Queen's Economics Department Working Paper No. 1091

Size Matters: Covariance Matrix Estimation Under the Alternative

Jason Allen
Bank of Canada; Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

8-2005

Size Matters: Covariance Matrix Estimation Under the Alternative

Jason Allen*

Department of Economics, Queen's University at Kingston

Department of Monetary and Financial Analysis, The Bank of Canada

August 18, 2005

ABSTRACT

The purpose of this paper is to investigate, using Monte Carlo methods, whether or not Hall's (2000) centered test of overidentifying restrictions for parameters estimated by Generalized Method of Moments (GMM) is more powerful, once the test is size-adjusted, than the standard test introduced by Hansen (1982). The Monte Carlo evidence shows that very little size-adjusted power is gained over the standard uncentered calculation. Empirical examples using Epstein and Zin (1991) preferences demonstrates that the centered and uncentered tests sometimes lead to different conclusions about model specification.

Keywords: Size, Power, Generalized Method of Moments, Overidentifying Restrictions.

JEL *classification:* C15, C52, G12

*MFA, Bank of Canada, Ottawa, ON, K1A 0G9, Phone: 613-782-8712, email: allenj@qed.econ.queensu.ca. I have benefited from discussions with Allan Gregory, comments by Gregor Smith, Don Andrews, Kim Huynh, Thanasis Stengos, and Greg Tkacz, and appreciate Russell Davidson's comments on an earlier version of this paper. All errors are my own. The views presented are those of the author. No responsibility should be attributed to the Bank of Canada.

1. Introduction

Generalized Method of Moments (GMM) (Hansen, 1982) is widely used in applied economics to estimate and test macroeconomic models. In terms of testing for model misspecification, the most popular test is Hansen's (1982) J-test for overidentifying restrictions. While the test has widespread use, Altonji and Segal (1996), and Hall and Horowitz (1996), among others, have documented that it frequently over-rejects in small samples. Most recent research efforts have been directed at improving the size performance of the J-test. On the other hand, Hall (2000) proposes a centered version of the popular J-test with the aim of gaining power. Smith (1999) provides Monte Carlo evidence that the standard J-test can have low size-adjusted power in a nonlinear asset pricing model. Hall (2000)'s Monte Carlo evidence shows a modest power gain of 10 percent assuming a linear data generating process.

Hall observes that the standard estimator of the long-run covariance matrix with unknown heteroscedasticity and serial correlation (HAC) used in Hansen's test statistics is consistent only under the null. If the null hypothesis is false, power could fall from using an inconsistent estimate. While power is not usually a problem in most practical situations, it is worthwhile to pursue the performance of the statistic under the alternative. Hall uses a two-step procedure to construct a new HAC matrix. The procedure amounts to subtracting the sample mean of the moment conditions from the second step weighting matrix. This gives a consistent estimator both under the null and alternative hypotheses. Hall's Monte Carlo experiments show a considerable size distortion of the centered calculation over the uncentered calculation as well as a 10 percent increase in the number of rejections when the null is false.

This paper presents Monte Carlo evidence on the properties of the two test statistics to show that the size-adjusted power gain of Hall's test is not significantly greater than the standard J-test. Two experimental designs are considered. The first is identical to Hall's and involves independent data. The second experiment augments the data generating process with serially correlated data, which better mimics most applications. Unfortunately, the Monte Carlo findings indicate that the centered J-test once adjusted for its size distortion is only a marginal improvement over the uncentered counterpart.

In addition to the Monte Carlo evidence, a nonlinear macroeconomic model assuming Epstein and Zin (1991) preferences is estimated as a way of showing the possibility of empirical discrepancies between the standard and centered J-tests which may lead to confusion.

2. Generalized Method of Moments Estimation

GMM estimates the parameters of a model, matching the moments of the theoretical model to those of the data as closely as possible. A weighting matrix determines the relative importance of matching each moment.

Let $\mathbf{X} = (x_1, \ldots, x_t)$, where $x_i \in \mathbb{R}^k$ is a $k \times 1$ random variable, and $t = 1, \ldots T$, be a set of observables from a stationary sequence. Suppose for some true parameter-value θ_0 ($k \times 1$) the following moment conditions (m equations) hold and $m \geq k$:

$$E\left[g(x_t, \theta_0)\right] = 0. \tag{1}$$

This is, of course, the usual set-up for GMM and leads to the estimator:

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} \left(T^{-1} \sum_{t=1}^{T} g(x_t, \theta)\right)' W_T \left(T^{-1} \sum_{t=1}^{T} g(x_t, \theta)\right), \tag{2}$$

where the positive semi-definite weighting matrix, W_T converges to a positive definite matrix of constants. The GMM estimator $\hat{\theta}$ is consistent for any arbitrary weighting matrix, subject to some regularity conditions. Hansen shows that the optimal weighting matrix converges to S_T^{-1} where $S_T = \sum_{j=-\infty}^{\infty} E g_t(\theta_0) g_{t-j}(\theta_0)' \equiv \Gamma_0 + \sum_{j=0}^{\infty} (\Gamma_j + \Gamma_j')$.

