1})$. Specifically, we have $E[U_1] = Q'B_n'\nu/(\nu'\nu)^{\frac{1}{2}} = \xi$, $E[U_2] = Q'B_n'P_2$, with U_1 and U_2 independent of each other. Using these transformations and writing $W = U_2U_2' \sim W_{K+1}(N-1, I_{K+1}, \Lambda)$, we have

$$\begin{aligned}
\tilde{\delta}^2 &= \nu'[I_N - \tilde{B}_nQ(Q'\tilde{B}_n'PP'\tilde{B}_nQ)^{-1}Q'\tilde{B}_n']\nu \\
&= \nu'\nu[1 - P_1'\tilde{B}_nQ(UU')^{-1}Q'\tilde{B}_n'P_1] \\
&= \nu'\nu[1 - U_1'(U_1U_1' + W)^{-1}U_1].
\end{aligned} \tag{A42}$$

Using the identity

$$(U_1U_1' + W)^{-1} = W^{-1} - \frac{W^{-1}U_1U_1'W^{-1}}{1 + U_1'W^{-1}U_1}, \tag{A43}$$

we have

$$\tilde{\delta}^2 = \nu'\nu\left[1 - U_1'W^{-1}U_1 + \frac{(U_1'W^{-1}U_1)^2}{1 + U_1'W^{-1}U_1}\right] = \frac{\nu'\nu}{1 + U_1'W^{-1}U_1}. \tag{A44}$$

Performing the same exercise on δ^2, we have

$$\begin{aligned}
\delta^2 &= \nu'[I_N - B_nQ(Q'B_n'PP'B_nQ)^{-1}Q'B_n']\nu \\
&= \nu'\nu[1 - P_1'B_nQ(Q'B_n'P_1P_1'B_nQ + Q'B_n'P_2P_2'B_nQ)^{-1}Q'B_n'P_1] \\
&= \nu'\nu[1 - \xi'(\xi\xi' + \Lambda)^{-1}\xi] \\
&= \frac{\nu'\nu}{1 + \xi'\Lambda^{-1}\xi}.
\end{aligned} \tag{A45}$$

This completes the proof. *Q.E.D.*

Proof of (80): Under the null hypothesis, 1_N is in the span of the columns of B, so ν is in the span of the columns of B_n. Writing $\nu = B_nh$, where h is a $(K+1)$-vector, we have $h = (B_n'B_n)^{-1}B_n'\nu$. As

$$B_n'[I_N - \nu(\nu'\nu)^{-1}\nu']B_nh = B_n'B_nh - B_n'B_nh(h'B_n'B_nh)^{-1}h'B_n'B_nh = 0_N, \tag{A46}$$

so h is proportional to q_{K+1}, which is the eigenvector associated with the zero eigenvalue of $B_n'[I_N - \nu(\nu'\nu)^{-1}\nu']B_n$, and we have $q_{K+1} = \pm h/(h'h)^{\frac{1}{2}}$. Therefore,

$$\xi_{K+1}^2 = \frac{(q_{K+1}'B_n'\nu)^2}{\nu'\nu} = \frac{(h'B_n'\nu)^2}{(h'h)(\nu'\nu)} = \frac{\nu'\nu}{h'h} = \frac{\nu'\nu}{\nu'B_n(B_n'B_n)^{-2}B_n'\nu}. \tag{A47}$$

This completes the proof. *Q.E.D.*

Proof of Proposition 5: From (73) and the definition of noncentral F-distribution, it suffices to show that $T\tilde{\delta}^2$ is approximately distributed as

$$\left(\frac{1 + \hat{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\hat{\gamma}_1^{HJ}}{(\hat{\gamma}_0^{HJ})^2}\right)\chi_{N-K-1}^2(d). \tag{A48}$$

Since $D = B(X'X)/T$, we can write

$$1_N = D\lambda^{HJ} + e_{HJ} = B(X'X)^{\frac{1}{2}}h + e^{HJ} \tag{A49}$$

by defining $h = (X'X)^{\frac{1}{2}}\lambda^{HJ}/T$. It follows that

$$\nu = \Sigma^{-\frac{1}{2}}B(X'X)^{\frac{1}{2}}h + \Sigma^{-\frac{1}{2}}e_{HJ} = B_nh + \Sigma^{-\frac{1}{2}}e_{HJ} = \tilde{B}_nh + (B_n - \tilde{B}_n)h + \Sigma^{-\frac{1}{2}}e_{HJ}. \tag{A50}$$

Since the first term is a linear combination of $\tilde{B}_n$, it will vanish when it is multiplied by $I_N - \tilde{B}_n(\tilde{B}_n'\tilde{B}_n)^{-1}\tilde{B}_n'$. Therefore, we can write

$$\tilde{\delta}^2 = \nu'[I_N - \tilde{B}_n(\tilde{B}_n'\tilde{B}_n)^{-1}\tilde{B}_n']\nu = Y'[I_N - \tilde{B}_n(\tilde{B}_n'\tilde{B}_n)^{-1}\tilde{B}_n']Y, \tag{A51}$$

where

$$Y = (B_n - \tilde{B}_n)h + \Sigma^{-\frac{1}{2}}e_{HJ} \sim N\left(\Sigma^{-\frac{1}{2}}e_{HJ}, (h'h)I_N\right). \tag{A52}$$

