Simulation-Based Multiple Testing for Many Non-Nested Multivariate Models

Lynda Khalaf* and Florian Richard†

December 10, 2020

Abstract

We propose a multivariate extension of exact specification tests for non-nested models. Our test is finite-sample exact under the assumption of Gaussian errors, and is easily generalized to a multiple-model hypothesis via a combined alternative. We obtain valid inference results using bootstrapped Monte Carlo p-values, even when the distribution under the null hypothesis is intractable. We consider both Gaussian and non-Gaussian error structures through bootstrapping, and we show that our test possesses good size and power properties via simulations. Finally, we present empirical applications to asset pricing by testing benchmark factor models against single and multiple alternatives.

JEL classification: C12, C15, C30, C52, G12

Keywords: Multivariate linear regression, uniform mixed linear hypothesis, Monte Carlo methods, factor models.

*Email: lynda.khalaf@carleton.ca. Carleton University, 1125 Colonel By Drive, A-801 Loeb, Ottawa ON, Canada K1S 5B6

†Email: florian.richard@carleton.ca. Carleton University, 1125 Colonel By Drive, D-886 Loeb, Ottawa ON, Canada K1S 5B6.

1 Introduction

In this paper, we propose multivariate extensions of exact specification tests for non-nested models and an extension in the case of multiple non-nested alternatives. The test is exact in finite samples when the errors follow a Gaussian distribution. Our test yields valid results even when the design matrix does not have full column rank, and non-Gaussian models are also considered via bootstrap methods. Our empirically relevant simulations demonstrate that the test enjoys good size and power properties. Finally, in our empirical analysis, we apply our multivariate test to the problem of factor model specification in asset pricing.

Model specification is a crucial aspect of applied research. Accurate specification allows for correct inference about the distribution of the data generating process (DGP) under study. In the case of non-nested models, where neither model can be expressed as a restricted version of the other, univariate specification tests exists, but in the multivariate context, the properties of these tests have generally been overlooked, particularly when testing against many competing non-nested alternatives.[1] Likelihood ratio-type tests from the nested model specification testing literature are available, but they lead to unreliable inference.[2]

The seminal work by Cox (1961, 1962) pioneered model specification testing with a likelihood ratio-based test. Despite applications of this test to non-nested models by Pesaran (1974) and the generalisation to multivariate non-linear regression models by Pesaran and Deaton (1978), its computational complexity makes it unpopular among econometricians. A simpler approach is to use the results of Milliken and Graybill (1970), by augmenting the *a priori* model with some possibly non-linear function of the expected value of the dependent variable, via the principle of artificial nesting; see Hoel (1947). Prominent tests for non-nested models which use this idea include Davidson and MacKinnon (1981)'s J test and its multivariate counterpart discussed in Davidson and MacKinnon (1983).[3] However, the artificial term in the J test only depends on a projection of the competing regressors and on the dependent variable. Moreover, McAleer (1981) suggests that the J test often leans toward the model with the least amount of regressors. Considering that the J test overrejects in finite samples, Fisher and McAleer (1981) proposes a modified J test, the J_A test, whose artificial term is a consistent estimate of the expected value of the dependent variable when the alternative model generates the data, and corrects for the size distortion. This approach complies with the framework of Milliken and Graybill (1970), following Atkinson (1970). The J_A test, however, is thought to generally lack power compared to the J test. Stewart

[1] Pesaran and Weeks (2001) provides a broad review of non-nested model testing.

[2] For model selection of non-nested hypotheses using the Kullback-Leibler Information Criterion, see Vuong (1989).

[3] Variations of the J test are provided in Bernanke et al. (1988) and Hagemann (2012).

(1997) presents multivariate generalisations of univariate specification tests (including the J_A test) by applying Rao (1951)'s F-test to a set of Berndt and Savin (1977)'s uniform mixed linear (UML) restrictions. Yet, the small sample properties of these multivariate tests for non-nested alternatives are left unexplored. Our paper makes several contributions to this literature.

First, we provide an extension of the J_A test to a multivariate setting, which is finite-sample exact under normality, using a regularization approach. Our test statistic, computed using UML restrictions, is asymptotically valid under weak conditions and invariant to the parameterisation of the distribution's covariance matrix, under the normality of the error terms.[4] The use of a pseudoinverse with this portmanteau test allows us to circumvent the issue of design matrices that do not have full column rank, conceivably because of high correlation between dependent variables used as regressors. Second, we draw attention to the multivariate J test of Davidson and MacKinnon (1983), which requires regularization in finite samples. We revisit this test with two important modifications: (i) a bootstrap procedure to correct for size distortions, and (ii) a pseudoinverse to bypass regularization problems. All bootstrap methods are corrected for the fact that the Moore-Penrose inverse is not necessarily smooth. Our simulation study shows that our version of the multivariate J test does not produce size distortions, in spite of not being an exact test in small samples. Third, we generalize the multivariate J_A test to allow for a formal comparison of a single model against the union of multiple models without compounding type I errors. Motivated by this idea, our test addresses the multiple-model inference problem via a compound alternative hypothesis, suitable to big data applications. Fourth, we present an extensive and empirically relevant simulation study, and consider designs with Gaussian errors, t-distributed errors with various degrees of freedom, as well as a wild bootstrap with Rademacher errors. We present the empirical size and power of our tests, for small and large sample sizes and varying number of dependent variables. Finally, in our empirical section, the applications of our test address the growing problem of model specification in the asset pricing literature. Harvey et al. (2016) documents 316 factor models since 1964.[5] Therefore, the problem of model specification in asset pricing factor models arises naturally. We apply our methodology to several model specifications, including the Fama and French (2015) five-factor model, and variations of the Fama and French (1993) model, as in Pástor and Stambaugh (2003). In addition to the models mentioned above, we examine models incorporating consumption and housing risk factors. While factor selection procedures in asset pricing often make use of machine learning techniques, we employ an inferential approach to this problem.[6] In contrast with

[4]The proof of asymptotic validity will be provided in the next draft of this paper.

[5]Campbell Harvey and Yan Liu keep an updated list of factors at this link.

[6]See Feng et al. (2017), Kozak et al. (2019), Freyberger et al. (2017), Chen et al. (2019) and Gu et al.

this strand of the literature, our procedure can result in rejecting or accepting all models under study. An important motivation of our paper is that the majority of machine learning methods are not designed for non-nested models specification testing. We make use of Monte Carlo (MC) p-values, and test the Fama and French (2015) five-factor model against single and multiple competing models. We find that the Fama and French (2015) and the Pástor and Stambaugh (2003) models are misspecified for most time periods in our sample. Finally, when testing the Fama and French (2015) model against multiple models, particularly against consumption-based asset pricing models, we find that it is almost never rejected.

The paper is structured as follows. Section 2 develops the framework of multivariate exact tests. Section 3 details the simulation study, for tests against both single and multiple alternatives. Section 4 presents empirical results. Section 5 concludes the paper.

2 Econometric Framework

Consider the collection of competing multivariate regression models $Y = f_k(X_k; \theta_k) + U_k$, where $Y \subset \mathbb{R}^{T\times n}$ is a vector of T observations for n endogenous variables, f_k are (possibly non-linear) functions, and X_k are $(T \times K_k)$ matrices of exogenous variables, respectively, for $k = 0, \ldots, K$. This ordering is motivated by the fact that $k = 0$ will refer to the model the null hypothesis H_0. θ_k are vectors of unknown parameters associated with compact parameter spaces $\Theta_k \subset \mathbb{R}^{K_k\times n}$, and U_k are matrices of errors. Denote a given model $Y = f_k(X_k; \theta_k)+U_k$ as $\mathcal{M}_k$, for some k. We make the following assumptions about the regressors and the endogenous variables:

Assumption 1. *The regressors X_k are non-stochastic for all k.*

Assumption 2. *The regressors X_k have full-column rank, so that $rk(X_k) = K_k$ for all k.*

Assumption 3. *U_k follows an absolutely continuous distribution conditional on X_k, for all k.*

Assumption 4. *$T > K_0 + n$.*

Assumption 1 is necessary for finite-sample validity but will later be relaxed to derive our asymptotic results. An additional assumption about the distribution of endogenous variables allows us to state the following Lemma. The proof is stated in Appendix A.1.

Lemma 2.1. *Under assumption 3, Y has full-column rank with probability 1 conditional on X_k.*

(2020) for applications of machine learning algorithms to factor selection.

Without loss of generality, let $\mathcal{M}_0$ be the null model and $\mathcal{M}_k$ be the alternative models, for $k = 1, \ldots, K$. We desire to test the following hypotheses:

$$H_0: \quad \mathcal{M}_0, \quad \text{vs.} \tag{2.1}$$

$$H_1: \quad \bigcup_{k=1}^{K} \mathcal{M}_k. \tag{2.2}$$

In the context of multivariate linear regression models, (2.1) and (2.2) can be written as:

$$H_0: \quad Y = X_0\theta_0 + U_0, \qquad U_0 \overset{i.i.d.}{\sim} (\mathbf{0}_{n\times 1}, \mathbf{\Sigma_0}), \text{ vs.} \tag{2.3}$$

$$H_1: \quad \text{the union of models of the form } Y = X_k\theta_k + U_k, \qquad U_k \overset{i.i.d.}{\sim} (\mathbf{0}_{n\times 1}, \mathbf{\Sigma_{\boldsymbol{k}}}), \tag{2.4}$$

where $\mathbf{\Sigma_0}$ and $\mathbf{\Sigma_{\boldsymbol{k}}}$ denote the scale matrices of the distribution of the error terms, which corresponds to the covariance matrices in the Gaussian and Student-t distributions. Consider the case of *non-nested* regressors as in Vuong (1989): the regressors can be strictly non-nested ($\bigcap_{k=0}^{K} X_k = \varnothing$) or overlapping ($X_0 \cap X_k \neq \varnothing$ with $X_0 \not\subset X_k$ and $X_k \subset X_0$ for at least one $k \in \{1, \ldots, K\}$). In that sense, X_k contains at least a unique regressor for each k. Consequently, no model can be expressed as a restricted version of the others. To circumvent this problem, the principle of artificial nesting can be applied to construct a comprehensive regression

$$Y = X_0\theta_0(I_n - \sum_{k=1}^{K} A_k) + \sum_{k=1}^{K} X_k\theta_k A_k + \bar{U}, \tag{2.5}$$

where $\bar{U} = (I_n - \sum_{k=1}^{K} A_k)U_0 + \sum_{k=1}^{K} A_k U_k$ and A_k are $(n \times n)$ matrices. One can interpret this comprehensive regression a weighted average of equations (2.3) and (2.4). Under the null hypothesis, $H_0 : A_1 = \ldots = A_K = \mathbf{0}_{n\times n}$, and under the alternative hypothesis, H_1 : at least one $A_k \neq \mathbf{0}_{n\times n}$. This regression is non-linear in parameters through the term $X_0\theta_0(I_n - \sum_{k=1}^{K} A_k)$. An alternative specification which is linear and that can be easily estimated via ordinary least squares (OLS) is

$$Y = X_0\theta_0 + \sum_{k=1}^{K} X_k\theta_k A_k + U. \tag{2.6}$$

However, as pointed out by Davidson and MacKinnon (1981) and Fisher and McAleer (1981) in the univariate case and single model case, it is impossible to identify $\theta_0, \ldots, \theta_K$ as well as $A_0, \ldots, A_K$. Indeed, $\sum_{k=0}^{K} K_k n + Kn^2$ parameters need to be identified, but one can only identify up to $\sum_{k=0}^{K} K_k n$ parameters if the regressors are strictly non-nested, and even fewer parameters if they have regressors in common. The regressors being uncorrelated with U when H_0 is true, the unknown quantities in (2.6) can be identified with consistent estimates of the θ_k.[7] The OLS estimator of θ_k is given by $\hat{\theta}_k = X_k^+ Y$, where X_k^+ is the Moore-Penrose inverse of X_k. This substitution yields