The estimation is usually done in two steps. The initial weighting matrix uses the instruments and the final estimate uses the optimal weighting matrix. The centered and uncentered estimators differ in the second stage. The centered estimator subtracts the sample mean from the moment condition which is non-zero in expectation under the alternative. The centered HAC estimator is consistent under the null and alternative, whereas the standard estimator is consistent only under the null.

2.1. Hypothesis Testing

A primary objective of this paper is to compare the size-adjusted power of Hall's centered J-test with that of the standard J-test introduced by Hansen. The standard way of testing is to take the second step estimate of the parameters, $\hat{\theta}_T$, and construct a test statistic that is distributed χ^2_{m-k}:

$$\hat{J}_T = T g_T(\hat{\theta}_T)' \hat{S}_T^{-1} g_T(\hat{\theta}_T). \tag{3}$$

The centered test statistic is essentially the same, except $g_T(\hat{\theta}_T)$ is demeaned for the covariance calculation. However, Hall demonstrates that the centered J-test is more powerful than the standard J-test in large samples. This is because the HAC matrix continues to be consistent under the alternative. I wish to examine the finite sample power issue when the two tests are size-adjusted.

3. Monte Carlo Experiment

Hall demonstrates an asymptotic gain in the divergence of the centered J-test over the uncentered for a local alternative. He also presents Monte Carlo results that suggest this test is 10% more powerful that Hansen's original test. An unrealistic feature of Hall's Monte Carlo design is the independence of the data so that there is no need to calculate a weight matrix based on autocovariances. Hall's comparisons are based on asymptotic critical values for the centered test, even though it has considerably more size distortion than does the standard test. I replicate Hall's experiment below and find that the power gain is almost completely lost once there is an adjustment for size. I also consider simple simulations using a model with dependent data, that nests Hall's experiments:

$$\begin{aligned} y_t &= x_t + \gamma z_{1,t} + \mu_t, \\ x_t &= z_{1,t} + z_{2,t} + \varepsilon_t, \\ z_{1,t} &= \rho_1 z_{1,t-1} + \omega_{1,t}, \\ z_{2,t} &= \rho_2 z_{2,t-1} + \omega_{2,t}, \end{aligned} \tag{4}$$

with $t = 1,2,...,T$, and $(z_{1,t}, z_{2,t}, \mu_t, \varepsilon_t)' \sim N(0,\Sigma)$ with Σ having elements $\sigma_{ii} = 1$, $i = 1,2,3,4$ and $\sigma_{12} = \sigma_{34} = 0.5$. For a model with independent data ρ_1 and ρ_2 are set equal to zero. For dependent data $\rho_1 = \rho_2 = 0.9$. The following moment condition is tested:

$$E[z_t(y_t - x_t\theta)] = 0. \tag{5}$$

The model is estimated by GMM where $\hat{S}_T = \Omega_0 + \sum_{j=1}^{N} \hat{k}_j(\hat{\Omega}_j + \hat{\Omega}_j')$ is a consistent estimator of S_T. The kernel weights are determined using the Newey and West (1994) automatic selection method with $N = c \times int[(T/100)^{2/9}]$ and $c = 4,12$. The moment condition holds only if $\theta = 1$ and $\gamma = 0$. I vary γ from 0.00 to 10.00 to measure the power of the J-test with 10,000 replications and sample size, $T = 300$. Results are presented in table 1 and 2. The median test statistic and the power of each test are reported in Table 1 for *i.i.d* data and Table 2 for serially correlated data. The median of the automatic lag length criteria of Newey and West using the Bartlett kernel is also reported for the standard and centered tests (b, b_c). For both *i.i.d* and serially correlated data, the power of the centered J-test (J_c) is greater than the power of the standard J-test (J). Interestingly, the median test statistic and power of the test statistic are larger when the null hypothesis is true. This is the case even though Hall's procedure is for estimation under the alternative hypothesis. The centered J-test may therefore over-reject a true null hypothesis.