Note that $I_N - \tilde{B}_n(\tilde{B}_n'\tilde{B}_n)^{-1}\tilde{B}_n'$ is idempotent with rank $N - K - 1$. If we ignore the fact that Y and $\tilde{B}_n$ are correlated (which is a good approximation when K is small relative to N),[22] then we have

$$T\tilde{\delta}^2 \sim T(h'h)\chi_{N-K-1}^2\left(\frac{e_{HJ}'\Sigma^{-1}e_{HJ}}{h'h}\right). \tag{A53}$$

Using the reparameterization of

$$\lambda^{HJ} = \frac{1}{\gamma_0^{HJ}}\begin{bmatrix} 1 + \hat{\mu}_1'\hat{V}_{11}^{-1}\gamma_1^{HJ} \\ -\hat{V}_{11}^{-1}\gamma_1^{HJ} \end{bmatrix}, \tag{A54}$$

[22]Alternatively, we can follow the same argument as in Shanken (1985) by replacing $\tilde{B}_n$ in the idempotent matrix by B_n.

we can simplify $T(h'h)$ to

$$\begin{aligned} T(h'h) &= \lambda^{HJ\prime}\left(\frac{X'X}{T}\right)\lambda^{HJ} \\ &= \begin{bmatrix} \lambda_0^{HJ} \\ \lambda_1^{HJ} \end{bmatrix}' \begin{bmatrix} 1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{V}_{11}+\hat{\mu}_1\hat{\mu}_1' \end{bmatrix} \begin{bmatrix} \lambda_0^{HJ} \\ \lambda_1^{HJ} \end{bmatrix} \\ &= \frac{1+\bar{\gamma}_1^{HJ\prime}\hat{V}_{11}^{-1}\bar{\gamma}_1^{HJ}}{(\gamma_0^{HJ})^2}, \end{aligned} \tag{A55}$$

where $\bar{\gamma}_1^{HJ} = \gamma_1^{HJ} + \hat{\mu}_1 - \mu_1$ is the *ex post* risk premium. Replacing γ_0^{HJ} and $\bar{\gamma}_1^{HJ}$ by their consistent estimates and using the fact that $\delta^2 = e_{HJ}'\Sigma^{-1}e_{HJ}$, we obtain the approximate F-distribution. This completes the proof. *Q.E.D.*

Proof of Lemma 5: Partition W into four blocks

$$W \equiv \begin{bmatrix} A & b \\ b' & c \end{bmatrix} \sim W_{K+1}(N-1, I_{K+1}, \Lambda), \tag{A56}$$

where A is the upper $K \times K$ submatrix of W. Under the null hypothesis, we have $\lambda_{K+1} = 0$. Using Lemma 2.1 in Gleser (1976), we can show that $A \sim W_K(N-1, I_K, \Lambda_1)$, where $\Lambda_1 = \text{Diag}(\lambda_1, ..., \lambda_K)$, $c - b'A^{-1}b \sim \chi^2_{N-K-1}$, $z = A^{-\frac{1}{2}}b \sim N(0_K, I_K)$, and they are independent of each other. Using the partitioned matrix inverse formula, the last diagonal element of W^{-1} is given by

$$W^{K+1,K+1} = (c - b'A^{-1}b)^{-1} \sim \frac{1}{\chi^2_{N-K-1}}. \tag{A57}$$

As $X'X = O_p(T)$, we have $B_n = O_p(T^{\frac{1}{2}})$ and this implies $\lambda_i = O_p(T)$ for $i = 1, ..., K$ and $\xi = O_p(T^{\frac{1}{2}})$. It follows that $A = O_p(T)$ and $A^{-\frac{1}{2}} = O_p(T^{-\frac{1}{2}})$. Since the upper $K \times K$ submatrix of W^{-1} is

$$A^{-1} + A^{-1}bW^{K+1,K+1}b'A^{-1} = A^{-\frac{1}{2}}[I_K + zW^{K+1,K+1}z']A^{-\frac{1}{2}}, \tag{A58}$$

its elements are $O_p(T^{-1})$. Similarly, the first K elements of the last column of W^{-1} is

$$-A^{-1}bW^{K+1,K+1} = -A^{-\frac{1}{2}}zW^{K+1,K+1} = O_p(T^{-\frac{1}{2}}). \tag{A59}$$

As $\xi = O_p(T^{\frac{1}{2}})$, we have $U_1 = O_p(T^{\frac{1}{2}})$ and

$$1 + U_1'W^{-1}U_1 = U_{1,K+1}^2 W^{K+1,K+1} + O_p(T^{\frac{1}{2}}). \tag{A60}$$

Note that the $O_p(T^{\frac{1}{2}})$ term in the above expression is

$$1 + \sum_{i=1}^{K} \sum_{j=1}^{K} W^{ij} U_{1i} U_{1j} + 2U_{1,K+1} \sum_{i=1}^{K} W^{i,K+1} U_{1i}, \tag{A61}$$

where W^{ij} is the (i,j) element of W^{-1} and U_{1i} is the ith element of U_1. The second term in this expression is a quadratic form, so it is positive. For the last term, the first K elements of the last column of W^{-1} is given by $A^{-\frac{1}{2}} z W^{K+1,K+1}$. As $E[z] = 0_K$ and it is independent of U_1, A, and $W^{K+1,K+1}$, the last term has zero expected value. Therefore, the $O_p(T^{\frac{1}{2}})$ term has a positive expected value.