$$Y = X_0\theta_0 + \sum_{k=1}^{K} X_k \hat{\theta}_k A_k + U, \tag{2.7}$$

which is equivalent to

$$Y = X_0\theta_0 + \sum_{k=1}^{K} P_{X_k} Y A_k + U, \tag{2.8}$$

where $P_{X_k} = X_k X_k^+$ is the orthogonal projection of X_k. This is the multivariate counterpart of the artificial regression associated with Davidson and MacKinnon (1981)'s J test using the Berndt and Savin (1977) framework. Under normality, $P_{X_0}Y$ will be independent from U and (weakly) exogenous otherwise, following Davidson and MacKinnon (1981) and Stewart (1997). Taking in account the parameter restriction $A_1 = \ldots = A_K = \mathbf{0}_{n\times n}$ for the null model, it is desirable to estimate the expected value of Y under the null hypothesis as well. Such an estimate is given by the artificial term $P_{X_k}P_{X_0}Y$, where $P_{X_0} = X_0 X_0^+$ is the orthogonal projection of X_0. Then, the corresponding comprehensive model is consistent with the procedure that Milliken and Graybill (1970) developed for univariate models, in that the artificial term is some function of $X_0\hat{\theta}_0$:

$$Y = X_0\theta_0 + \sum_{k=1}^{K} P_{X_k} P_{X_0} Y A_k + U, \tag{2.9}$$

or, if we define $\widetilde{X} = [X_0 \quad P_{X_1}P_{X_0}Y \ \ldots \ P_{X_K}P_{X_0}Y]$ and $\Pi = [\theta_0 \quad A_1 \ \ldots \ A_K]'$:

$$Y = \widetilde{X}\Pi + U, \tag{2.10}$$

where $\widetilde{X}$ and Π are $(T\times(K_0+nK))$ and $((K_0+nK)\times n)$, respectively. This is the multivariate counterpart of Fisher and McAleer (1981)'s J_A test given by Stewart (1997), again using the Berndt and Savin (1977) restrictions. We assume the errors U have the following form:

Assumption 5. $U = WJ'$

[7] See Atkinson (1970).

where J is invertible, so $\boldsymbol{\Sigma} = JJ'$. W has a known distribution so that the joint distribution of the rows of U is known up the unknown matrix J. In the multivariate Gaussian case, each row of W follows an i.i.d. multivariate Gaussian distribution:

Assumption 6. $W_t \sim N(\boldsymbol{0}_{n\times 1}, I_n), \quad t = 1, \ldots, T.$

We test the hypothesis:

$$H_0 : A_1 = \ldots = A_K = \mathbf{0}_{n\times n}. \tag{2.11}$$

In the context of specification testing, to fully understand the meaning of a rejection of the null hypothesis, and conversely, a failure to reject, it becomes evident that one should also test the reverse of these hypotheses, as suggested by MacKinnon (1983). If $A_1 = \ldots = A_K = \mathbf{0}_{n\times n}$, we fail to reject H_0 and the null model is the true model; however, a rejection of H_0 does not necessarily imply that the alternative model is the true model. Only after testing the reverse of this hypothesis (with the null model in the alternative hypothesis and vice versa) can one draw conclusions about H_0 and H_1. In the case that $K = 1$, there are 4 possible outcomes: both the null and the alternative models are misspecified; the null model is misspecified and the alternative model is correctly specified; the null model is correctly specified and the alternative model is misspecified; and both models are correctly specified. The extension of this argument to multiple alternatives to raises a multiple comparison problem addressed in Richard (2020). The hypothesis (2.11) can be tested via a set of UML restrictions:

$$R\Pi C = L, \tag{2.12}$$

where $R = [\mathbf{0}_{n\times K_0} \quad I_n \, \ldots \, I_n]$, which is of dimension $(n \times (K_0 + nK))$, $C = I_n$, $L = \mathbf{0}_{n\times n}$. We focus on the widely-used Wilks (1938)'s lambda criterion (see Dufour and Khalaf (2002) and references therein):

$$\Lambda = |\widehat{U}'\widehat{U}|/|\widehat{U}_0'\widehat{U}_0|, \tag{2.13}$$

where $\widehat{U}'\widehat{U}$ and $\widehat{U}_0'\widehat{U}_0$ are the unconstrained and constrained sum of squared errors (SSE), respectively. In particular, Wilk's lambda can be expressed as the product of the eigenvalues λ_i:

$$\Lambda = \prod_{i=1}^{n} \lambda_i, \tag{2.14}$$

which in turn, are the roots of the determinantal equation

$$|\widehat{U}'\widehat{U} - \lambda \widehat{U}_0'\widehat{U}_0| = 0. \tag{2.15}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)$, with $\lambda_1 \geq \ldots \geq \lambda_n$. Furthermore, we can define the likelihood ratio (LR) criterion as

$$LR = -T \ln(\Lambda). \tag{2.16}$$

Dufour and Khalaf (2002) show that, when $\widetilde{X}$ has full-column rank, Wilk's lambda is a pivotal quantity. We generalize this result to the present case where $\widetilde{X}$ is rank-deficient, using the Moore-Penrose inverse. For the unconstrained model, the SSE is given by

$$\widehat{U}'\widehat{U} = U'M(\widetilde{X})U = JW'M(\widetilde{X})WJ', \tag{2.17}$$

where $M(\widetilde{X}) = I_T - P_{\widetilde{X}}$ is the residual maker matrix, and $P_{\widetilde{X}} = \widetilde{X}\widetilde{X}^+$. For the constrained model, the SSE is

$$\widehat{U}_0'\widehat{U}_0 = U'M_0U = JW'M_0WJ', \tag{2.18}$$

where $M_0 = M(\widetilde{X}) + \widetilde{X}(\widetilde{X}'\widetilde{X})^+R'\left[R(\widetilde{X}'\widetilde{X})^+R'\right]^{-1}R\widetilde{X}^+$. We can condition on $\widetilde{X}$ because $\widetilde{X}$ is independent of U under normality. Outside of the Gaussian distribution, the asymptotic case satisfies the regularity conditions of Andrews (1987). We can now state generalizations of Theorem 3.1, Corollary 3.2 and 3.3 of Dufour and Khalaf (2002) using the Moore-Penrose inverse.

Theorem 2.2. *Under assumptions 1 to 6 and hypothesis (2.11), the eigenvalues* $\lambda_1, \ldots, \lambda_n$ *are distributed, conditional on* $\widetilde{X}$*, like the roots of*

$$|W'M(\widetilde{X})W - \lambda W'M_0W| = 0. \tag{2.19}$$

Proof. Equation (2.15) can be written as

$$|JW'M(\widetilde{X})WJ' - \lambda JW'M_0WJ| = 0 \tag{2.20}$$
$$|J||W'M(\widetilde{X})W - \lambda W'M_0W||J'| = 0 \tag{2.21}$$
$$|W'M(\widetilde{X})W - \lambda W'M_0W| = 0 \tag{2.22}$$

which does not depend on J. □

Corollary 2.2.1. *Under assumptions 1 to 6 and hypothesis (2.11),* Λ *is distributed like the product of the roots of* $|W'M(\widetilde{X})W - \lambda W'M_0W| = 0$.

Corollary 2.2.2. *Under assumptions 1 to 6 and hypothesis (2.11),* Λ *follows the same distribution as the ratio*

$$|W'M(\widetilde{X})W|/W'M_0W|. \tag{2.23}$$

When the problem does not require regularization, i.e., when $\widetilde{X}$ and the partition $[P_{X_1}P_{X_0}Y \ \dots \ P_{X_K}P_{X_0}Y]$ have full column rank, another available statistic to test (2.11) under UML restrictions is the monotonic transformation of Λ due to Rao (1951):

$$F = \frac{1-\Lambda^{1/\tau}}{\Lambda^{1/\tau}}\frac{\rho\tau - 2\lambda}{pq} \tag{2.24}$$

where $\rho = T - K_0 - (p-q+1)/2$, $\lambda = (pq-2)/4$, $\tau = \sqrt{(p^2q^2-4)/(p^2+q^2-5)}$ when $p^2+q^2-5>0$ and 1 otherwise. q is the number of restrictions per equation. For non-integer degrees of freedom, the degrees of freedom are rounded to the nearest integer.[8] In our case, $p = q = n$. Then, $\rho = T - K_0 - 1/2$, $\lambda = (n^2-2)/4$, and $\tau = \sqrt{(n^4-4)/(2n^2-5)}$. For the special case where $\min\{p,q\} \leq 2$, Rao's F statistic will follow an F-distribution exactly. In the univariate case with exogenous regressors, Fisher and McAleer (1981) and Godfrey (1983) have pointed out the J_A test is exact under normality of the errors. Outside of normality, $\widetilde{X}$ and U will not be independent and the statistic will not be pivotal in finite samples. We have not made any assumptions with regards to the inversion of $\widetilde{X}'\widetilde{X}$. Indeed, there is no guarantee that the design matrix $\widetilde{X}$ has full column rank, and in many cases $\widetilde{X}$ is rank deficient. Under Assumption 4, the full-rank condition is $rk(\widetilde{X}) = \min(T, K_X+n) = K_X+n$, which requires that $\min(rk(P_Z), rk(P_X), rk(Y)) - dim\,\mathscr{C}(X)\cap\mathscr{C}(P_ZP_XY) = n$. This is equivalent to $n < \min(K_Z, K_X)$ and $dim\,\mathscr{C}(X)\cap\mathscr{C}(P_ZP_XY) = 0$. Consequently, the OLS estimator $\widehat{\Pi} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'Y$ may be difficult to compute using traditional methods. Generalized inverses can be used for the purpose of regularization in linear models, as in Milliken and Graybill (1970). For the same reason that has motivated the procedure of Milliken and Graybill (1970), we can condition on $\widetilde{X}$, and the statistic will be pivotal under the normality assumption, following the argument of Dufour and Khalaf (2002), using (2.17) and (2.18). It is straightforward to show that the Wilk statistic is invariant to J, hence conditioning on $\widetilde{X}$ provides pivotality, even when $\widehat{\Pi}$ is obtained via a pseudoinverse.

Davidson and MacKinnon (1983) develops a multivariate J test and presents the linear case using generalized least squares (GLS). Let $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{Z}}$ the variance-covariance matrices

[8] See Rao (1973), p. 556.

of models (2.3) and (2.4). Define $\widehat{Q}_X = \text{chol}(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{X}})$ and $\widehat{Q}_Z = \text{chol}(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{Z}})$ as the Cholesky decompositions of the variance-covariance matrices, and let $\hat{f} = X\widehat{\theta}$, $\hat{g} = Z\widehat{\gamma}$, and $h = \widehat{\boldsymbol{\Omega}}_{\boldsymbol{X}}(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{Z}})^{-1}(\hat{g}-\hat{f})'$. Then, define $y = \text{vec}((\widehat{Q}_X(Y-\hat{f})')')$ as the vectorization of the dependent variable matrix. Denoting $\widetilde{\mathcal{W}} = [I_n \otimes (\widehat{Q}_X\hat{f}')' \quad \text{vec}((\widehat{Q}_X h)')]$, we can write the regression succinctly as

$$y = \widetilde{\mathcal{W}}B + E, \tag{2.25}$$

where $B = [b^v \quad \lambda]'$, and E denotes a stochastic error term. We desire to test the hypothesis $\lambda = 0$ using a t-statistic. The GLS estimate of B is simply

$$\widehat{B} = (\widetilde{\mathcal{W}}'\widetilde{\mathcal{W}})^{-1}\widetilde{\mathcal{W}}'y, \tag{2.26}$$

and the residuals for equation (2.25) are

$$\widehat{E} = y - \widetilde{\mathcal{W}}\widehat{B}. \tag{2.27}$$

The standard errors are then given by

$$S = ((\widehat{E}'\widehat{E})(\widetilde{\mathcal{W}}'\widetilde{\mathcal{W}})^{-1})/(nT). \tag{2.28}$$

Let S_e denote the bottom right elements of the $(n^2+1)\times(n^2+1)$ standard error matrix S. The desired t-statistic is given by the ratio of the estimated $\widehat{\lambda}$ and the square root of S_e

$$t = \frac{\widehat{\lambda}}{\sqrt{S_e}}. \tag{2.29}$$

We find through our simulation study that the matrix is almost singular in small samples ($T = 60$). Thus, we also apply our regularization technique to the multivariate J test. In our case, we use the Moore-Penrose inverse to compute the estimates of Π and Π_c for the J_A test, and the estimate of B for the J test. In the context of the linear regression (2.10), the Moore-Penrose inverse of $\widetilde{X}'\widetilde{X}$ is denoted $(\widetilde{X}'\widetilde{X})^+$. In the case that $\widetilde{X}'\widetilde{X}$ is invertible, the Moore-Penrose inverse is simply the matrix inverse, so $(\widetilde{X}'\widetilde{X})^+ = (\widetilde{X}'\widetilde{X})^{-1}$. In addition to the Moore-Penrose inverse, we use a Monte Carlo (MC) p-value method to obtain a tractable simulated distribution under the null hypothesis, which is detailed in the following subsection.