Since the centered test has a greater size distortion than the non-centered test I compare the tests using size-power tradeoff curves as described by Davidson and MacKinnon (1998). Figure 1 to 4 present several of these curves. The dotted line represents values for the centered test and the solid line represents the standard test. The 45° line represents a test with size equal to power. A curve below the 45° line represents a biased test and a curve above the line represents a test with power greater than size. The tradeoff curves are generated by varying the critical value of the tests. For each critical value, size is measured as the percentage of rejections under the null hypothesis, and power is measured as the percentage of rejections under the alternative hypothesis. Thus power is adjusted by size. Figure 1 presents size-power tradeoff curves with $\gamma = 0.125$ for the case of *i.i.d* observations. The centered test is only slightly more powerful than the standard test, and only at very low size. Figure 2 has the same γ but with dependent data. Again the centered test is slightly more powerful than the standard test. The overidentifying restrictions tests are more powerful with independent data than dependent data. Figure

Table 1
Summary Statistics for Overidentifying Restrictions Tests, $\rho_1 = \rho_2 = 0$

c	γ	Med(b_T)	Med(J)	P(J)	Med(b_{cT})	Med(J_c)	P(J_c)
4	0.000	5	0.475	0.045	5	0.482	0.060
	0.125	5	1.330	0.183	5	1.386	0.219
	0.250	6	4.698	0.598	5	5.394	0.639
	0.375	7	9.048	0.926	5	12.170	0.940
	0.500	8	12.703	0.996	5	21.086	0.997
	10.00	13	18.489	1.000	5	111.21	1.0000
12	0.000	14	0.509	0.040	15	0.527	0.085
	0.125	14	1.344	0.166	15	1.508	0.250
	0.250	16	4.193	0.553	15	5.879	0.666
	0.375	20	6.784	0.897	15	13.152	0.944
	0.500	23	8.238	0.993	15	22.628	0.997
	10.00	28	9.806	1.000	15	117.92	1.000

Note: Med$(\cdot)$ denotes the median of the statistic, and $P(\cdot)$ denotes the probability of rejecting the null with nominal size 5 percent. c is a truncation parameter in the estimation of the long-run covariance estimator.

Table 2
Summary Statistics for Overidentifying Restrictions Tests, $\rho_1 = \rho_2 = 0.9$

c	γ	Med(b_T)	Med(J)	P(J)	Med(b_{cT})	Med(J_c)	P(J_c)
4	0.000	6	0.493	0.040	6	0.503	0.058
	0.125	6	2.254	0.294	6	2.431	0.340
	0.250	7	6.432	0.805	6	8.104	0.842
	0.375	10	9.333	0.983	7	14.163	0.990
	0.500	11	10.738	0.999	8	19.085	0.999
	10.00	13	12.298	1.000	12	30.114	1.0000
12	0.000	13	0.539	0.034	15	0.568	0.083
	0.125	15	2.195	0.248	15	2.619	0.368
	0.250	18	5.238	0.742	14	8.296	0.845
	0.375	21	6.786	0.959	14	13.991	0.988
	0.500	23	7.422	0.992	14	18.606	0.999
	10.00	26	8.147	0.999	16	29.245	1.000

Note: Med$(\cdot)$ denotes the median of the statistic, and $P(\cdot)$ denotes the probability of rejecting the null with nominal size 5 percent. c is a truncation parameter in the estimation of the long-run covariance estimator.

3 presents size-power tradeoff curves with $\gamma = 0.250$ for the case of *i.i.d* observations. Figure 4 has $\gamma = 0.250$ and dependent data. The tests are more powerful for independent data than dependent data.

This may be because estimation is more difficult with serially correlated data. The Monte Carlo results indicate that once adjusted for size the centered J-test is not as powerful as initially believed.

4. Empirical Example

GMM is extensively applied in the asset pricing literature. Testing for overidentifying restrictions is a first step in determining whether a model is misspecified. Hansen's J-test is widely used in this context. Many asset pricing models can be written as $1 = E_t[(1+R_{i,t+1})m_{t+1}]$ where m_{t+1} is the pricing kernel or stochastic discount factor. This result lends itself to estimating and testing this class of asset pricing models by GMM. The empirical example follows Epstein and Zin (1991), who assume state-nonseparable preferences. The Epstein and Zin (1991) model is attractive because it breaks the link between risk aversion and intertemporal substitution. This model has been examined extensively in an attempt to explain the equity risk premium puzzle. The stochastic discount factor is given by:

$$m_{t+1} = \left(\beta\left(c_{t+1}/c_t\right)^{-\frac{1}{\phi}}\right)^{\theta}\left(1/R_{m,t+1}\right)^{1-\theta}, \tag{6}$$

where c_{t+1}/c_t is consumption growth, R_m is the risk free rate and $\{\beta,\phi,\theta\}$ parameters to be estimated. The specification separates risk aversion from intertemporal substitution. Epstein and Zin (1991) show that the performance of the model is sensitive to the measure of consumption and choice of instrumental variables. However, for the most part they do not reject the model using the standard tests at conventional levels. Using monthly stock return data from January 1970 to December 2002 I compare Hansen's J-test to Hall's centered J-test. The test statistics are not adjusted for size because the empirical distribution function cannot be traced under both the null and alternative hypotheses. As reported in the Monte Carlo, the size distortions of Hall's centered J-test are larger than the standard J-test. The nonlinearity of the moment conditions in this empirical example almost certainly exacerbates the distortion. Clark (1996) provides Monte Carlo results on size distortions of the J-test for nonlinear models. Since the centered test statistic demeans the moment condition, the estimate of the covariance matrix will be more precise. This increases the over-rejection rate of the centered test statistic.

The Euler equations used in the estimation are:

$$E_t\left[\left(\beta\left(c_{t+1}/c_t\right)^{-\frac{1}{\phi}}\right)^{\theta} R_{m,t+1}^{\theta-1} R_{i,t+1}\right] - 1 = 0 \qquad i = 1,...N.$$

Monthly data sets are constructed for consumption and asset returns. Two measures of consumption are considered: real per capita expenditure growth on nondurables goods and real per capita expenditures on nondurables plus services. For real asset returns four value-weighted indexes are included: returns on the NYSE/AMEX for SIC codes A,B,C; E; F,G; and H,I. The market portfolio (R_m) is a value-weighted index of the NYSE/AMEX returns. There are five equations and three parameters. There are two sets of instruments (Z1,Z2) with nine and thirteen overidentifying restrictions, respectively. Descriptive statistics are presented in table 3 and estimation results in table 4. Similar to Epstein and Zin (1991), the time discount factor, β, is not significantly different from one. The elasticity of intertemporal substitution in consumption, ϕ varies significantly across the set of instruments and is imprecisely estimated. Relative risk aversion is given by α where $\theta = (1-\alpha)/(1-1/\phi)$ and is near one. It is imprecisely estimated when the Treasury bill is used as an instrument.

Table 3
Descriptive Statistics, 1970:1-2002:12

Variable	Nondurables	Nondurables and Services
c_{t+1}/c_t	1.0009 (0.0070)	1.0013 (0.0038)
R_{mt}	1.0062 (0.0547)	1.0057 (0.0452)
R_{1t}	1.0047 (0.0611)	1.0042 (0.0610)
R_{2t}	1.0053 (0.0427)	1.0048 (0.0423)
R_{3t}	1.0069 (0.0575)	1.0064 (0.0569)
R_{4t}	1.0065 (0.0525)	1.0060 (0.0520)
$Tbill_t$	1.0020 (0.0042)	1.0015 (0.0027)

Note: Nominal returns are deflated using the prices corresponding to the definition of consumption. R_{1t} corresponds to the return on SIC codes A,B,C; R_{2t} to the return on SIC code E; R_{3t} to the return on SIC codes F,G; and R_{4t} corresponds to the return on SIC code H,I. Standard errors are in parentheses.

The overidentification tests give similar results. There are competing conclusions about the non-expected utility model at the one or five percent significance level. The centered J-test is larger than

the standard test in every case. The non-expected utility model is rejected at the ten percent level, regardless of whether the centered or uncentered covariance matrix is used.

5. Conclusion

Hall (2000) has suggested using a centered J-test to increase the power of the existing J-test for overidentification. Hall presents results that suggest that his test is 10 percent more powerful that Hansen's (1982) original test. In this paper, I show that in finite samples the gain in power is mostly due to a greater size distortion present in the centered test. The empirical example using Epstein and Zin (1991) preferences illustrates that the two J-tests can yield conflicting conclusions about model specification.