Finally, $U_{1,K+1}^2 \sim \chi_1^2(\xi_{K+1}^2)$, so its expectation is $1 + \xi_{K+1}^2$. Under the null hypothesis, we have from the proof of (80) that $\xi_{K+1}^2 = (\nu'\nu)/(h'h)$, where $h = (B_n' B_n)^{-1}(B_n'\nu)$. Then, using a similar proof of Proposition 5, we obtain $T(h'h) = (1 + \bar{\gamma}_1' \hat{V}_{11}^{-1} \bar{\gamma}_1)/\gamma_0^2$. This completes the proof. *Q.E.D.*

References

Ahn, Seung, and Christopher Gadarowski, 2004, Small sample properties of the GMM specification test based on the Hansen-Jagannathan distance, *Journal of Empirical Finance* 11, 109–132.

Bansal, Ravi, and S. Viswanathan, 1993, No arbitrage and arbitrage pricing: A new approach, *Journal of Finance* 48, 1231–1262.

Black, Fischer, 1972, Capital market equilibrium with restricted borrowing, *Journal of Business* 45, 444–454.

Breeden, Douglas T., 1979, An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* 7, 265–296.

Cheung, C. Sherman, Clarence C. Y. Kwan, and Dean C. Mountain, 2000, Spanning tests of asset pricing: a minimal encompassing approach, working paper, McMaster University.

Campbell, John Y., and John H. Cochrane, 2000, Explaining the poor performance of consumption-based asset pricing models, *Journal of Finance* 55, 2863–2878.

Cochrane, John H., 1996, A cross-sectional test of an investment-based asset pricing model, *Journal of Political Economy* 104, 572–621.

Dittmar, Robert F., 2002, Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns, *Journal of Finance* 57, 369–403.

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

Farebrother, R. W., 1984, The distribution of a positive linear combination of χ^2 random variables, *Applied Statistics* 33, 332–339.

Farnsorth, Heber, Wayne E. Ferson, David Jackson, and Steven Todd, 2002, Performance evaluation with stochastic discount factors, *Journal of Business* 75, 473–503.

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica* 57, 1121–1152.

Gleser, Leon Jay, 1976, A canonical representation for the noncentral Wishart distribution useful for simulation, *Journal of the American Statisical Association* 71, 690–695.

Grinblatt, Mark, and Sheridan Titman, 1987, The relation between mean-variance efficiency and arbitrage pricing, *Journal of Business* 60, 97–112.

Hansen, Lars Peter, 1982, Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029–1054.

Hansen, Lars Peter, and Ravi Jagannathan, 1991, Implications of security market data for models of dynamic economies, *Journal of Political Economy* 99, 225–262.

Hansen, Lars Peter, and Ravi Jagannathan, 1997, Assessing specification errors in stochastic discount factor model, *Journal of Finance* 52, 557–590.

Hodrick, Robert J., and Xiaoyan Zhang, 2001, Evaluating the specification errors of asset pricing models, *Journal of Financial Economics* 62, 327–376.

Huberman, Gur, and Shmuel Kandel, 1987, Mean-variance spanning, *Journal of Finance* 42, 873–888.

Huberman, Gur, Shmuel Kandel, and Robert F. Stambaugh, 1987, Mimicking portfolios and exact arbitrage pricing, *Journal of Finance* 42, 1–9.

Jagannathan, Ravi, Keiichi Kubota, and Hitoshi Takehara, 1998, Relationship between labor-income risk and average return: empirical evidence from the Japanese stock market, *Journal of Business* 71, 319–348.

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

Kan, Raymond, and Guofu Zhou, 2002, Tests of mean-variance spanning, working paper, University of Toronto and Washington University in St. Louis.

Lettau, Martin, and Sidney Ludvigson, 2001, Resurrecting the (C)CAPM: a cross-section test when risk premia are time-varying, *Journal of Political Economy* 109, 1238–1287.

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

MacKinlay, A. Craig, 1985, Analysis of multivariate financial tests, Ph.D. dissertation, Graduate School of Business, University of Chicago.

Merton, Robert C., 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.

Muirhead, Robb J., 1982, *Aspects of multivariate statistical theory* (Wiley, New York).

Roll, Richard, 1985, A note on the geometry of Shanken's CSR T^2 test for mean/variance efficiency, *Journal of Financial Economics* 14, 349–357.

Ross, Stephen A., 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–360.

Shanken, Jay, 1982, An analysis of the traditional risk-return model, unpublished Ph.D. dissertation (Carnegie-Mellon University, Pittsburgh, PA).