2.1 Bootstrap Procedure

To obtain reliable inference results when applying the Moore-Penrose inverse, we perform a parametric bootstrap which corresponds to a MC test with a consistent estimate of the nuisance parameters, as in Dufour (2006). In particular, this method yields an exact test, when the null distribution of the test statistic is shown to be pivotal, which occurs in our case under normality for the multivariate J_A test. The MC p-value procedure is as follows. Let S_0 denote the F statistic computed from the sample, called the *observed statistic*, and let $S_1, \ldots, S_B$ denote B exchangeable replications of the observed statistic, called the *simulated statistics*.[9] The simulated statistics can be obtained via bootstrapping or simulations, so that the null distributions of $S_0, S_1, \ldots, S_B$ are identical. We use a parametric bootstrap with the same parameterisation as S_0 and a wild bootstrap. The following steps outline the parametric bootstrap procedure:

1. Estimate the model parameters $\widehat{\theta}_0$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}$ under H_0.
2. Compute S_0 under H_0, using (2.24).
3. Draw U^b from multivariate normal or multivariate t distributions with covariance $\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}$.
4. Compute $Y_b = X_0\widehat{\theta}_0 + U^b$.
5. Compute the simulated statistic S_b using Y_b as in (2.24), conditioning on X_k, for $k = 1, \ldots, K$.
6. Repeat steps 3 through 5 B times to obtain the simulated statistics $S_1, \ldots, S_B$.

We consider errors drawn from a multivariate Gaussian distribution $N(0_{n\times 1}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{0}})$, as well as multivariate Student-t errors with 5 degrees of freedom. We generate the Student-t errors as follows:

$$\boldsymbol{t} = \frac{\mathcal{Z}_1}{(\mathcal{Z}_2/\kappa)^{1/2}}, \tag{2.30}$$

where $\mathcal{Z}_1$ follows a multivariate Gaussian distribution parameterised like above, and $\mathcal{Z}_2$ follows a chi-squared distribution with κ degrees of freedom. In the Gaussian case, pivotality implies that the test will be exact. The asymptotic validity of this method beyond the Gaussian case depends on the fact that the statistic is asymptotically pivotal; see Dufour (2006) and Davidson and MacKinnon (2002) and references therein. We also apply the

[9] A sequence of random variables is *exchangeable* if all permutations of that sequence have the same joint distribution as the original sequence; see de Finetti's representation theorem.

approach of Davidson and Flachaire (2008) to perform a multivariate wild bootstrap. The wild bootstrap allows the errors to exhibit heteroskedasticity. The wild bootstrap procedure is as follows:

1. Estimate the model parameters $\widehat{\theta}_0$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{0}}$ under H_0.
2. Compute S_0 under H_0, using (2.24).
3. Draw a $T \times 1$ vector of random variables ε such that $E(\varepsilon) = 0$ and $E(\varepsilon^2) = 1$.
4. For each row $\widehat{U}_t$ of $\widehat{U}$, compute $h(\widehat{U}_t)$, where $h(.)$ is some transformation of $\widehat{U}_t$.
5. For $t = 1, \ldots, T$, the bootstrap disturbances are $U_t^* = h(\widehat{U}_t) \circ \varepsilon_t$, where $\circ$ denotes the Hadamard product.
6. Compute $Y_b^* = X\widehat{\theta} + U^*$.
7. Compute the simulated statistic S_b using Y_b^* as in (2.24), conditioning on X and Z.
8. Repeat steps 3 through 7 B times to obtain the simulated statistics $S_1, \ldots, S_B$.

For ε_t, we draw from a Rademacher distribution, where ε_t takes on values 1 and -1 with probability 0.5, and we set $h(\widehat{U}_t) = \widehat{U}_t$.

Following the framework of Dufour (2006), under the assumption that $P(S_0 = S_b) \neq 0$ for all $b = 1, \ldots, B$, we draw $B + 1$ random variables $\widetilde{Z}_0, \widetilde{Z}_1, \ldots, \widetilde{Z}_B$ that follow a uniform distribution, independent from $S_0, S_1, \ldots, S_B$. The statistics are not automatically continuous, as the Moore-Penrose inverse is not always smooth. A smooth regularization is not a necessity as the MC p-value procedure allows for discrete statistics. $(S_0, S_1, \ldots, S_B)$ and $(\widetilde{Z}_0, \widetilde{Z}_1, \ldots, \widetilde{Z}_B)$ are then organised in pairs according to a lexicographic order:

$$(S_b, \widetilde{Z}_b) \geq (S_c, \widetilde{Z}_c) \Leftrightarrow [S_b > S_c \quad \text{or} \quad (S_b = S_c \quad \text{and} \quad \widetilde{Z}_b \geq \widetilde{Z}_c)], \tag{2.31}$$

for all $b, c = 1, \ldots, B$. The uniform random variables $(\widetilde{Z}_0, \widetilde{Z}_1, \ldots, \widetilde{Z}_B)$ serve the purpose of breaking the tie when $S_b = S_c$. This yields the MC p-value:

$$\widetilde{p}_B(x) = \frac{B\widetilde{G}_B(x) + 1}{B + 1}, \qquad \text{where} \tag{2.32}$$

$$\widetilde{G}_B(x) = 1 - \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{[0,\infty)}(x - S_b) + \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{[0]}(S_b - x)\mathbf{1}_{[0,\infty)}(\widetilde{Z}_b - \widetilde{Z}_0). \tag{2.33}$$

As long as $\alpha(B+1)$ is an integer, the test will have level α:

$$P\left[\widetilde{p}_B(S_0) \leq \alpha\right] = \frac{I\left[\alpha(B+1)\right]}{B+1}, \quad \text{for} \quad 0 \leq \alpha \leq 1, \tag{2.34}$$

where $I\left[\cdot\right]$ denotes the integer part.

3 Simulation Study

We present the results of an empirically-relevant simulation study. Each model is calibrated using the estimated parameters of asset pricing models that have received significant support in the literature. The test depends on the value of the regressors X_k, as well as on the multidimensional distribution parameters Σ_k. For non-nested tests, size and power depend on the distance between models, the regressors X_k, and multidimensional parameters, θ_k and Σ_k. We have chosen to vary X_k as a function of observations. As it is not obvious to track the effects of these changes on power, we change the sample size T and the number of test portfolios n, which are described below. To remain empirically relevant, we simulate the models based on the regressors and observed values of the considered models. Power will change with the Kullback-Leibler distance, and variations in our set of empirically relevant choices provide a substitute for the usual power curves.[10] We compute the Kullback-Leibler divergence as

$$D_{KL}(P_{\theta_k}, P_{\theta_0}) = \frac{T}{2}\log\left(\widehat{\Sigma}_0/\widehat{\Sigma}_k\right) - \frac{n}{2} + \frac{1}{2}tr\left(\widehat{\Sigma}_0^{-1}\widehat{\Sigma}_k\right) + \frac{1}{2}tr\left(\widehat{\Sigma}_0^{-1}\left(X_k\widehat{\theta}_k - X_0\widehat{\theta}_0\right)'\left(X_k\widehat{\theta}_k - X_0\widehat{\theta}_0\right)\right),$$

where P_{θ_k} and P_{θ_0} denote the distributions under the alternative and the null models, respectively.

We repeat the MC p-value procedure for our size and power calculations, using a parametric bootstrap and a wild bootstrap. For the parametric bootstrap, we consider both Gaussian errors and Student-t errors with 5 degrees of freedom, generated as in (2.30). For the wild bootstrap, we consider a Rademacher distribution. We consider both the full sample period, as well as 5 and 10 year subsamples. We perform 10,000 simulations for each MC p-value procedure, and we set B equal to 999 replications. We use a nominal significance level of $\alpha = 0.05$.

[10] For a treatment of the interpretation of the Kullback-Leibler divergence in a Neyman-Pearson framework, see Eguchi and Copas (2006).

3.1 Data Description

We use the research portfolio monthly returns available on Professor French's website as dependent variables, R_{it}.[11] The return series fare value-weighted monthly portfolio returns of U.S. stocks on the New York Stock Exchange (NYSE), the American Stock Exchange (AMEX), and the NASDAQ Stock Market. Portfolios are rebalanced each June and are sorted by characteristics. We perform our analysis for the market equity (size) portfolios ($n = 18$), the size and book value portfolios ($n = 25$), and the industry portfolios ($n = 5$ and $n = 12$), as suggested by Lewellen et al. (2010). The portfolios formed on size are sorted according to the firm's market equity value, and the portfolios formed on size and book value-to-market are the intersection of portfolios sorted according to the firm's market equity value and book value-to-market equity ratio. For both the size portfolios, we use the lowest 30%, the middle 40%, and the top 40% portfolios returns, along with the quintile and the decile portfolios returns. For the size and book portfolios, we use the NYSE quintiles. The industry portfolios are sorted according to the industry that the issuing firm falls under, for $n = 5$ (consumer goods, manufacturing, business equipment, healthcare, and others) and $n = 12$ (consumer nondurables, consumer durables, manufacturing, energy, chemicals, business equipment, telecommunications, utilities, wholesale and retail, healthcare, financial services, and others). The portfolio return data spans a period from January 1968 to August 2018.