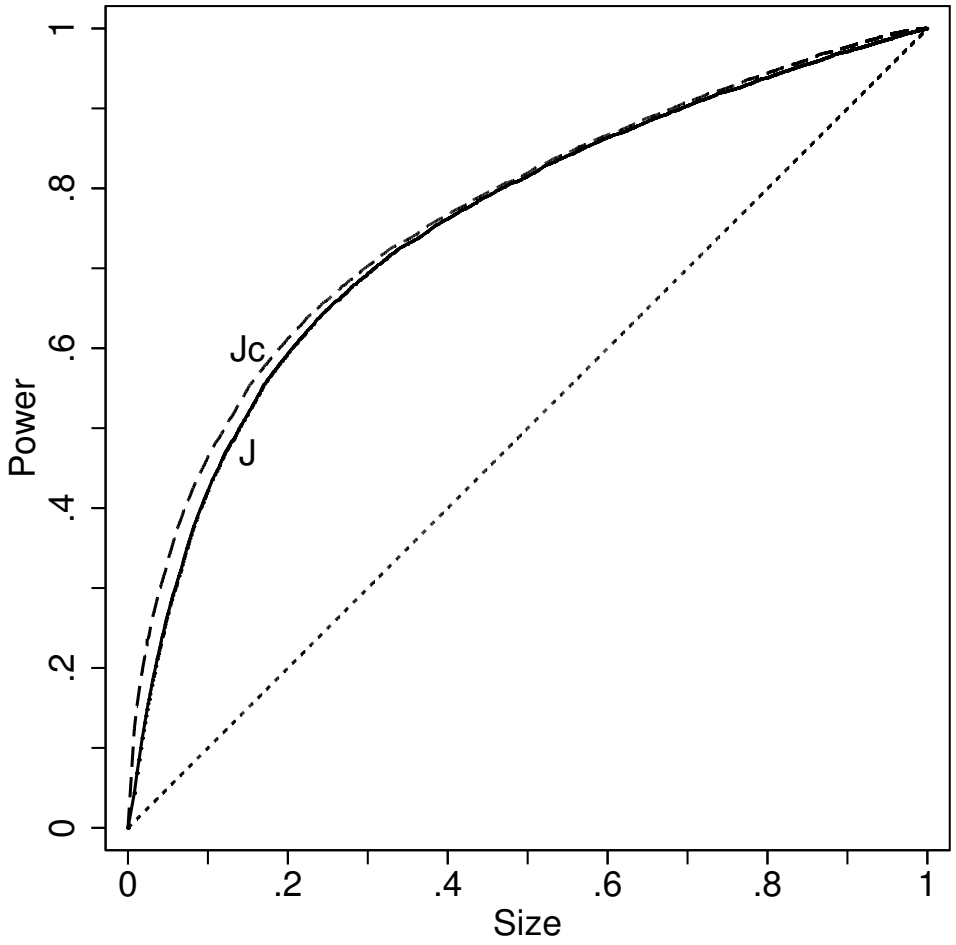
Given the performance of Hall's J-test once it is adjusted for size, one must search elsewhere for size-corrected power gains. One prospect is to apply the bootstrap to the centered J-test. The bootstrap is now a common approach to hypothesis testing that has been shown to reduce approximation error. For example, see Hall and Horowitz (1996). MacKinnon (2002) points out that although at a small number of bootstrap samples the bootstrap does lose some power, at a reasonably large number of bootstrap samples this is not problematic. In fact, the loss of power in bootstrapping the centered J-test would arise precisely because it corrects the tendency of the test to over-reject.

Another avenue is to use more robust estimators than GMM and/or more robust statistics. With respect to estimation, advances with generalized empirical likelihood (GEL) methods provide one alternative to using GMM. For example, see the statistics derived in Imbens et al. (1998). GEL has the advantage over GMM in that (i) the asymptotic bias of the parameter estimates does not increase in the number of overidentifying restrictions, and (ii) the moment conditions hold exactly in finite samples. This second point implies there is no need to re-center the moment conditions when estimating the long-run covariance matrix.

Finally, weak identification is a potential issue in the empirical example which is overlooked. Test-statistics based on GMM estimation in the presence of weak identification have non-standard limiting distributions. Several new tests have been developed using empirical likelihood (or one of its variants) estimation. These statistics are robust to weak identification and should be used under those conditions.