Shanken, Jay, 1985, Multivariate tests of the zero-beta CAPM, *Journal of Financial Economics* 14, 327–348.

Shanken, Jay, 1986, Testing portfolio efficiency when the zero-beta rate is unknown: a note, *Journal of Finance* 41, 269–276.

Sharpe, William F., 1964, Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* 19, 425–442.

Velu, Raja, and Guofu Zhou, 1999, Testing multi-beta asset pricing models, *Journal of Empirical Finance* 6, 219–241.

Zhou, Guofu, 1993, Asset pricing test under alternative distributions, *Journal of Finance* 48, 1925–1942.

Zhou, Guofu, 1995, Small sample rank tests with applications to asset pricing, *Journal of Empirical Finance* 2, 71–93.

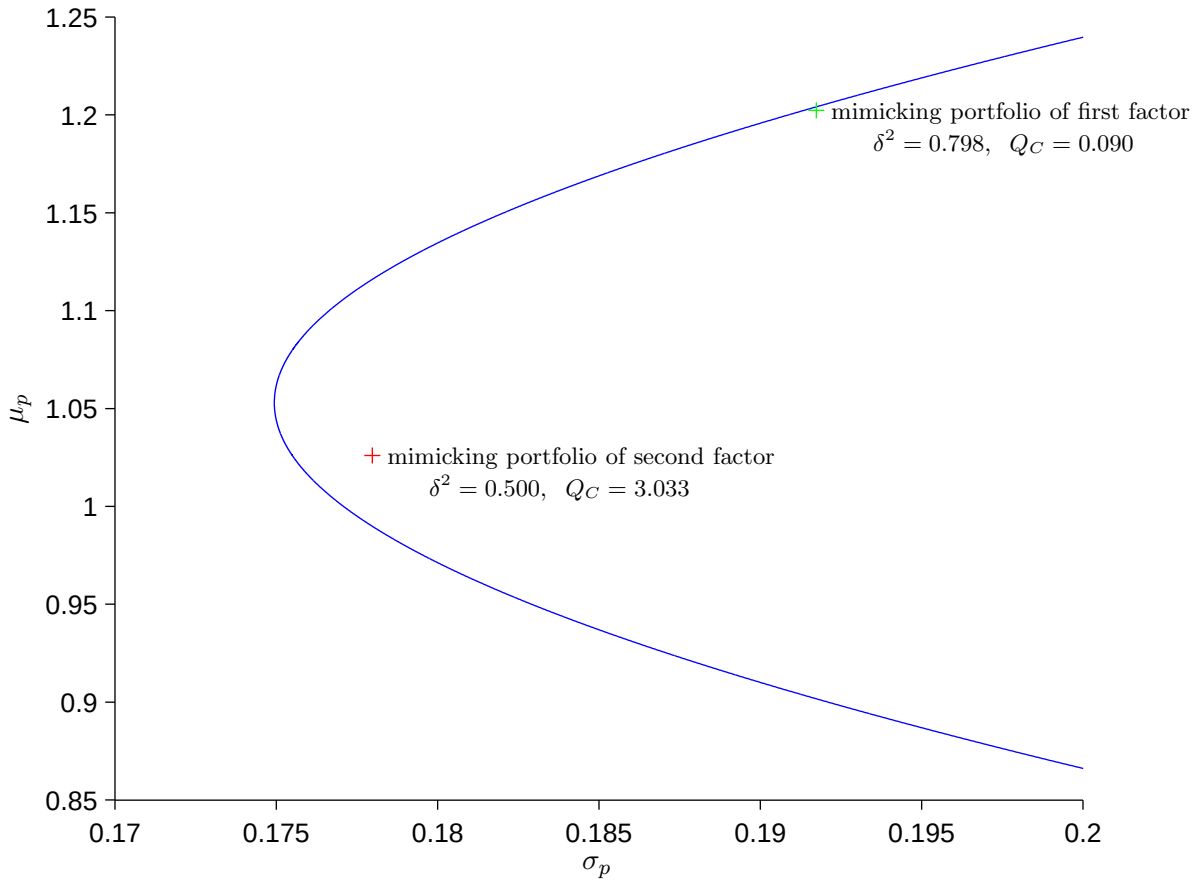

Figure 1
Rankings of Two Models Using HJ-Distance and Aggregate Expected Return Errors
The figure plots the two factor mimicking portfolios as well as the minimum-variance frontier hyperbola of four test assets. The mimicking portfolio of the first factor produces small errors in expected returns but large pricing errors for the four test assets. The mimicking portfolio of the second factor produces large errors in expected returns but small pricing errors for the four test assets.