For the independent variables, we use risk factors from the asset pricing literature. Factor models attempt to explain variation in asset returns from variation in economic variables, called risk factors. Such economic variables includes returns of (possibly non-traded) factors, macroeconomic quantities (Chen et al. (1986) and Shanken and Weinstein (2006)), and even behavioral factors (Daniel et al. (2019)). Factors can be created from non-tradable assets via a two-pass regression procedure (see Fama and MacBeth (1973)). The Fama and French (2015) model is represented by the following regression:

$$R_{it} = \alpha_i + \beta_{1i}(r_{mt} - r_{ft}) + \beta_{2i}SMB_t + \beta_{3i}HML_t + \beta_{4i}RMW_t + \beta_{5i}CMA_t + e_{it} \tag{FF5}$$

where $R_{it} = r_{it} - r_{ft}$ denotes the excess return of the test portfolio i over the risk-free return r_{ft} in period t, and $r_{mt} - r_{ft}$ denotes the market risk premium, i.e., the excess of the market portfolio r_{mt} over the risk-free rate. The FF5 risk factors are the SMB (Small Minus Big) factor, the HML (High Minus Low) factor, the RMW (Robust Minus Weak) factor, and the

[11] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Research

CMA (Conservative Minus Aggressive) factor. They are computed as follows:

$$
\begin{aligned}
SMB &= (\text{Small Value} + \text{Small Neutral} + \text{Small Growth})/3 \\
&\quad - (\text{Big Value} + \text{Big Neutral} + \text{Big Growth})/3 \\
HML &= (\text{Small Value} + \text{Big Value})/2 - (\text{Small Growth} + \text{Big Growth})/2 \\
RMW &= (\text{Small Robust} + \text{Big Robust})/2 - (\text{Small Weak} + \text{Big Weak})/2 \\
CMA &= (\text{Small Conservative} + \text{Big Conservative})/2 - (\text{Small Aggressive} + \text{Big Aggressive})/2
\end{aligned}
$$

where "Small" and "Big" denote the stocks of firms with small and large market capitalization, "Value" and "Growth" denote stocks with high book value-to-price (B/P) ratio and stocks with low B/P ratio, "Robust" and "Weak" denote the stocks of firms with high and low profitability, and "Conservative" and "Aggressive" denotes the stocks of firms that invest conservatively and aggressively. We test the Pástor and Stambaugh (2003), Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005) models against the FF5 model, either individually or jointly:

$$R_{it} = \alpha_i + \beta_{1i}(r_{mt} - r_{ft}) + \beta_{2i}SMB_t + \beta_{3i}HML_t + \beta_{4i}\mathcal{L}_t + \epsilon_{it} \tag{PS}$$

$$R_{it} = \alpha_i + \beta_{1i}cay_t + \beta_{2i}\Delta c_t + \beta_{3i}cay_t \times \Delta c_t + \epsilon_{it} \tag{LL}$$

$$R_{it} = \alpha_i + \beta_{1i}my_t + \beta_{2i}\Delta c_t + \beta_{3i}my_t \times \Delta c_t + \epsilon_{it} \tag{LvN}$$

$\mathcal{L}$ is the Pástor and Stambaugh (2003) liquidity factor. For a given day d in month t, the liquidity factor is formed by regressing the next day's market excess return $r^e_{i,d+1,t}$ on the signed volume and the portfolio return for that given day $r_{i,d,t}$: $r^e_{i,d+1,t} = \theta_{i,t} + \phi_{i,t}r_{i,d,t} + \gamma_{i,t}\text{sign}(r^e_{i,d,t}) \cdot v_{i,d,t} + \epsilon_{i,d+1,t}$. The estimated coefficient on the signed volume is expected to be negative when liquidity is low. This represents the effect of returns reversals induced by trading volumes at the security level. The first difference of the estimated coefficient for each period is then scaled by an estimate of liquidity cost (m_t/m_1) and averaged over the number of stocks N_t: $\Delta\hat{\gamma}_t = (m_t/m_1)\sum_{i=1}^{N_t}(\hat{\gamma}_{i,t} - \hat{\gamma}_{i,t-1})/N_t$. The resulting measure is then regressed on its own lag and the liquidity cost: $\Delta\hat{\gamma}_t = a + b\Delta\hat{\gamma}_{t-1} + c(m_{t-1}/m_t)\hat{\gamma}_t + u_t$. The liquidity factor is the residual from this regression, divided by 100: $\mathcal{L} = \hat{u}_t/100$. cay is the consumption-to-wealth ratio from the Lettau and Ludvigson (2001) consumption capital asset pricing model (C-CAPM), and Δc is the log consumption growth. cay is estimated as $\widehat{cay}_t = c_t - \hat{\beta}_a a_t - \hat{\beta}_y y_t$, where c is consumption, a is asset wealth, and y is labor income. my is the housing collateral ratio from Lustig and Van Nieuwerburgh (2005). It is computed as $my_t = \log(hv_t) + \hat{\omega}\log(y_t) + \hat{v}t + \hat{\chi}$, hv_t is the per household real estate wealth, y_t is the labor income plus transfers, $\hat{v}t$ accounts for a time trend, and $\hat{\chi}$ is a constant.

The factor returns are monthly for the Fama and French (2015), and Pástor and Stambaugh (2003) models, from January 1968 to August 2018 ($T = 608$), and quarterly for the Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005) models, from Q1 1968 to Q1 2005 ($T = 149$). When testing the Fama and French (2015) model against multiple alternatives, we compute the quarterly returns by compounding monthly returns.

3.2 Simulation Design

Experiments *I*, *II*, and *III* explore testing against the alternative hypothesis of a single model, while experiments *IV* and *V* consider the multiple testing aspect of our method. Experiment *VI* offers a finite-sample comparison between the multivariate J test from Davidson and MacKinnon (1983) (DM1983) using equation (2.29), the multivariate extension of the univariate J test with Berndt and Savin (1977) restrictions (DMBS), using equation (2.8), and the multivariate J_A test (JABS), using equation (2.9).

- *Experiment I*: Fama and French (2015) vs. Pástor and Stambaugh (2003).

  Hypothesis: $H_0 : A = \mathbf{0}_{n\times n}$.

  DGP for empirical size: Fama and French (2015) model.

  DGP for empirical power: Pástor and Stambaugh (2003) model.

  Test: JABS.

- *Experiment II:* Fama and French (2015) vs. Lettau and Ludvigson (2001).

  Hypothesis: $H_0 : A = \mathbf{0}_{n\times n}$.

  DGP for empirical size: Fama and French (2015) model.

  DGP for empirical power: Lettau and Ludvigson (2001) model.

  Test: JABS.

- *Experiment III:* Fama and French (2015) vs. Lustig and Van Nieuwerburgh (2005).

  Hypothesis: $H_0 : A = \mathbf{0}_{n\times n}$.

  DGP for empirical size: Fama and French (2015) model.

  DGP for empirical power: Lustig and Van Nieuwerburgh (2005) model.

  Test: JABS.

- *Experiment IV:* Fama and French (2015) vs. Pástor and Stambaugh (2003), Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005).

Hypothesis: $H_0 : A_1 = A_2 = A_3 = \mathbf{0}_{n\times n}$.

DGP for empirical size: Fama and French (2015) model.

DGP for empirical power: We generate the data from the estimation of each of the 3 alternative models.

Test: JABS.

- *Experiment V:* Fama and French (2015) vs. Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005).

Hypothesis: $H_0 : A_1 = A_2 = \mathbf{0}_{n\times n}$.

DGP for empirical size: Fama and French (2015) model.

DGP for empirical power: We generate the data from the estimation of each of the 2 alternative models.

Test: JABS.

- *Experiment VI:* Fama and French (2015) vs. Pástor and Stambaugh (2003).

Hypothesis: $H_0 : A = \mathbf{0}_{n\times n}$.

DGP for empirical size: Fama and French (2015) model.

DGP for empirical power: Pástor and Stambaugh (2003) model.

Test: DM1983, DMBS, JABS.

3.3 Discussion of Simulation Results

The results of experiment I show that the multivariate J_A test exhibits appropriate size for the Gaussian case. Additionally, using a thick-tailed distribution such at a Student-t distribution with 5 degrees of freedom, does not create size distortions. A larger number of test portfolios does not to affect empirical size, as the test remains correctly sized as n increases. With a wild bootstrap, size is still close to the nominal level, but the test slightly overrejects in the 68-78 period. Power is generally close to 1 and increases with the number of portfolios, n. Power significantly lowers during the 1988 to 1998 period, which is explained by the fact that the distance between the alternative and the null models is small, as shown in Table 2. The period from 1998Q1 to 2005Q1 is not shown in subsequent tables, as it violates Assumption 4 that $T > K_0 + n$. Table 3 shows the simulation results of testing the Fama and French (2015) model against the Lettau and Ludvigson (2001) C-CAPM. Even with a small sample size ($T = 29$) because of the quarterly frequency, the test remains properly sized with Gaussian and $t(5)$ errors, but overrejects when using a wild bootstrap, providing support for the former distributional assumptions. For the 5-industry portfolios, periods of lower power (1978Q1 - 1987Q4) again stem from the proximity between the distributions ($D_{KL} = 22.49$).

Tables 7 through 16 show that even when testing the null hypothesis against the compound alternative of multiple models, size is still controlled for Gaussian and $t(5)$ errors, and overrejects with the Wild bootstrap, providing further support for the assumption of Gaussian and $t(5)$ errors. Power is satisfactory for the Gaussian and $t(5)$ cases, and increases with n. When performing a wild bootstrap, however, power is generally lower than under Gaussian and $t(5)$ distributions.

Additionally, Table 17 presents empirical size and power for the DM1983, the DMBS, and the JABS tests, for a design where sample size is small ($T = 60$) and the number of dependent variables is large ($n = 25$). While the J test is not exact in small samples, we do not observe size distortions, and the deviation from 5% is negligible. This finding is generally due to the use of bootstrap methods.[12] In fact, the test appears to reject the null hypothesis closer to the nominal significance level when n increases. All 3 tests are adequately sized in small samples. The DM1983 test dominates marginally the DMBS and JABS, whose power is identical.

[12]The use of bootstrap methods with the J test is discussed at length in Davidson and MacKinnon (2002).

Table 1: *Experiment I:* Fama and French (2015) vs. Pástor and Stambaugh (2003). Empirical size and power for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Empirical size				Empirical power			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
	Gaussian errors							
68 - 18	0.0464	0.0475	0.0481	0.0457	0.9945	0.9998	0.9999	0.9916
68 - 78	0.0541	0.0536	0.0525	0.0527	1.0000	1.0000	0.9935	0.9957
78 - 88	0.0510	0.0505	0.0487	0.0493	0.8770	0.9554	0.9838	0.9460
88 - 98	0.0505	0.0512	0.0496	0.0501	0.3823	0.6418	0.6938	0.8240
98 - 08	0.0529	0.0567	0.0490	0.0548	0.9762	0.9832	0.9935	0.9921
08 - 18	0.0506	0.0515	0.0515	0.0507	0.9995	0.9960	0.9930	0.9903
	$t(5)$ errors							
68 - 18	0.0567	0.0553	0.0515	0.0536	0.9864	0.9910	0.9845	0.8918
68 - 78	0.0534	0.0538	0.0517	0.0526	0.9974	1.0000	0.9207	0.9513
78 - 88	0.0495	0.0502	0.0492	0.0465	0.6754	0.8055	0.8878	0.8128
88 - 98	0.0486	0.0487	0.0524	0.0479	0.2538	0.4412	0.4941	0.6267
98 - 08	0.0468	0.0468	0.0458	0.0490	0.8424	0.8760	0.9224	0.9328
08 - 18	0.0486	0.0497	0.0548	0.0507	0.9760	0.9417	0.9310	0.9176
	Wild bootstrap							
68 - 18	0.0531	0.0554	0.0529	0.0535	0.9884	0.9969	0.9899	0.9309
68 - 78	0.0645	0.0644	0.0671	0.0632	1.0000	1.0000	0.9406	0.9799
78 - 88	0.0497	0.0524	0.0467	0.0518	0.7553	0.8880	0.9546	0.8598
88 - 98	0.0526	0.0515	0.0487	0.0544	0.3579	0.5689	0.6015	0.7027
98 - 08	0.0578	0.0556	0.0534	0.0528	0.9330	0.9303	0.9751	0.9456
08 - 18	0.0550	0.0612	0.0594	0.0667	0.9984	0.9806	0.9401	0.8335

Table 2: *Experiment I:* Kullback-Leibler distance between the distribution of errors of the Pástor and Stambaugh (2003) and Fama and French (2015) models, assuming a Gaussian distribution, and the pseudo Kullback-Leibler distance for a Student-$t(5)$ distribution obtained via simulation, and using a Wild bootstrap with a Rademacher distribution obtained via simulation.