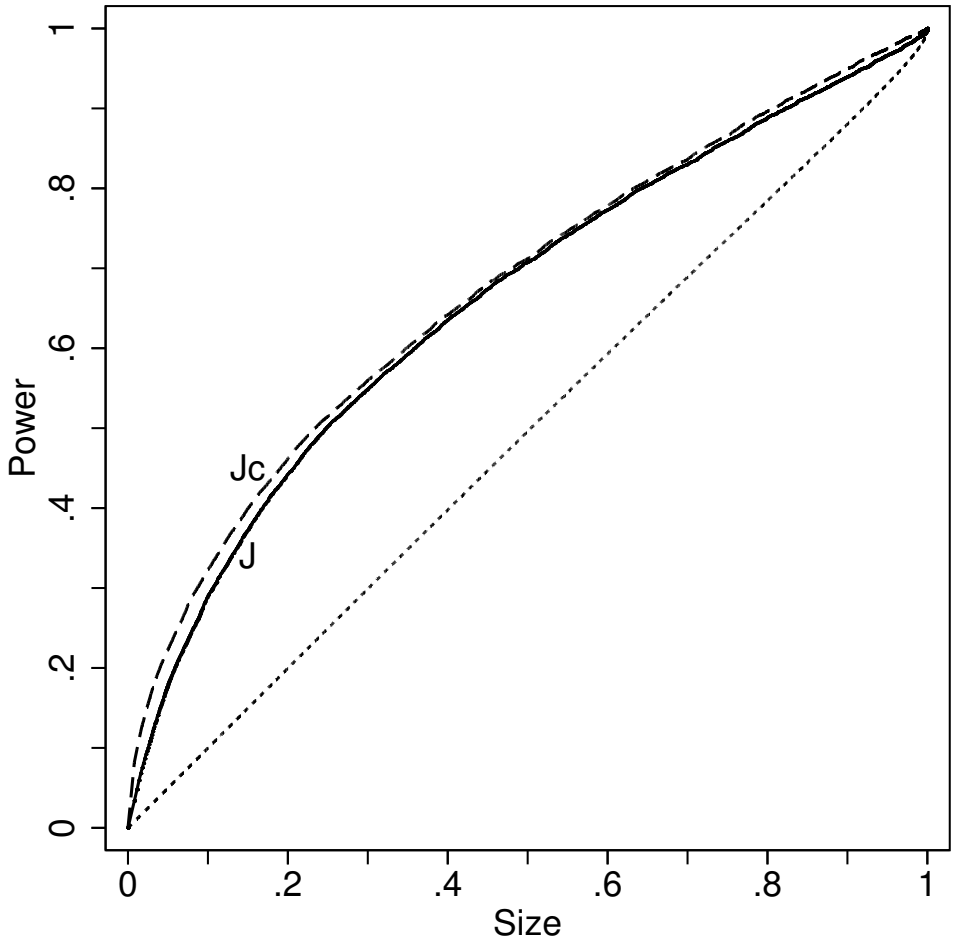
In a nonlinear framework examples include Stock and Wright (2000), Kleibergen (2005), and Otsu (2004). More research in these areas seem fruitful.

Figure 1. Size-Power Curves: Independent data



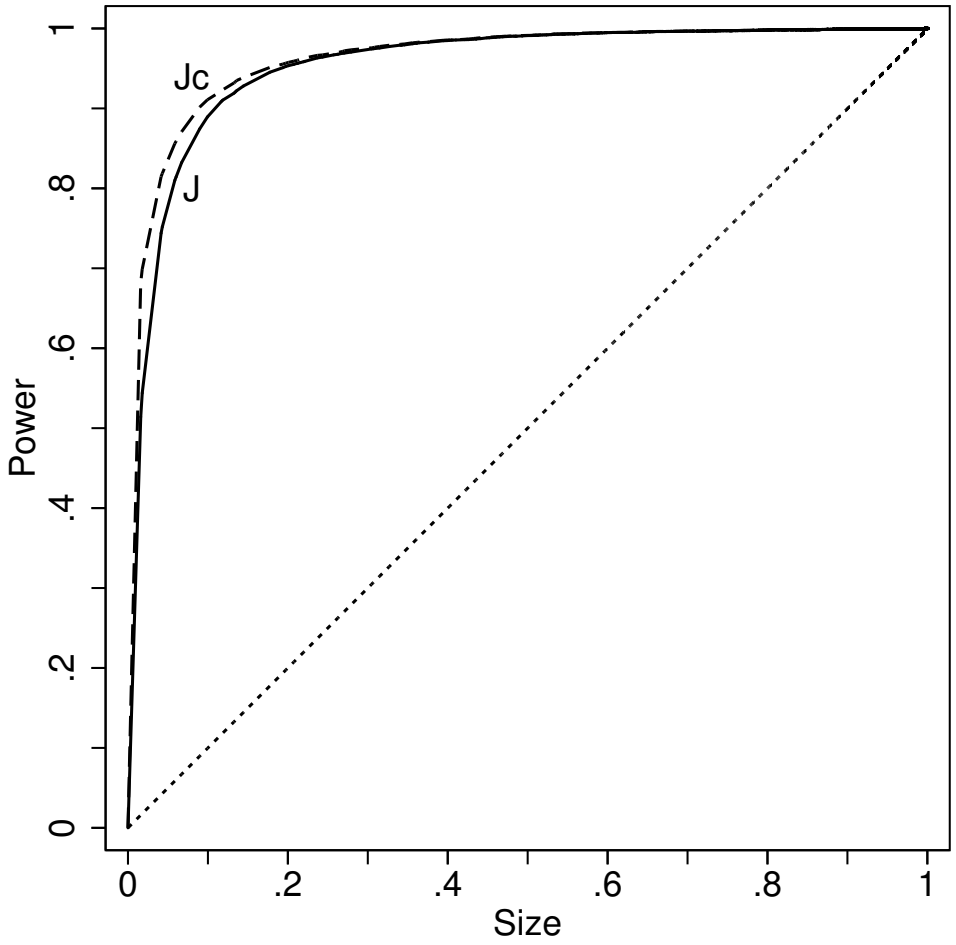
Note: $\gamma = 0.125$, $T = 300$, Replications= 10000. *Jc* is the centered test for overidentifying restrictions and is given by the long-dashed line. *J* is the standard test and given by the straight line. A test above the 45 degree line is one with power greater than size.