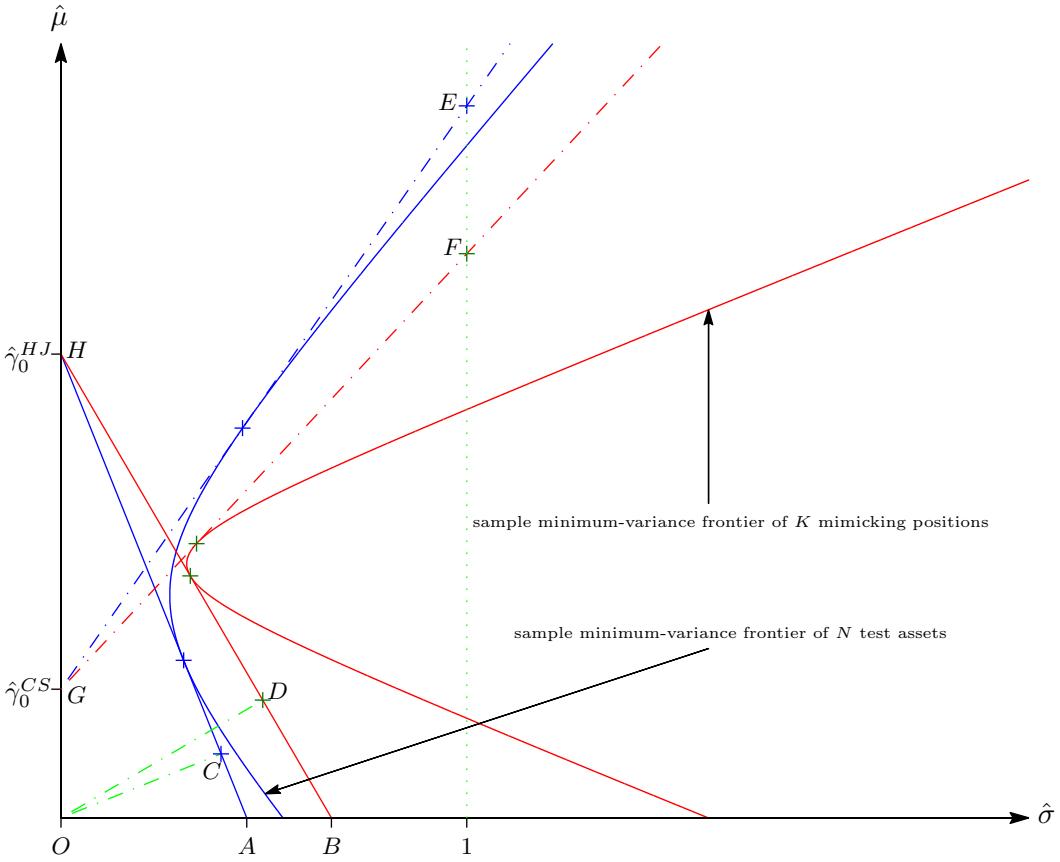

Figure 2
The Geometry of Hansen-Jagannathan Distance and CSRT Statistic
The figure plots the *ex post* minimum-variance frontier hyperbola of K mimicking positions and that of N test assets on the $(\hat{\sigma}, \hat{\mu})$ space. $\hat{\gamma}_0^{HJ}$ is the estimated zero-beta rate that minimizes the sample HJ-distance. The straight line HA is the tangent line to the frontier of the N test assets and its slope is equal to $\hat{\theta}_2(\hat{\gamma}_0^{HJ})$. Point C is the point on the tangent line HA that is closest to the origin. The straight line HB is the tangent line to the frontier of the K mimicking positions and its slope is equal to $\hat{\theta}_1(\hat{\gamma}_0^{HJ})$. Point D is the point on the tangent line HB that is closest to the origin. The squared sample Hansen-Jagannathan distance is given by $1/(OA)^2 - 1/(OB)^2$ or $1/(OC)^2 - 1/(OD)^2$. $\hat{\gamma}_0^{CS}$ is the estimated zero-beta rate from a generalized least squares cross-sectional regression of $\hat{\mu}_2$ on 1_N and $\hat{\beta}$. The straight line GE is the tangent line to the frontier of the N test assets and its slope is equal to $\hat{\theta}_2(\hat{\gamma}_0^{CS})$. The length of GE is $\sqrt{1+\hat{\theta}_2^2(\hat{\gamma}_0^{CS})}$. The straight line GF is the tangent line to the frontier of the K mimicking positions and its slope is equal to $\hat{\theta}_1(\hat{\gamma}_0^{CS})$. The length of GF is $\sqrt{1+\hat{\theta}_1^2(\hat{\gamma}_0^{CS})}$. The CSRT statistic is equal to $GE^2 - GF^2$.