	Industry		Size	Size & Book
	$n = 5$	$n = 12$	$n = 18$	$n = 25$
T	Gaussian distribution			
68 - 18	123.27	131.55	67.16	114.95
68 - 78	44.62	70.33	50.73	59.79
78 - 88	19.01	38.38	52.45	39.90
88 - 98	9.12	33.41	25.67	35.67
98 - 08	25.83	39.67	51.26	65.18
08 - 18	44.72	65.87	54.52	65.42
	$t(5)$ distribution			
68 - 18	20.56	23.74	18.76	8.41
68 - 78	30.52	41.60	9.00	6.22
78 - 88	5.18	5.59	5.73	3.19
88 - 98	-1.47	-3.61	-6.68	-5.94
98 - 08	9.62	8.51	9.09	11.00
08 - 18	18.59	13.43	10.27	9.53
	Wild bootstrap			
68 - 18	24.48	26.44	14.55	3.46
68 - 78	44.49	53.13	12.17	4.13
78 - 88	9.60	10.91	8.43	5.22
88 - 98	-2.83	-5.65	-8.80	-5.17
98 - 08	14.23	9.28	11.11	10.33
08 - 18	27.38	18.94	15.03	11.89

Table 3: *Experiment II:* Fama and French (2015) vs. Lettau and Ludvigson (2001). Empirical size and power for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Empirical size				Empirical power			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
	Gaussian errors							
1968Q1 - 2005Q1	0.0512	0.0492	0.0483	0.0488	0.8694	0.9953	0.9927	0.9991
1968Q1 - 1977Q4	0.0529	0.0513	0.0493	0.0490	0.7042	0.9704	0.9991	0.9952
1978Q1 - 1987Q4	0.0531	0.0510	0.0555	0.0541	0.3966	0.8108	0.9949	1.0000
1988Q1 - 1997Q4	0.0499	0.0512	0.0516	0.0537	0.6436	0.9197	0.9993	0.9996
	$t(5)$ errors							
1968Q1 - 2005Q1	0.0520	0.0451	0.0469	0.0464	0.6030	0.8908	0.8893	0.9621
1968Q1 - 1977Q4	0.0525	0.0500	0.0483	0.0507	0.4803	0.8625	0.9811	0.9776
1978Q1 - 1987Q4	0.0510	0.0498	0.0515	0.0513	0.2297	0.5647	0.9492	0.9971
1988Q1 - 1997Q4	0.0493	0.0502	0.0526	0.0504	0.3920	0.7173	0.9797	0.9893
	Wild bootstrap							
1968Q1 - 2005Q1	0.0561	0.0556	0.0613	0.0566	0.8056	0.9842	0.9670	0.9869
1968Q1 - 1977Q4	0.0614	0.0597	0.0760	0.0737	0.6984	0.8717	0.9399	0.7353
1978Q1 - 1987Q4	0.0519	0.0465	0.0512	0.0448	0.2268	0.4973	0.6921	0.8033
1988Q1 - 1997Q4	0.0521	0.1109	0.0877	0.0898	0.5326	0.7833	0.9586	0.8573

Table 4: *Experiment II:* Kullback-Leibler distance between the distribution of errors of the Lettau and Ludvigson (2001) and Fama and French (2015) models, assuming a Gaussian distribution, and the pseudo Kullback-Leibler distance for a Student-$t(5)$ distribution obtained via simulation, and using a Wild bootstrap with a Rademacher distribution obtained via simulation.

	Industry		Size	Size & Book
	$n = 5$	$n = 12$	$n = 18$	$n = 25$
T	Gaussian distribution			
1968Q1 - 2005Q1	42.34	88.08	142.54	199.84
1968Q1 - 1977Q4	31.77	83.38	154.40	239.39
1978Q1 - 1987Q4	22.49	71.77	143.91	310.15
1988Q1 - 1997Q4	23.36	85.64	130.40	287.27
	$t(5)$ distribution			
1968Q1 - 2005Q1	3.71	10.09	3.66	7.26
1968Q1 - 1977Q4	2.05	10.49	27.75	26.94
1978Q1 - 1987Q4	-3.31	-1.68	21.90	62.31
1988Q1 - 1997Q4	0.98	5.40	29.89	42.45
	Wild bootstrap			
1968Q1 - 2005Q1	2.75	-11.04	-28.43	-9.68
1968Q1 - 1977Q4	-2.40	-20.70	-66.07	-185.09
1978Q1 - 1987Q4	-2.62	14.62	-79.36	-75.40
1988Q1 - 1997Q4	12.62	-21.41	29.24	-34.26

Table 5: *Experiment III:* Fama and French (2015) vs. Lustig and Van Nieuwerburgh (2005). Empirical size and power for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Empirical size				Empirical power			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
	Gaussian errors							
1968Q1 - 2005Q1	0.0492	0.0480	0.0503	0.0488	0.7809	0.9774	0.9968	0.9977
1968Q1 - 1977Q4	0.0527	0.0517	0.0521	0.0482	0.8706	0.9961	1.0000	0.9997
1978Q1 - 1987Q4	0.0491	0.0521	0.0506	0.0507	0.5879	0.8745	0.9962	0.9741
1988Q1 - 1997Q4	0.0500	0.0491	0.0531	0.0525	0.4979	0.8998	0.9859	0.9930
	$t(5)$ errors							
1968Q1 - 2005Q1	0.0522	0.0495	0.0498	0.0510	0.4973	0.7996	0.9220	0.9268
1968Q1 - 1977Q4	0.0529	0.0523	0.0526	0.0525	0.6280	0.9397	0.9990	0.9935
1978Q1 - 1987Q4	0.0496	0.0473	0.0456	0.0474	0.3685	0.6773	0.9607	0.9239
1988Q1 - 1997Q4	0.0505	0.0542	0.0536	0.0535	0.3036	0.7061	0.9115	0.9613
	Wild bootstrap							
1968Q1 - 2005Q1	0.0603	0.0600	0.0556	0.0543	0.7097	0.9399	0.9751	0.9807
1968Q1 - 1977Q4	0.0562	0.0559	0.0711	0.0718	0.8035	0.9617	0.9632	0.6646
1978Q1 - 1987Q4	0.0389	0.0469	0.0572	0.0462	0.5297	0.6735	0.7587	0.4584
1988Q1 - 1997Q4	0.0599	0.0892	0.1102	0.0998	0.4677	0.7599	0.8146	0.7250

Table 6: *Experiment III:* Kullback-Leibler distance between the distribution of errors of the Lustig and Van Nieuwerburgh (2005) and Fama and French (2015) models, assuming a Gaussian distribution, and the pseudo Kullback-Leibler distance for a Student-$t(5)$ distribution obtained via simulation, and using a Wild bootstrap with a Rademacher distribution obtained via simulation.

	Industry		Size	Size & Book
	$n = 5$	$n = 12$	$n = 18$	$n = 25$
T	Gaussian distribution			
1968Q1 - 2005Q1	41.68	85.26	140.47	197.50
1968Q1 - 1977Q4	32.62	90.99	179.45	255.90
1978Q1 - 1987Q4	25.88	75.41	148.95	236.24
1988Q1 - 1997Q4	21.65	81.13	107.42	271.27
	$t(5)$ distribution			
1968Q1 - 2005Q1	0.90	4.24	7.33	-1.01
1968Q1 - 1977Q4	6.80	19.84	50.56	45.62
1978Q1 - 1987Q4	0.25	2.32	22.81	11.68
1988Q1 - 1997Q4	-1.90	2.58	12.54	21.97
	Wild bootstrap			
1968Q1 - 2005Q1	-9.36	-1.02	-39.98	-54.24
1968Q1 - 1977Q4	6.03	-7.32	-26.98	-54.82
1978Q1 - 1987Q4	0.41	-5.09	-57.76	-63.48
1988Q1 - 1997Q4	8.84	-4.54	11.55	-35.26

Table 7: *Experiment IV:* Fama and French (2015) vs. Pástor and Stambaugh (2003), Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005). Empirical size for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.0494	0.0461	0.0483	0.0481	0.0494	0.0466	0.0467	0.0460
1968Q1 - 1977Q4	0.0524	0.0512	0.0487	0.0499	0.0526	0.0495	0.0530	0.0530
1978Q1 - 1987Q4	0.0489	0.0540	0.0516	0.0515	0.0522	0.0475	0.0484	0.0491
1988Q1 - 1997Q4	0.0485	0.0515	0.0520	0.0522	0.0526	0.0560	0.0538	0.0539
	Wild bootstrap							
1968Q1 - 2005Q1	0.0554	0.0552	0.0594	0.0575				
1968Q1 - 1977Q4	0.0607	0.0696	0.0761	0.0728				
1978Q1 - 1987Q4	0.0550	0.0532	0.0607	0.0560				
1988Q1 - 1997Q4	0.0581	0.1091	0.1082	0.0847				

Table 8: *Experiment IV:* Fama and French (2015) vs. Pástor and Stambaugh (2003), Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005). Empirical power (DGP: Pástor and Stambaugh (2003)) for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.7911	0.9822	0.9925	0.9992	0.4966	0.8233	0.8823	0.9658
1968Q1 - 1977Q4	0.5697	0.9439	0.9943	0.9699	0.3686	0.7931	0.9531	0.9279
1978Q1 - 1987Q4	0.4125	0.7967	0.9901	0.9888	0.2444	0.5838	0.9358	0.9611
1988Q1 - 1997Q4	0.7605	0.9345	0.9962	0.9414	0.4984	0.7626	0.9622	0.8743
	Wild bootstrap							
1968Q1 - 2005Q1	0.7253	0.9574	0.9714	0.9876				
1968Q1 - 1977Q4	0.5158	0.7662	0.8580	0.5668				
1978Q1 - 1987Q4	0.2825	0.5271	0.7656	0.5755				
1988Q1 - 1997Q4	0.7227	0.8416	0.9432	0.5656				

Table 9: *Experiment IV:* Fama and French (2015) vs. Pástor and Stambaugh (2003), Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005). Empirical power (DGP: Lettau and Ludvigson (2001)) for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.7262	0.9529	0.9406	0.9831	0.4402	0.7224	0.7075	0.8348
1968Q1 - 1977Q4	0.4872	0.8073	0.9242	0.7590	0.3095	0.6075	0.8101	0.6789
1978Q1 - 1987Q4	0.2516	0.5541	0.8925	0.9199	0.1651	0.3705	0.7431	0.8468
1988Q1 - 1997Q4	0.4305	0.6964	0.9341	0.8464	0.2571	0.4758	0.8082	0.7458
	Wild bootstrap							
1968Q1 - 2005Q1	0.6698	0.9048	0.8747	0.9262				
1968Q1 - 1977Q4	0.4450	0.6234	0.6502	0.4003				
1978Q1 - 1987Q4	0.1556	0.3442	0.4521	0.4010				
1988Q1 - 1997Q4	0.4127	0.6204	0.8002	0.4562				

Table 10: *Experiment IV:* Fama and French (2015) vs. Pástor and Stambaugh (2003), Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005). Empirical power (DGP: Lustig and Van Nieuwerburgh (2005)) for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.6082	0.8923	0.9670	0.9622	0.3584	0.6068	0.7699	0.7546
1968Q1 - 1977Q4	0.6775	0.9225	0.9900	0.8632	0.4474	0.7559	0.9412	0.7801
1978Q1 - 1987Q4	0.3899	0.6363	0.8896	0.6550	0.2314	0.4374	0.7347	0.5439
1988Q1 - 1997Q4	0.3266	0.6644	0.8148	0.7342	0.2003	0.4406	0.6441	0.6280
	Wild bootstrap							
1968Q1 - 2005Q1	0.5677	0.8087	0.8765	0.8673				
1968Q1 - 1977Q4	0.5672	0.7374	0.7871	0.3764				
1978Q1 - 1987Q4	0.3287	0.4406	0.4875	0.2331				
1988Q1 - 1997Q4	0.2870	0.5880	0.5900	0.3665				

Table 11: *Experiment IV:* Kullback-Leibler distance between the distribution of errors of the Pástor and Stambaugh (2003) and Fama and French (2015) models, assuming a Gaussian distribution, and the pseudo Kullback-Leibler distance for a Student-$t(5)$ distribution obtained via simulation, and using a Wild bootstrap with a Rademacher distribution obtained via simulation.

