Model Confidence Sets in Multivariate Systems

Florian Richard*
Département de finance, assurance et immobilier, Université Laval†

July 25, 2023

Abstract

This paper provides a model selection procedure for multivariate models, generalizing the model confidence set (MCS) procedure to systems of $N > 1$ dependent variables. A $(1 - \alpha)$ level MCS collects the set of models with equal predictive ability, based on a sequential elimination procedure that relies on an equivalence test. The latter relies on supremum-type t and Hotelling-type T^2 statistics which account for correlation between loss differentials. The procedure shows good size and power properties in simulations. The performance of 14 candidate asset pricing models is assessed using the Fama and French research portfolios, with monthly data for the period 1972-2013. Under quadratic loss, the MCS contains at most one model for out-of-sample tests, but it often includes multiple competing models for in-sample tests: models are much harder to distinguish. Overall, out-of-sample tests and a larger number of more heterogenous test assets provide more information to disentangle models. The market-based capital asset pricing model is never included in the MCS.

Keywords: Equal predictive ability, confidence sets, factor models, multiple testing.
JEL Codes: C52, C53, G12.

*I am thankful to James MacKinnon and to the participants of the 2022 Econometric Society European Meetings, the 2022 Australasian Meeting of the Econometric Society, New York Camp Econometrics XVI, the 37th Annual Meeting of the Canadian Econometrics Study Group, the 60th Annual Meeting of the Société canadienne de science économique, the 2021 Conference of the International Association for Applied Econometrics, the 27th International Conference Computing in Economics and Finance, the 55th Annual Conference of the Canadian Economics Association, and the Sixteenth CIREQ Ph.D. Students' Conference for the insightful comments. I am grateful to Kevin Sheppard for providing for the code for the univariate model confidence set procedure.

†Email: florian.richard@fsa.ulaval.ca. Website: https://sites.google.com/view/florianrichard. Université Laval, 2325 rue de la Terrasse, Local 3602, Québec QC, Canada G1V 0A6.

1 Introduction

Choosing among competing models and associated multiple testing is pervasive in economics and finance [see Harvey et al. (2016), Harvey and Liu (2020), Giglio et al. (2021), Grønborg et al. (2021), and Wang et al. (2023)]. Both empirical and theoretical research typically provide support for multiple models, and challenges related to multiple testing remain [see for example Romano and Lehmann (2005), Romano and Wolf (2016), List et al. (2019), and Spreng and Urga (2022)]. In this context, a natural question is whether empirical models can be differentiated statistically. Moreover, the ability for empirical models to explain multiple dependent variables is crucial in data-rich environments, whereby systems of $N > 1$ predictive equations raise dimensionality problems that have received less attention to date.

This paper makes three contributions: (*i*) I generalize the work of Hansen et al. (2011) to systems of $N > 1$ dependent variables, providing a statistically meaningful solution to unresolved multiple comparisons issues by using a confidence set approach; (*ii*) in view of the dimensionality, I propose supremum t and Hotelling T^2 statistics, which account for potential correlations across the N-variate loss differentials, to test the equal predictive ability of candidate models; and (*iii*) empirically, I provide an alternative perspective on underlying asset pricing issues, in particular, on the relevance of various anomalies, the information content of in- and out-of-sample assessments, and of predictions across short and long term horizons.

Sequential pairwise testing introduces well-known statistical problems, often characterized by the inflation of the family-wise error rate (FWER), *i.e.* the probability of making at least one type I error. Although Bonferroni- and Šidák-type corrections offer a solution, the former often results in decreased statistical power when dealing with a large number of tests, and the latter can be either excessively liberal or too conservative when tests are dependent; see Montiel Olea and Plagborg-Møller (2019) for a recent review of these issues. I propose to use the MCS procedure, which controls the asymptotic FWER at level α, to address these issues. The MCS procedure, consisting of a series of tests of equal predictive ability, outputs a set of models with equal statistical performance and known confidence level: the $(1-\alpha)$ level MCS, denoted $\widehat{\mathcal{M}}^*_{1-\alpha}$, is the set of *best* models from a collection of candidate models $\mathcal{M}^0$, *i.e.*, the set of models that survived a sequential selection procedure based on an equivalence test $\delta_{\mathcal{M}}$, assessed in- or out-of-sample, and an elimination rule $e_{\mathcal{M}}$, both determined by the user. Similarly to confidence intervals for point estimates, the MCS selects a set of models for some confidence level $(1-\alpha)$: $\widehat{\mathcal{M}}^*_{1-\alpha}$ covers the set of models with equal predictive ability with probability $(1-\alpha)$.[1]

The usefulness of the MCS procedure can best be illustrated through the many published empirical models in the beta-pricing context. Lewellen et al. (2010) qualifies these findings as "*an embarrassment of riches*". In his seminal contribution, Harvey (2017) draws serious attention to the underlying multiple testing problems, and suggests raising the hurdle for the discovery of new factors: "*our standard testing methods are often ill-equipped to answer the questions that we pose*". Harvey et al. (2016) also documents more than 300 possible risk factors since 1964. As recent empirical work points to more stringent statistical standards to achieve factor significance [Lewellen et al. (2010), Harvey and Liu (2019), Gospodinov and Robotti (2021)], the MCS procedure provides a formal confidence set with known confidence level $(1-\alpha)$, to infer the set of models with equal predictive ability.[2] The MCS procedure offers the additional benefit of accommodating misspecification among the candidate models. As the null hypothesis of the MCS is that the loss functions are indistinguishable, it is possible to select a misspecified model. The MCS procedure yields a non-empty set, ensuring the identification of a winning model, even when all candidates

[1] The MCS can be interpreted as a test inversion of a sequential procedure, although Hansen et al. (2011) do not present it in this way.