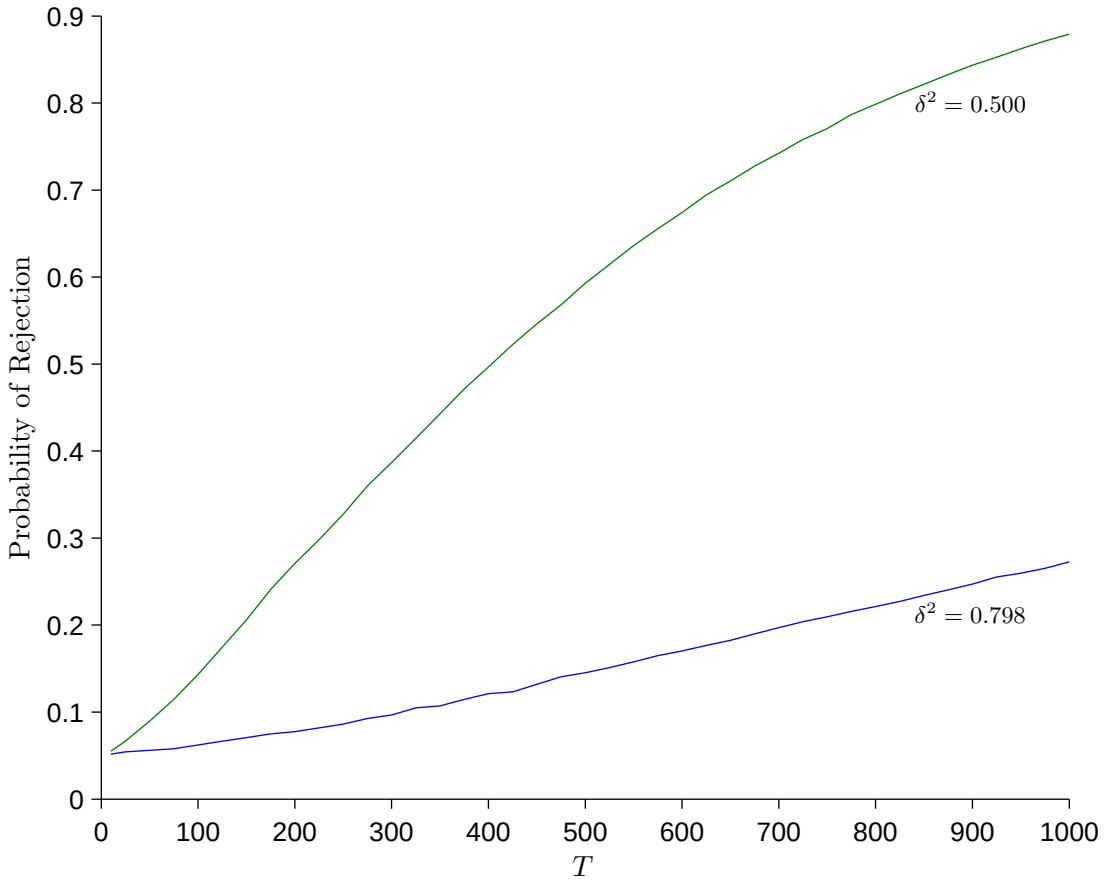

Figure 3
Power Function of Sample HJ-Distance for Two Models
The figure plots the probabilities of rejecting the null hypothesis $H_0 : \delta = 0$ as a function of the length of time series observations (T) for two different models. The returns of four test assets are generated using a two factor model. The first model contains only the first factor and it has a $\delta^2 = 0.798$. The second model contains only the second factor and it has a $\delta^2 = 0.500$. The size of the test is 5% and the distributions of the sample HJ-distance under the null and the alternatives are computed using the exact distribution, assuming the sample mean and the sample variance of the factors are equal to their population counterparts.