Figure 2. Size-Power Curves: Dependent Data



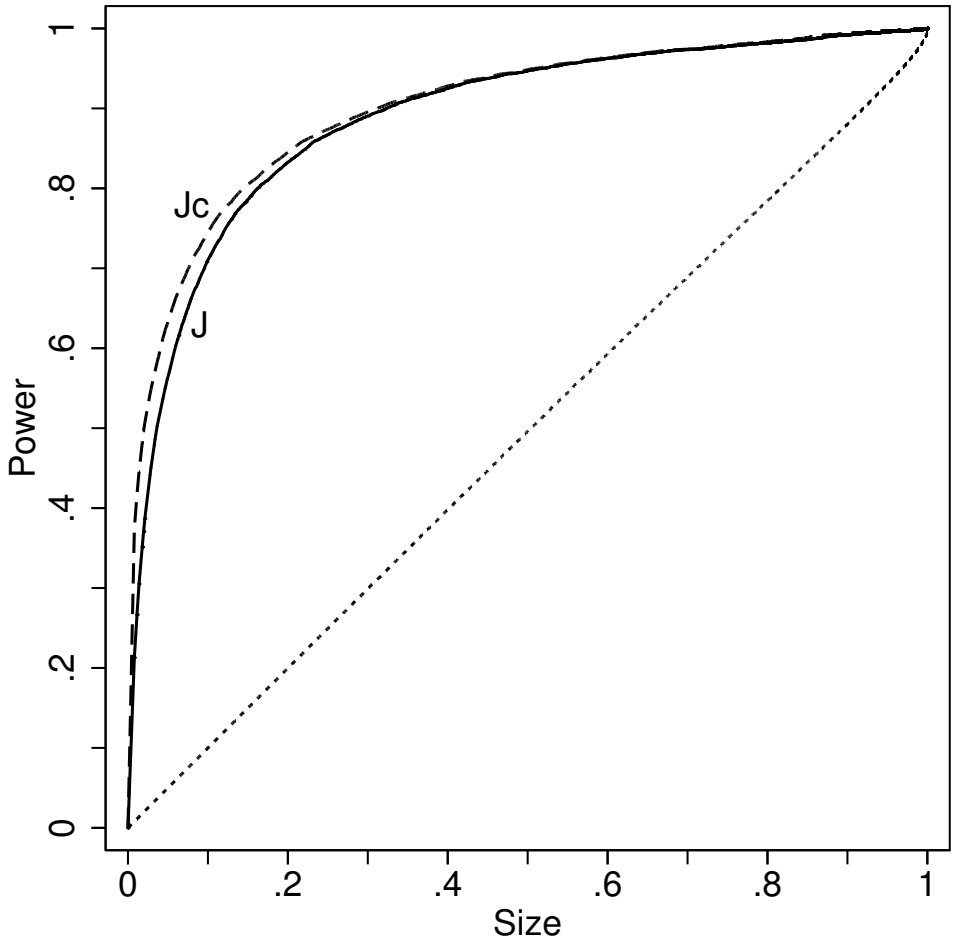
Note: $\gamma = 0.125$, $T = 300$, Replications= 10000. *Jc* is the centered test for overidentifying restrictions and is given by the long-dashed line. *J* is the standard test and given by the straight line. A test above the 45 degree line is one with power greater than size.