	Industry		Size	Size & Book
	$n = 5$	$n = 12$	$n = 18$	$n = 25$
T	Gaussian distribution			
1968Q1 - 2005Q1	8.29	24.60	46.94	78.78
1968Q1 - 1977Q4	13.73	43.50	79.77	178.67
1978Q1 - 1987Q4	10.25	46.93	63.78	211.34
1988Q1 - 1997Q4	27.33	56.99	94.71	359.38
	$t(5)$ distribution			
1968Q1 - 2005Q1	-0.60	1.40	11.03	21.70
1968Q1 - 1977Q4	4.13	18.10	25.37	39.88
1978Q1 - 1987Q4	0.53	18.25	19.46	49.40
1988Q1 - 1997Q4	18.65	26.81	58.82	58.61
	Wild bootstrap			
1968Q1 - 2005Q1	-2.28	17.14	-12.47	28.68
1968Q1 - 1977Q4	0.51	31.92	-26.62	-62.58
1978Q1 - 1987Q4	10.61	43.29	-40.60	26.14
1988Q1 - 1997Q4	29.97	21.69	-1.81	37.86

Table 12: *Experiment IV:* Kullback-Leibler distance between the distribution of errors of the Lettau and Ludvigson (2001) and Fama and French (2015) models, assuming a Gaussian distribution, and the pseudo Kullback-Leibler distance for a Student-$t(5)$ distribution obtained via simulation, and using a Wild bootstrap with a Rademacher distribution obtained via simulation.

	Industry		Size	Size & Book
	$n = 5$	$n = 12$	$n = 18$	$n = 25$
T	Gaussian distribution			
1968Q1 - 2005Q1	23.77	56.37	72.23	111.90
1968Q1 - 1977Q4	21.54	67.18	141.50	299.77
1978Q1 - 1987Q4	10.21	43.02	129.36	323.68
1988Q1 - 1997Q4	16.25	56.49	101.20	477.46
	$t(5)$ distribution			
1968Q1 - 2005Q1	3.71	10.09	3.66	7.26
1968Q1 - 1977Q4	2.05	10.49	27.75	26.94
1978Q1 - 1987Q4	-3.31	-1.68	21.90	62.31
1988Q1 - 1997Q4	0.98	5.40	29.89	42.45
	Wild bootstrap			
1968Q1 - 2005Q1	9.84	-11.04	-28.43	-9.68
1968Q1 - 1977Q4	-2.40	-20.70	-66.07	-185.09
1978Q1 - 1987Q4	-2.62	14.62	-79.36	-75.40
1988Q1 - 1997Q4	12.62	-21.41	29.24	-34.26

Table 13: *Experiment IV:* Kullback-Leibler distance between the distribution of errors of the Lustig and Van Nieuwerburgh (2005) and Fama and French (2015) models, assuming a Gaussian distribution, and the pseudo Kullback-Leibler distance for a Student-$t(5)$ distribution obtained via simulation, and using a Wild bootstrap with a Rademacher distribution obtained via simulation.

	Industry		Size	Size & Book
	$n = 5$	$n = 12$	$n = 18$	$n = 25$
T	Gaussian distribution			
1968Q1 - 2005Q1	18.97	45.61	75.09	101.48
1968Q1 - 1977Q4	27.74	82.29	160.38	299.85
1978Q1 - 1987Q4	17.28	53.60	123.11	293.85
1988Q1 - 1997Q4	12.68	52.22	86.45	449.41
	$t(5)$ distribution			
1968Q1 - 2005Q1	0.91	4.24	7.33	-1.01
1968Q1 - 1977Q4	6.80	19.84	50.56	45.62
1978Q1 - 1987Q4	0.25	2.32	22.81	11.68
1988Q1 - 1997Q4	-1.90	2.58	12.54	21.97
	Wild bootstrap			
1968Q1 - 2005Q1	-9.36	-1.02	-39.98	-54.23
1968Q1 - 1977Q4	6.03	-7.32	-26.98	-54.82
1978Q1 - 1987Q4	0.41	-5.09	-57.76	-63.48
1988Q1 - 1997Q4	8.84	-4.54	11.55	-35.26

Table 14: *Experiment V:* Fama and French (2015) vs. Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005). Empirical size for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.0483	0.0471	0.0473	0.0468	0.0518	0.0444	0.0462	0.0488
1968Q1 - 1977Q4	0.0525	0.0506	0.0472	0.0458	0.0530	0.0481	0.0505	0.0509
1978Q1 - 1987Q4	0.0492	0.0530	0.0534	0.0553	0.0511	0.0473	0.0472	0.0491
1988Q1 - 1997Q4	0.0489	0.0486	0.0504	0.0534	0.0520	0.0539	0.0534	0.0500
	Wild bootstrap							
1968Q1 - 2005Q1	0.0573	0.0530	0.0559	0.0528				
1968Q1 - 1977Q4	0.0537	0.0608	0.0695	0.0689				
1978Q1 - 1987Q4	0.0512	0.0518	0.0550	0.0548				
1988Q1 - 1997Q4	0.0596	0.1155	0.1158	0.0914				

Table 15: *Experiment V:* Fama and French (2015) vs. Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005). Empirical power (DGP: Lettau and Ludvigson (2001)) for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n=5$	$n=12$	$n=18$	$n=25$	$n=5$	$n=12$	$n=18$	$n=25$
1968Q1 - 2005Q1	0.7734	0.9725	0.9630	0.9924	0.4806	0.7688	0.7606	0.8765
1968Q1 - 1977Q4	0.5467	0.8734	0.9668	0.8814	0.3464	0.6838	0.8761	0.8085
1978Q1 - 1987Q4	0.2884	0.6190	0.9435	0.9781	0.1761	0.4138	0.8137	0.9358
1988Q1 - 1997Q4	0.4925	0.7721	0.9718	0.9407	0.2857	0.5312	0.8700	0.8641
	Wild bootstrap							
1968Q1 - 2005Q1	0.7239	0.9397	0.9103	0.9504				
1968Q1 - 1977Q4	0.4920	0.6704	0.7276	0.4603				
1978Q1 - 1987Q4	0.1674	0.3788	0.5125	0.5015				
1988Q1 - 1997Q4	0.4535	0.6802	0.8662	0.5912				

Table 16: *Experiment V:* Fama and French (2015) vs. Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005). Empirical power (DGP: Lustig and Van Nieuwerburgh (2005)) for the parametric bootstrap with Gaussian and $t(5)$ errors, and the wild bootstrap with a Rademacher distribution.

	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n=5$	$n=12$	$n=18$	$n=25$	$n=5$	$n=12$	$n=18$	$n=25$
1968Q1 - 2005Q1	0.6606	0.9247	0.9807	0.9794	0.3928	0.6565	0.8221	0.8048
1968Q1 - 1977Q4	0.7407	0.9601	0.9966	0.9493	0.5003	0.8237	0.9744	0.8925
1978Q1 - 1987Q4	0.4445	0.7128	0.9433	0.7883	0.2591	0.4837	0.8071	0.6724
1988Q1 - 1997Q4	0.3663	0.7429	0.8899	0.8611	0.2192	0.4934	0.7178	0.7512
	Wild bootstrap							
1968Q1 - 2005Q1	0.6161	0.8610	0.9066	0.9055				
1968Q1 - 1977Q4	0.6416	0.8160	0.8534	0.4303				
1978Q1 - 1987Q4	0.3500	0.4895	0.5433	0.2728				
1988Q1 - 1997Q4	0.3479	0.6678	0.6677	0.4795				

Table 17: *Experiment VI:* Fama and French (2015) vs. Pástor and Stambaugh (2003). Empirical size and power for the parametric bootstrap with Gaussian errors, $n = 5$.

	Empirical Size			Empirical Power		
T	DM1983	DMBS	JABS	DM1983	DMBS	JABS
68 - 73	0.0584	0.0518	0.0518	0.7763	0.7019	0.7019
73 - 78	0.0523	0.0471	0.0471	1.0000	1.0000	1.0000
78 - 83	0.0508	0.0461	0.0461	0.4274	0.3384	0.3384
83 - 88	0.0605	0.0540	0.0540	0.9144	0.8611	0.8611
88 - 93	0.0513	0.0509	0.0509	0.7744	0.7270	0.7270
93 - 98	0.0634	0.0500	0.0500	0.9560	0.9433	0.9433
98 - 03	0.0658	0.0511	0.0511	0.7933	0.7519	0.7519
03 - 08	0.0488	0.0530	0.0530	1.0000	1.0000	1.0000
08 - 13	0.0617	0.0516	0.0516	0.9991	0.9977	0.9977
13 - 18	0.0533	0.0496	0.0496	0.5954	0.4877	0.4877

4 Empirical Results

We present empirical applications of the multivariate J_A test. Each table shows the MC p-values for various test portfolios, computed using (2.32), for an observed statistic computed using (2.24) with the LR criterion. We set the number of bootstrap replications B equal to 999. We test the following single hypotheses:

- **Single Hypotheses:**

  Hypothesis I.a: H_0: Fama and French (2015) vs. H_1: Pástor and Stambaugh (2003).

  Hypothesis I.b: H_0: Pástor and Stambaugh (2003) vs. H_1: Fama and French (2015).

  Hypothesis II.a: H_0: Fama and French (2015) vs. H_1: Lettau and Ludvigson (2001).

  Hypothesis II.b: H_0: Lettau and Ludvigson (2001) vs. H_1: Fama and French (2015).

  Hypothesis III.a: H_0: Fama and French (2015) vs. H_1: Lustig and Van Nieuwerburgh (2005).

  Hypothesis III.b: H_0: Lustig and Van Nieuwerburgh (2005) vs. H_1: Fama and French (2015).

The rejection decision of a given null model should not be interpreted in isolation, but rather in conjunction with the rejection decision of the associated alternative model. Thus, each single hypothesis consists of two parts; the first with what we have referred to as the null model in H_0 and the alternative model in H_1, and the second with the alternative model in H_0 and the null model in H_1. We also test the following multiple hypotheses:

- **Multiple Hypotheses:**

  Hypothesis IV: H_0: Fama and French (2015) vs. H_1: Pástor and Stambaugh (2003), Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005).

  Hypothesis V: H_0: Pástor and Stambaugh (2003) vs. H_1: Fama and French (2015), Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005).

  Hypothesis VI: H_0: Lettau and Ludvigson (2001) vs. H_1: Fama and French (2015), Pástor and Stambaugh (2003), and Lustig and Van Nieuwerburgh (2005).

  Hypothesis VII: H_0: Lustig and Van Nieuwerburgh (2005) vs. H_1: Fama and French (2015), Pástor and Stambaugh (2003), and Lettau and Ludvigson (2001).

  Hypothesis VIII: H_0: Fama and French (2015) vs. H_1: Lettau and Ludvigson (2001) and Lustig and Van Nieuwerburgh (2005).

Overall, there is no distinct pattern relating to n, the number dependent variables in the multivariate regression. When T is larger, the p-values are generally lower, meaning that the null hypothesis is rejected more strongly for the full sample.

Table 18 shows the result of testing H_0 : Fama and French (2015) vs. H_1: Pástor and Stambaugh (2003), and vice-versa. Regardless of portfolios type and error structure, we fail to reject the null hypothesis of the Fama and French (2015) model at the 5% level in the 1988-1998 period, while the Pástor and Stambaugh (2003) model is rejected at the 5% level. This implies that the profitability and investment factors of the Fama and French (2015) model are better suited to describe asset returns than the liquidity factor during this period. Outside of the 1988-1998 period, we rejected both models most of the time, suggesting that both models suffer from misspecification.