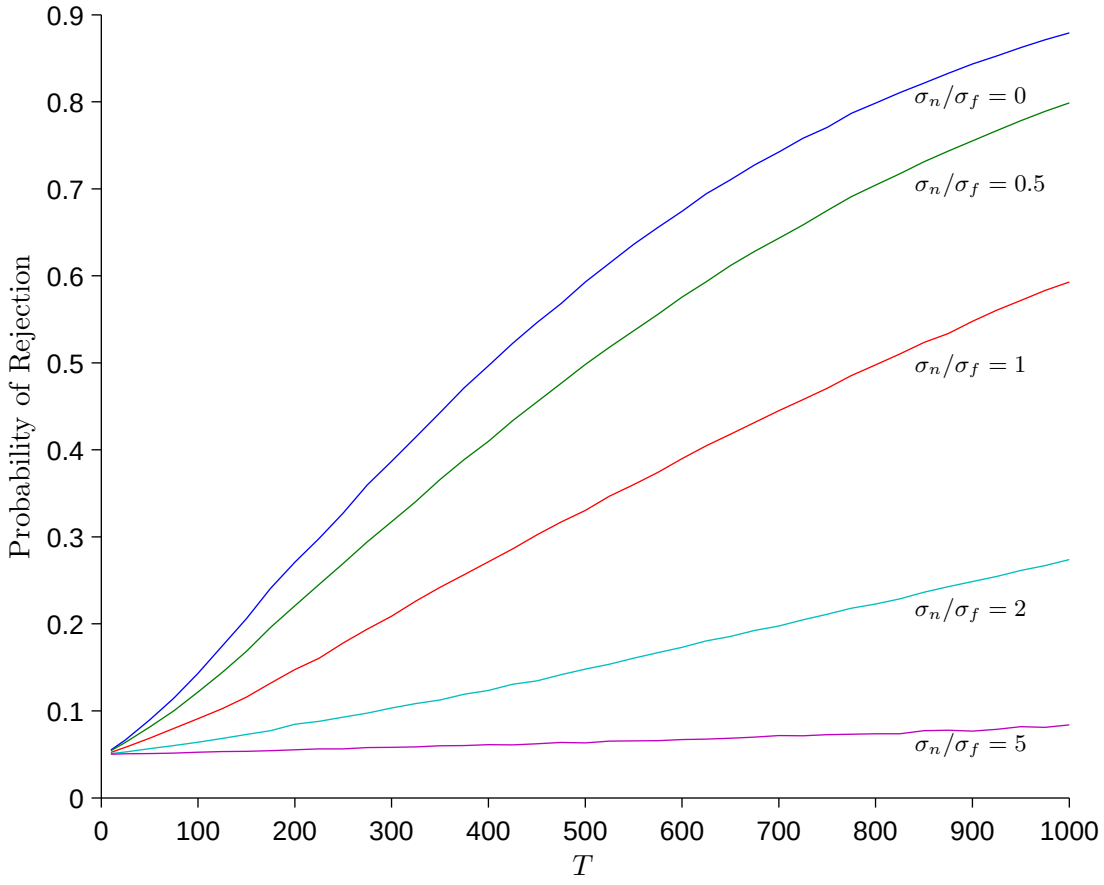

Figure 4
Power Function of Sample HJ-Distance for Models with Noisy Factor
The figure plots the probabilities of rejecting the null hypothesis $H_0 : \delta = 0$ as a function of the length of time series observations (T) for five different models. The returns of four test assets are generated using a two factor model. A noisy version of the second factor is used in each one of the five tested models, which is generated as $f_t^* = f_{2t} + n_t$, where n_t is a measurement error with mean zero and independent of the returns and the factors. All five models have the same $\delta^2 = 0.500$ but they differ in terms of their noise to signal ratio σ_n/σ_f, where σ_n/σ_f is the ratio of the standard deviation of the noise to the standard deviation of the second factor. The size of the test is 5% and the distributions of the sample HJ-distance under the null and the alternatives are computed using the exact distribution, assuming the sample mean and the sample variance of the factors are equal to their population counterparts.

Table I
Sizes of Asymptotic Test of HJ-Distance Under Normality

The table presents the actual probabilities of rejection for the asymptotic χ^2-test of $H_0 : \delta = 0$ with different levels of significance under the null hypothesis, assuming the residuals are generated from a multivariate normal distribution. The rejection decision is based on an asymptotic χ^2-test of the sample HJ-distance assuming conditional homoskedasticity. Results for different values of the number of factors (K), test assets (N), and time series observations (T) are based on 100,000 simulations.

		$N = 10$			$N = 25$			$N = 100$		
		Level of Significance			Level of Significance			Level of Significance		
K	T	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	120	0.117	0.063	0.016	0.343	0.242	0.108	1.000	1.000	1.000
	240	0.103	0.053	0.011	0.201	0.122	0.038	0.971	0.950	0.880
	360	0.101	0.051	0.011	0.163	0.094	0.026	0.803	0.715	0.518
	480	0.100	0.051	0.011	0.144	0.080	0.021	0.634	0.516	0.302
	600	0.099	0.050	0.010	0.137	0.075	0.018	0.510	0.386	0.192
3	120	0.040	0.017	0.002	0.300	0.205	0.085	1.000	1.000	1.000
	240	0.040	0.017	0.002	0.176	0.104	0.031	0.950	0.918	0.821
	360	0.045	0.019	0.003	0.142	0.078	0.020	0.738	0.637	0.428
	480	0.049	0.020	0.003	0.130	0.071	0.017	0.555	0.434	0.233
	600	0.054	0.024	0.003	0.123	0.065	0.015	0.439	0.321	0.146
5	120	0.023	0.008	0.001	0.230	0.148	0.054	1.000	1.000	1.000
	240	0.025	0.009	0.001	0.128	0.071	0.018	0.933	0.893	0.778
	360	0.029	0.011	0.001	0.105	0.054	0.012	0.690	0.582	0.372
	480	0.036	0.014	0.001	0.095	0.048	0.010	0.504	0.384	0.194
	600	0.041	0.016	0.002	0.090	0.045	0.009	0.394	0.279	0.120

Table II
Sizes of Approximate F-test of HJ-Distance Under Normality

The table presents the actual probabilities of rejection for the approximate F-test of $H_0 : \delta = 0$ with different levels of significance under the null hypothesis, assuming the residuals are generated from a multivariate normal distribution. The rejection decision is based on an approximate F-test of the sample HJ-distance. Results for different values of the number of factors (K), test assets (N), and time series observations (T) are based on 100,000 simulations.