Figure 3. Size-Power Curves: Independent data



Note: $\gamma = 0.250$, $T = 300$, Replications= 10000. *Jc* is the centered test for overidentifying restrictions and is given by the long-dashed line. *J* is the standard test and given by the straight line. A test above the 45 degree line is one with power greater than size.

Figure 4. Size-Power Curves: Dependent Data



Note: $\gamma = 0.250$, $T = 300$, Replications= 10000. *Jc* is the centered test for overidentifying restrictions and is given by the long-dashed line. *J* is the standard test and given by the straight line. A test above the 45 degree line is one with power greater than size.

Table 4
Empirical Results

We estimate the following moment conditions for the period 1970 : 01 to 2002 : 12

$$E_t\left[\left(\beta\left(c_{t+1}/c_t\right)^{-\frac{1}{\phi}}\right)^{\theta} R_{m,t+1}^{\theta-1} R_{i,t+1}\right] - 1 = 0 \qquad i = 1,...N$$

		Standard		Centered	
Consumption		Z1	Z2	Z1	Z2
ND	β	0.9963	0.9992	0.9962	0.9990
		(0.0065)	(0.0059)	(0.0065)	(0.0056)
	ϕ	1.2502	0.3875	1.2504	0.4268
		(6.384)	(0.2633)	(6.341)	(0.3070)
	α	0.8021	1.1776	0.8032	1.1584
		(6.3844)	(0.2294)	(6.3406)	(0.2310)
	J	19.80	24.31		
		[0.0192]	[0.0284]		
	J_c			21.47	28.02
				[0.0107]	[0.009]
ND + SV	β	1.0000	0.9963	0.9997	0.9958
		(0.0623)	(0.0053)	(0.0581)	(0.0055)
	ϕ	0.2323	2.5204	0.2649	2.5171
		(2.651)	(18.90)	(3.188)	(19.71)
	α	0.3383	0.8962	0.3942	0.9018
		(8.496)	(0.5116)	(8.761)	(0.5060)
	J	20.79	20.07		
		[0.0136]	[0.0934]		
	J_c			24.19	27.18
				[0.0040]	[0.0118]

Note: The specific notation is the following: ND=nondurables; SV=services. $Z1 = \{\iota, c_t/c_{t-1}, R_{m,t-1}\}$, and $Z2 = \{\iota, c_t/c_{t-1}, R_{m,t-1}, Tbill_{t-1}\}$. Standard errors are in parentheses and p-values are in brackets

References

Altonji, J.G., and L.M. Segal (1996) 'Small-sample bias in GMM estimation of covariance structures.' *Journal of Business & Economic Statistics* 14, 353–366

Clark, T.E. (1996) ‘Small-sample properties of estimators of nonlinear models of covariance structure.’ *Journal of Business and Economic Statistics* 14, 367–373

Davidson, Russell, and James MacKinnon (1998) ‘Graphical methods for investigating the size and power of test statistics.’ *The Manchester School* 66, 1–26

Epstein, L.G., and S.E. Zin (1991) ‘Substitution, risk averson, and the temporal beahvior of consumption and asset returns: An empirical analysis.’ *Journal of Political Economy* 99, 263–286

Hall, A.R. (2000) ‘Covariance matrix estimation and the power of the overidentifying restrictions test.’ *Econometrica* 68, 1517–1527

Hall, P., and J.L. Horowitz (1996) ‘Bootstrap critical values for tests based on generalized-method-of-moments estimators.’ *Econometrica* 64, 891–916

Hansen, L.P. (1982) ‘Large sample properties of generalized method of moments estimators.’ *Econometrica* 50, 1029–1054

Imbens, G.W., R.H. Spady, and P. Johnson (1998) ‘Information theoretic approaches to inference in moment condition models.’ *Econometrica* 66, 333–357

Kleibergen, Frank (2005) ‘Testing parameters in gmm without assuming that they are identified.’ forthcoming Econometrica

MacKinnon, James (2002) ‘Bootstrap inference in econometrics.’ *Canadian Journal of Economics* 35, 615–645

Newey, W.K., and K.D. West (1994) ‘Automatic lag selection in covariance matrix estimation.’ *Review of Economic Studies* 61, 631–654

Otsu, Taisuke (2004) ‘Generalized empirical likelihood inference under weak identification.’ working paper

Smith, David (1999) ‘Finite-sample properties of tests of the Epstein-Zin asset pricing model.’ *Journal of Econometrics* 93, 113–148

Stock, James, and Jonathan Wright (2000) ‘Gmm with weak identification.’ *Econometrica* 68, 1055–1096