Table 19 displays the p-values for the hypotheses H_0: Fama and French (2015) vs. H_1: Lettau and Ludvigson (2001) and H_0: Lettau and Ludvigson (2001) vs. H_1: Fama and French (2015). In the case of Gaussian errors, we fail to reject the Fama and French (2015) model at the 5% level for all time periods. The lowest p-values are 0.0220, for the full sample and $n = 12$, and 0.0270, for the 1998Q1 - 2005Q1 period, with $n = 25$. On the other hand, we fail to reject the Lettau and Ludvigson (2001) model 5% for all but 5 cases. The results for the $t(5)$ distribution are similar. For the wild bootstrap, we fail to reject H_0 at the 5% level in all cases except for the 12-industry portfolios with the full sample under hypothesis *II.a*, and for the 25 size and book portfolios during the 1988Q1 - 1997Q4 period under hypothesis *II.b*. These results are consistent with Prescription 1 in Lewellen et al. (2010): it is much more difficult to reject a model when the dependent variables are sorted by the same characteristics as the right-hand side variables. Using portfolios based on exogenous sorts such as industry characteristics results in rejecting of the said models, at least in large samples.

Table 20 shows the p-values for hypothesis *III.a* and *III.b*. In the Gaussian case, the Lustig and Van Nieuwerburgh (2005) model is rejected in favour of the Fama and French (2015) model for $n = 5$, $n = 18$, and $n = 25$, at the 5% level for the full sample ($T = 149$). For smaller sample sizes ($T = 40$), the Fama and French (2015) is rejected in the direction of the Lustig and Van Nieuwerburgh (2005) model in the 1968Q1 - 1977Q4 period, for $n = 18$ ($p = 0.0280$). In turn, the Lustig and Van Nieuwerburgh (2005) model is rejected in the direction of the Fama and French (2015) model for the size portfolios during the 1978Q1 - 1987Q4 period. The failure to reject the Fama and French (2015) model is not as pronounced for this type of portfolios ($p = 0.0640$). Apart from these cases, we fail to reject both the Fama and French (2015) and the Lustig and Van Nieuwerburgh (2005) models at the 5% level. These decisions hold under the $t(5)$ error case, with the exception of the 5-industry portfolios for the full sample under hypothesis *III.b* ($p = 0.0750$). Accounting for heteroskedasticity in the errors via a wild bootstrap reveals that neither model can be rejected at the 5% level.

Table 21 shows the results of testing the Fama and French (2015) model against the union of the Pástor and Stambaugh (2003), Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005) models. Under the assumption of Gaussian errors, we fail to reject the null model at the 5% level, with the exception of the 1988Q1 - 1997Q4 with $n = 5$ and $n = 12$. The decision for the $t(5)$ case is the same as in the Gaussian case, except that for $n = 12$, we barely fail to reject the null model at the 5% level ($p = 0.0520$). For the wild bootstrap, we reject the Fama and French (2015) model for the 5-industry portfolios during the 1988Q1 - 1997Q4 period, and we fail to reject H_0 the rest of the time.

Tables 22 to 25 display the Monte Carlo p-values from testing a single model against a compound hypothesis, via the hypotheses *IV* through *VIII*. The Monte Carlo p-values are interpreted in a similar fashion as in Table 21. A caveat to this analysis remains, however. The reverse of these multiple model hypotheses is unclear. To know whether the null is rejected in the direction of the alternative, that is, in favour of the union of multiple alternative models, could require the use of a model confidence set, of which the study is left for future research.

Table 18: MC p-values for the multivariate J_A test, H_0: Fama and French (2015) vs. H_1: Pástor and Stambaugh (2003) and H_0: Pástor and Stambaugh (2003) vs. H_1: Fama and French (2015), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis I.a				*Hypothesis I.b*			
	Gaussian errors							
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
68 - 18	0.0010	0.0010	0.0010	0.0070	0.0010	0.0010	0.0010	0.0010
68 - 78	0.0010	0.0010	0.0200	0.0240	0.0010	0.0010	0.0010	0.0800
78 - 88	0.0030	0.0170	0.0390	0.1880	0.0040	0.0010	0.0010	0.0020
88 - 98	0.4330	0.3750	0.5470	0.5010	0.0010	0.0010	0.0040	0.0010
98 - 08	0.0050	0.0110	0.0100	0.0190	0.0750	0.0080	0.0130	0.0010
08 - 18	0.0010	0.0010	0.0100	0.0390	0.0010	0.0010	0.0010	0.0010
	$t(5)$ errors							
68 - 18	0.0010	0.0010	0.0010	0.0020	0.0010	0.0010	0.0010	0.0010
68 - 78	0.0010	0.0010	0.0310	0.0360	0.0010	0.0010	0.0010	0.0730
78 - 88	0.0040	0.0190	0.0370	0.2030	0.0010	0.0010	0.0010	0.0010
88 - 98	0.4430	0.3700	0.5890	0.5380	0.0010	0.0010	0.0110	0.0010
98 - 08	0.0040	0.0060	0.0050	0.0090	0.0880	0.0170	0.0160	0.0010
08 - 18	0.0010	0.0010	0.0230	0.0330	0.0010	0.0010	0.0010	0.0010
	Wild bootstrap							
68 - 18	0.0010	0.0010	0.0010	0.0860	0.0010	0.0010	0.0030	0.0010
68 - 78	0.0010	0.0010	0.1300	0.0910	0.0020	0.0010	0.0010	0.1330
78 - 88	0.0150	0.0450	0.0590	0.2530	0.0050	0.0010	0.0010	0.0030
88 - 98	0.5500	0.4500	0.5170	0.5360	0.0010	0.0010	0.0050	0.0010
98 - 08	0.0170	0.0350	0.0370	0.0270	0.2200	0.0560	0.0320	0.0010
08 - 18	0.0010	0.0020	0.0770	0.2830	0.0010	0.0010	0.0010	0.0010

Table 19: MC p-values for the multivariate J_A test, H_0: Fama and French (2015) vs. H_1: Lettau and Ludvigson (2001) and H_0: Lettau and Ludvigson (2001) vs. H_1: Fama and French (2015), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis II.a				*Hypothesis II.b*			
	Gaussian errors							
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.0870	0.0220	0.3540	0.3380	0.0170	0.0040	0.1630	0.0220
1968Q1 - 1977Q4	0.1160	0.2120	0.0810	0.3230	0.0790	0.2380	0.2570	0.1140
1978Q1 - 1987Q4	0.8860	0.9450	0.4090	0.2680	0.3540	0.4370	0.2080	0.3270
1988Q1 - 1997Q4	0.5680	0.1930	0.0980	0.7780	0.7190	0.0680	0.2660	0.0180
	$t(5)$ errors							
1968Q1 - 2005Q1	0.0870	0.0270	0.3720	0.3050	0.0220	0.0070	0.1730	0.0280
1968Q1 - 1977Q4	0.1380	0.2390	0.1010	0.3630	0.0840	0.2610	0.2800	0.1320
1978Q1 - 1987Q4	0.8900	0.9240	0.4690	0.3060	0.3730	0.4290	0.1740	0.3560
1988Q1 - 1997Q4	0.5650	0.2470	0.1410	0.7740	0.7200	0.0780	0.3020	0.0240
	Wild bootstrap							
1968Q1 - 2005Q1	0.1090	0.0490	0.5410	0.4690	0.1590	0.0830	0.8120	0.3280
1968Q1 - 1977Q4	0.2230	0.2840	0.0950	0.4270	0.4630	0.7150	0.6480	0.4740
1978Q1 - 1987Q4	0.9380	0.9590	0.7320	0.3170	0.4540	0.4430	0.2810	0.4290
1988Q1 - 1997Q4	0.5980	0.3650	0.0980	0.6800	0.6980	0.1820	0.3940	0.0140

Table 20: MC p-values for the multivariate J_A test, H_0: Fama and French (2015) vs. H_1: Lustig and Van Nieuwerburgh (2005) and H_0: Lustig and Van Nieuwerburgh (2005) vs. H_1: Fama and French (2015), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis III.a				*Hypothesis III.b*			
	Gaussian errors							
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.2870	0.2400	0.3130	0.6550	0.0470	0.1170	0.0120	0.0010
1968Q1 - 1977Q4	0.1030	0.0610	0.0280	0.6730	0.7970	0.9830	0.2440	0.4670
1978Q1 - 1987Q4	0.2790	0.4560	0.0640	0.7930	0.1590	0.4500	0.0390	0.3660
1988Q1 - 1997Q4	0.8350	0.5790	0.2500	0.4740	0.9210	0.8140	0.4900	0.2100
	$t(5)$ errors							
1968Q1 - 2005Q1	0.2750	0.2650	0.3660	0.5940	0.0750	0.1700	0.0190	0.0010
1968Q1 - 1977Q4	0.1010	0.0770	0.0340	0.6910	0.7810	0.9880	0.2960	0.4890
1978Q1 - 1987Q4	0.2690	0.4800	0.1020	0.8170	0.1790	0.4570	0.0350	0.3720
1988Q1 - 1997Q4	0.8190	0.6030	0.2940	0.4950	0.9190	0.7960	0.4910	0.2170
	Wild bootstrap							
1968Q1 - 2005Q1	0.3420	0.3920	0.3980	0.7010	0.3030	0.6320	0.6430	0.0860
1968Q1 - 1977Q4	0.1740	0.1750	0.1210	0.8360	0.9840	0.9990	0.5840	0.7580
1978Q1 - 1987Q4	0.3840	0.5640	0.1580	0.7560	0.2280	0.4940	0.0780	0.4540
1988Q1 - 1997Q4	0.7170	0.7130	0.3100	0.4140	0.9110	0.8510	0.6440	0.1030

Table 21: MC p-values for the multivariate J_A test, H_0: Fama and French (2015) vs. H_C: Pástor and Stambaugh (2003), Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis IV							
	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n=5$	$n=12$	$n=18$	$n=25$	$n=5$	$n=12$	$n=18$	$n=25$
1968Q1 - 2005Q1	0.5510	0.2750	0.2430	0.4650	0.5160	0.2810	0.2580	0.4140
1968Q1 - 1977Q4	0.0840	0.0890	0.1500	0.3500	0.0930	0.1090	0.2050	0.3820
1978Q1 - 1987Q4	0.5230	0.6670	0.2130	0.8050	0.5300	0.6840	0.2520	0.8280
1988Q1 - 1997Q4	0.0430	0.0440	0.2600	0.4820	0.0380	0.0520	0.3200	0.5220
	Wild bootstrap							
1968Q1 - 2005Q1	0.5490	0.3960	0.3710	0.6130				
1968Q1 - 1977Q4	0.1830	0.2700	0.3570	0.4940				
1978Q1 - 1987Q4	0.6370	0.7500	0.4070	0.8030				
1988Q1 - 1997Q4	0.0140	0.0990	0.2910	0.3760				

Table 22: MC p-values for the multivariate J_A test, H_0: Pástor and Stambaugh (2003) vs. H_C: Fama and French (2015), Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis V							
	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n=5$	$n=12$	$n=18$	$n=25$	$n=5$	$n=12$	$n=18$	$n=25$
1968Q1 - 2005Q1	0.1000	0.0740	0.0070	0.0490	0.1290	0.0940	0.0210	0.0570
1968Q1 - 1977Q4	0.0920	0.1920	0.2770	0.2570	0.1120	0.2280	0.3310	0.2710
1978Q1 - 1987Q4	0.3400	0.5550	0.3240	0.7610	0.3600	0.5650	0.3120	0.7840
1988Q1 - 1997Q4	0.3590	0.0310	0.7800	0.0580	0.3310	0.0550	0.7670	0.0690
	Wild bootstrap							
1968Q1 - 2005Q1	0.1290	0.0940	0.0210	0.0570				
1968Q1 - 1977Q4	0.1120	0.2280	0.3310	0.2710				
1978Q1 - 1987Q4	0.3600	0.5650	0.3120	0.7840				
1988Q1 - 1997Q4	0.3310	0.0550	0.7670	0.0690				