		$N = 10$			$N = 25$			$N = 100$		
		Level of Significance			Level of Significance			Level of Significance		
K	T	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	120	0.093	0.047	0.010	0.094	0.046	0.009	0.092	0.045	0.009
	240	0.101	0.052	0.011	0.097	0.049	0.010	0.098	0.048	0.010
	360	0.104	0.053	0.012	0.098	0.049	0.010	0.098	0.049	0.010
	480	0.104	0.054	0.012	0.099	0.049	0.010	0.100	0.050	0.011
	600	0.105	0.053	0.012	0.100	0.050	0.010	0.100	0.050	0.010
3	120	0.052	0.024	0.004	0.062	0.028	0.005	0.058	0.025	0.004
	240	0.071	0.034	0.007	0.072	0.034	0.006	0.099	0.050	0.010
	360	0.084	0.043	0.009	0.077	0.037	0.007	0.108	0.055	0.012
	480	0.094	0.050	0.011	0.081	0.039	0.007	0.111	0.058	0.013
	600	0.101	0.054	0.013	0.122	0.063	0.014	0.113	0.059	0.013
5	120	0.034	0.015	0.002	0.056	0.026	0.004	0.045	0.020	0.003
	240	0.052	0.024	0.005	0.067	0.032	0.006	0.097	0.048	0.009
	360	0.069	0.034	0.007	0.076	0.037	0.007	0.110	0.056	0.012
	480	0.084	0.043	0.010	0.081	0.039	0.008	0.117	0.061	0.013
	600	0.096	0.051	0.012	0.085	0.042	0.008	0.120	0.062	0.014

Table III
Sizes of Approximate Finite Sample Test of HJ-Distance Under Normality

The table presents the actual probabilities of rejection for the approximate finite sample test of $H_0 : \delta = 0$ with different levels of significance under the null hypothesis, assuming the residuals are generated from a multivariate normal distribution. The rejection decision is based on an approximate finite sample test of the sample HJ-distance using estimated nuisance parameters. Results for different values of the number of factors (K), test assets (N), and time series observations (T) are based on 100,000 simulations.

		$N = 10$			$N = 25$			$N = 100$		
		Level of Significance			Level of Significance			Level of Significance		
K	T	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	120	0.117	0.060	0.012	0.097	0.048	0.009	0.093	0.047	0.009
	240	0.113	0.058	0.011	0.099	0.050	0.010	0.098	0.049	0.010
	360	0.110	0.056	0.010	0.099	0.049	0.010	0.100	0.050	0.010
	480	0.108	0.054	0.010	0.099	0.050	0.010	0.100	0.051	0.010
	600	0.105	0.054	0.010	0.100	0.050	0.010	0.099	0.050	0.010
3	120	0.097	0.043	0.006	0.141	0.072	0.014	0.079	0.037	0.007
	240	0.116	0.055	0.009	0.135	0.070	0.014	0.149	0.078	0.017
	360	0.124	0.059	0.010	0.122	0.062	0.012	0.147	0.077	0.017
	480	0.129	0.063	0.011	0.112	0.057	0.011	0.142	0.074	0.017
	600	0.131	0.064	0.011	0.106	0.053	0.010	0.138	0.074	0.015
5	120	0.063	0.025	0.002	0.141	0.073	0.014	0.070	0.032	0.005
	240	0.084	0.036	0.004	0.149	0.078	0.017	0.164	0.088	0.020
	360	0.100	0.044	0.006	0.146	0.076	0.017	0.165	0.091	0.021
	480	0.109	0.050	0.007	0.141	0.074	0.015	0.164	0.089	0.021
	600	0.116	0.054	0.007	0.138	0.072	0.015	0.159	0.087	0.020

Table IV
Sizes of Finite Sample Test of HJ-Distance Under Nonnormality of Residuals

The table presents the actual probabilities of rejection for the finite sample test of $H_0 : \delta = 0$ with different levels of significance under the null hypothesis, assuming the residuals are generated from a multivariate Student-t distribution with five degrees of freedom. The rejection decision is based on the simulated exact distribution of the sample HJ-distance under the normality assumption. Results for different values of the number of factors (K), test assets (N), and time series observations (T) are based on 100,000 simulations.

		$N = 10$			$N = 25$			$N = 100$		
		Level of Significance			Level of Significance			Level of Significance		
K	T	10%	5%	1%	10%	5%	1%	10%	5%	1%
1	120	0.097	0.046	0.009	0.092	0.043	0.007	0.095	0.046	0.009
	240	0.097	0.048	0.009	0.094	0.045	0.008	0.089	0.041	0.007
	360	0.099	0.049	0.010	0.096	0.046	0.009	0.090	0.042	0.007
	480	0.098	0.050	0.009	0.095	0.046	0.009	0.090	0.042	0.008
	600	0.100	0.049	0.010	0.096	0.047	0.009	0.090	0.043	0.008
3	120	0.099	0.049	0.010	0.094	0.044	0.008	0.099	0.048	0.009
	240	0.099	0.049	0.009	0.095	0.046	0.008	0.096	0.046	0.008
	360	0.101	0.051	0.010	0.097	0.047	0.009	0.095	0.046	0.008
	480	0.100	0.050	0.010	0.097	0.048	0.009	0.095	0.046	0.008
	600	0.100	0.049	0.010	0.097	0.048	0.009	0.094	0.045	0.008
5	120	0.100	0.049	0.009	0.096	0.046	0.008	0.100	0.049	0.010
	240	0.099	0.050	0.010	0.095	0.046	0.008	0.098	0.047	0.009
	360	0.101	0.050	0.009	0.096	0.047	0.009	0.099	0.046	0.009
	480	0.100	0.050	0.010	0.099	0.048	0.009	0.096	0.047	0.009
	600	0.100	0.050	0.010	0.098	0.048	0.009	0.097	0.047	0.009