Table 23: MC p-values for the multivariate J_A test, H_0: Lettau and Ludvigson (2001) vs. H_C: Fama and French (2015), Pástor and Stambaugh (2003), and Lustig and Van Nieuwerburgh (2005), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis VI							
	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n=5$	$n=12$	$n=18$	$n=25$	$n=5$	$n=12$	$n=18$	$n=25$
1968Q1 - 2005Q1	0.1460	0.0930	0.1310	0.1270	0.1730	0.1060	0.1610	0.1250
1968Q1 - 1977Q4	0.0680	0.0830	0.1180	0.1710	0.0790	0.0880	0.1700	0.1850
1978Q1 - 1987Q4	0.2090	0.4160	0.2150	0.7890	0.2230	0.4050	0.2010	0.8110
1988Q1 - 1997Q4	0.1910	0.0200	0.5620	0.1920	0.1770	0.0270	0.5900	0.2120
	Wild bootstrap							
1968Q1 - 2005Q1	0.3250	0.3150	0.5930	0.4800				
1968Q1 - 1977Q4	0.3320	0.4100	0.5010	0.5250				
1978Q1 - 1987Q4	0.2540	0.4520	0.2520	0.7540				
1988Q1 - 1997Q4	0.2040	0.0780	0.6790	0.1670				

Table 24: MC p-values for the multivariate J_A test, H_0: Lustig and Van Nieuwerburgh (2005) vs. H_C: Fama and French (2015), Pástor and Stambaugh (2003), and Lettau and Ludvigson (2001), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis VII							
	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n=5$	$n=12$	$n=18$	$n=25$	$n=5$	$n=12$	$n=18$	$n=25$
1968Q1 - 2005Q1	0.2160	0.2170	0.0240	0.0020	0.2420	0.2310	0.0250	0.0010
1968Q1 - 1977Q4	0.5140	0.6840	0.5340	0.5540	0.4940	0.6960	0.5400	0.6030
1978Q1 - 1987Q4	0.2620	0.3810	0.0840	0.3650	0.2840	0.4170	0.0770	0.3760
1988Q1 - 1997Q4	0.0510	0.1440	0.3430	0.4910	0.0490	0.1620	0.3790	0.5440
	Wild bootstrap							
1968Q1 - 2005Q1	0.4340	0.5950	0.5330	0.1070				
1968Q1 - 1977Q4	0.8430	0.8730	0.7450	0.8070				
1978Q1 - 1987Q4	0.4020	0.5470	0.1780	0.3580				
1988Q1 - 1997Q4	0.0280	0.3310	0.5220	0.3550				

Table 25: MC p-values for the multivariate J_A test, H_0: Fama and French (2015) vs. H_C: Lettau and Ludvigson (2001), and Lustig and Van Nieuwerburgh (2005), parametric bootstrap with Gaussian and $t(5)$ errors, and wild bootstrap with Rademacher distribution.

	Hypothesis VIII							
	Gaussian errors				$t(5)$ errors			
	Industry		Size	Size & Book	Industry		Size	Size & Book
T	$n = 5$	$n = 12$	$n = 18$	$n = 25$	$n = 5$	$n = 12$	$n = 18$	$n = 25$
1968Q1 - 2005Q1	0.3810	0.2150	0.2380	0.5500	0.3540	0.2270	0.2680	0.4950
1968Q1 - 1977Q4	0.1150	0.0870	0.1140	0.4960	0.1070	0.1110	0.1340	0.5220
1978Q1 - 1987Q4	0.5120	0.5730	0.1880	0.7340	0.5360	0.6140	0.2430	0.7700
1988Q1 - 1997Q4	0.0970	0.1080	0.3420	0.7700	0.1170	0.1460	0.3960	0.7830
	Wild bootstrap							
1968Q1 - 2005Q1	0.3630	0.2770	0.3230	0.6050				
1968Q1 - 1977Q4	0.2100	0.1950	0.2230	0.6220				
1978Q1 - 1987Q4	0.6910	0.6930	0.3990	0.7640				
1988Q1 - 1997Q4	0.0540	0.2000	0.4260	0.6860				

5 Conclusion

We proposed multivariate extensions of exact specification tests for non-nested models that can accommodate both simple and multiple alternatives. Our approach, which uses pseudoinverses to bypass regularity problems, yields an exact test in the Gaussian setting, as was pointed in the literature for the univariate case. The MC p-value approach provides valid results, whether the null distribution depends on nuisance parameters or not. Our extension to multiple non-nested alternatives via a combined hypothesis addresses the growing problem of model selection in asset pricing, but is also applicable in any field where specification tests are necessary. Our simulation studies have shown that the multivariate J_A test enjoys good size and power properties, under both the Gaussian and non-Gaussian errors. Moreover, we have shown via simulations that applying our method to the multivariate J test helps to correct size distortions that occur in small samples. Our empirical results showed evidence of misspecification for the Fama and French (2015) and Pástor and Stambaugh (2003) models, as the test rejected these prominent models for most time periods. For most time periods, the Fama and French (2015) model was not rejected against a compound alternative hypothesis of multiple models.

A Appendix

A.1 Proof of Lemma 2.1

Proof. Let $y_i \in \mathbb{R}^T$ denote the i^{th} column of Y. Consider $i \leq n$ and let d_i denote the event that y_i is linearly dependent of y_j, for all $j \neq i$. Let B be a Borel set of Lesbegue measure zero. Since U_k is absolutely continuous, so is y_i. The event that y_i falls into any set $B \in \mathbb{R}^T$ has zero probability, by definition of an absolutely continuous random variable. Thus, conditioning on all X_k, $P(d_i) = 0$ for all i, and the probability that all y_i's are linearly dependent of each other is $P(\cup_{i=1}^n d_i)$. By the union bound, we have $P(\cup_{i=1}^n d_i) \leq \sum_{i=1}^n P(d_i)$ for all B. Then, the probability that Y has full-column rank is $1 - P(\cup_{i=1}^n d_i) \geq 1 - \sum_{i=1}^n P(d_i)$. Since $P(d_i) = 0$ for all i, Y has full-column rank with probability 1. $\square$

References

Andrews, D. W. (1987): "Asymptotic results for generalized Wald tests," *Econometric Theory*, 348–358.

Atkinson, A. C. (1970): "A method for discriminating between models," *Journal of the Royal Statistical Society: Series B (Methodological)*, 32, 323–345.

Bernanke, B., H. Bohn, and P. C. Reiss (1988): "Alternative non-nested specification tests of time-series investment models," *Journal of Econometrics*, 37, 293–326.

Berndt, E. R. and N. E. Savin (1977): "Conflict among criteria for testing hypotheses in the multivariate linear regression model," *Econometrica: Journal of the Econometric Society*, 1263–1277.

Chen, L., M. Pelger, and J. Zhu (2019): "Deep learning in asset pricing," *Available at SSRN 3350138.*

Chen, N.-F., R. Roll, and S. A. Ross (1986): "Economic forces and the stock market," *Journal of Business*, 383–403.

Cox, D. R. (1961): "Tests of separate families of hypotheses," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, 105, 23.

——— (1962): "Further results on tests of separate families of hypotheses," *Journal of the Royal Statistical Society: Series B (Methodological)*, 24, 406–424.

Daniel, K., D. Hirshleifer, and L. Sun (2019): "Short-and long-horizon behavioral factors," *The Review of Financial Studies.*

Davidson, R. and E. Flachaire (2008): "The wild bootstrap, tamed at last," *Journal of Econometrics*, 146, 162–169.

Davidson, R. and J. G. MacKinnon (1981): "Several tests for model specification in the presence of alternative hypotheses," *Econometrica: Journal of the Econometric Society*, 781–793.

——— (1983): "Testing the specification of multivariate models in the presence of alternative hypotheses," *Journal of Econometrics*, 23, 301–313.

——— (2002): "Bootstrap J tests of nonnested linear regression models," *Journal of Econometrics*, 109, 167–193.

Dufour, J.-M. (2006): "Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics," *Journal of Econometrics*, 133, 443–477.

Dufour, J.-M. and L. Khalaf (2002): "Simulation based finite and large sample tests in multivariate regressions," *Journal of Econometrics*, 111, 303–322.

Eguchi, S. and J. Copas (2006): "Interpreting kullback–leibler divergence with the neyman–pearson lemma," *Journal of Multivariate Analysis*, 97, 2034–2040.

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

——— (2015): "A five-factor asset pricing model," *Journal of Financial Economics*, 116, 1–22.

Fama, E. F. and J. D. MacBeth (1973): "Risk, return, and equilibrium: Empirical tests," *Journal of political economy*, 81, 607–636.

Feng, G., S. Giglio, and D. Xiu (2017): "Taming the factor zoo," *Fama-Miller Working Paper*, 24070.

Fisher, G. R. and M. McAleer (1981): "Alternative procedures and associated tests of significance for non-nested hypotheses," *Journal of Econometrics*, 16, 103–119.

Freyberger, J., A. Neuhierl, and M. Weber (2017): "Dissecting characteristics nonparametrically," *Review of Financial Studies (forthcoming)*.

Godfrey, L. G. (1983): "Testing non-nested models after estimation by instrumental variables or least squares," *Econometrica: Journal of the Econometric Society*, 355–365.

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 33, 2223–2273.

Hagemann, A. (2012): "A simple test for regression specification with non-nested alternatives," *Journal of econometrics*, 166, 247–254.

Harvey, C. R., Y. Liu, and H. Zhu (2016): ". . . and the cross-section of expected returns," *The Review of Financial Studies*, 29, 5–68.

Hoel, P. G. (1947): "On the choice of forecasting formulas," *Journal of the American Statistical Association*, 42, 605–611.

Kozak, S., S. Nagel, and S. Santosh (2019): "Shrinking the cross-section," *Journal of Financial Economics*.

Lettau, M. and S. Ludvigson (2001): "Consumption, aggregate wealth, and expected stock returns," *the Journal of Finance*, 56, 815–849.

Lewellen, J., S. Nagel, and J. Shanken (2010): "A skeptical appraisal of asset pricing tests," *Journal of Financial economics*, 96, 175–194.

Lustig, H. N. and S. G. Van Nieuwerburgh (2005): "Housing collateral, consumption insurance, and risk premia: An empirical perspective," *The Journal of Finance*, 60, 1167–1219.

MacKinnon, J. G. (1983): "Model specification tests against non-nested alternatives," *Econometric Reviews*, 2, 85–110.

McAleer, M. (1981): "A small sample test for non-nested regression models," *Economics Letters*, 7, 335–338.

Milliken, G. A. and F. A. Graybill (1970): "Extensions of the general linear hypothesis model," *Journal of the American Statistical Association*, 65, 797–807.

Pástor, L. and R. F. Stambaugh (2003): "Liquidity risk and expected stock returns," *Journal of Political Economy*, 111, 642–685.

Pesaran, M. H. (1974): "On the general problem of model selection," *The Review of Economic Studies*, 41, 153–171.

Pesaran, M. H. and A. S. Deaton (1978): "Testing non-nested nonlinear regression models," *Econometrica: Journal of the Econometric Society*, 677–694.

Pesaran, M. H. and M. Weeks (2001): "Non-nested hypothesis testing: an overview," *A companion to theoretical econometrics*, 279–309.

Rao, C. R. (1951): "An Asymptotic Expansion of the Distribution of Wilks' Criterion," *Bulletin of the International Statistical Institute*, 33, 177–180.

——— (1973): *Linear statistical inference and its applications*, vol. 2, Wiley New York.

Richard, F. (2020): "Simultaneous Inference in Multivariate Models," .

Shanken, J. and M. I. Weinstein (2006): "Economic forces and the stock market revisited," *Journal of Empirical Finance*, 13, 129–144.

Stewart, K. G. (1997): "Exact testing in multivariate regression," *Econometric reviews*, 16, 321–352.

Vuong, Q. H. (1989): "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: Journal of the Econometric Society*, 307–333.

Wilks, S. S. (1938): "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, 9, 60–62.