[2] For applications of the MCS procedure in finance, see Hansen et al. (2003), Liu et al. (2015), Varneskov (2016), and Grønborg et al. (2021).

are misspecified.

This paper primarily focuses on addressing the challenge of effectively combining and summarizing the information contained in multidimensional loss differentials while ensuring the validity of the MCS procedure. This paper offers three keys contributions.

First, I provide an extension of the MCS procedure to multivariate losses under general assumptions, and present the asymptotic theory in the N-equation case. Stacking the loss differential vectors preserves the necessary statistical properties to use a quadratic form statistic, and captures the information content of multidimensional dependent variables. Currently, an approach that can accommodate N-dimensional loss functions for the computation of the MCS has not been considered to date.

Second, I propose to use two statistics, a supremum t (or sup t) statistic, and a Hotelling T^2 statistic, for computing the MCS. The sup t statistic, computed on an equation-by-equation basis, allows for a large cross section (N) relative to the sample size (T), a useful feature when testing a large number of dependent variables. The Hotelling T^2 statistic, adapted from the multivariate test of equal predictive ability of Mariano and Preve (2012), effectively summarizes the information contained within the system of N equations. Both statistics consider the correlations between loss differentials through the application of a moving block bootstrap. I show the validity of the bootstrap implementation using the proposed statistics, and present the conventional size and power properties through two simulation designs: (i) a design based on dependent losses drawn from a multivariate normal distribution, encompassing varying parameter values for between-model and within-model correlations; and (ii) an empirically relevant design which uses regression models, parameterized according to estimated models found in the literature. The procedure works well in terms of both size and power. When considering a single "best" model based on a quadratic loss function, the MCS behaves as theoretically predicted in Corollary 1 of Hansen et al. (2011). For multiple "best" models, the procedure achieves the conventional coverage probability for reasonable sample sizes, and in most cases, eliminates all inferior models when the sample size is smaller than 1,000. Empirically relevant simulations underline the importance of the market factor.

Third, I provide an analysis of a large set of candidate factor models using the MCS approach, focusing on the Fama and French research portfolios as dependent variables, using monthly data spanning from 1972 to 2013. For both statistics, different candidate models are selected by the out-of-sample tests, including accounting-, mispricing- and liquidity-based factors, although at most one model is selected. For in-sample predictions, distinguishing candidate models becomes more challenging, particularly for a smaller number of test portfolios. In particular, the MCS always excludes the model that contains only the market risk premium. In light of the conflicting evidence regarding the effectiveness of the market factor, this result supports recent work on the success of the market factor, notably Harvey and Liu (2021), establishing its importance in conjunction with other risk factors, namely the size factor. Ultimately, using out-of-sample predictions proves valuable in discerning competing models.

Section 2 outlines the theoretical framework and the MCS procedure. I present simulation results in Section 3, and the empirical analysis in Section 4. Section 5 concludes.

1.1 Notation

Convergence in probability is denoted by $\xrightarrow{d}$. A m-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $N_m(\boldsymbol{\mu}_m, \boldsymbol{\Sigma})$, and $\mathbf{0}_N$ denotes the N-dimensional vector of zeros.

2 Framework

This section presents the econometric framework for multivariate losses, the model confidence set procedure, and the proposed sup t and Hotelling T^2 statistics.

2.1 Setting

To define the MCS procedure within the framework of a multivariate test of equal predictive ability, consider the multivariate stochastic process $\boldsymbol{W} \equiv \{\boldsymbol{W}_t : \Omega \to \mathbb{R}^{K_i+N}, i \in \{1, \ldots, m\}, t = 1, \ldots, T\}$ where K_i is the number of predictor variables in model i, N is the number of dependent variables, m is the number of models under consideration, and T is the sample size. $\boldsymbol{W}$ is defined on the complete probability space $(\Omega, \mathcal{F}, P)$, where Ω is a sample space, $\mathcal{F}$ is a σ-field on Ω, and P is a probability measure. Define $\boldsymbol{W}_t = (\boldsymbol{Y}_t', \boldsymbol{X}_t')$, where $\boldsymbol{Y}_t : \Omega \to \mathbb{R}^N$ denotes a vector of dependent variables, $\boldsymbol{X}_t : \Omega \to \mathbb{R}^{K_i}$ denotes a vector of predictors, and $\mathcal{F}_t = \sigma(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_t)$ denotes the σ-field generated from the history of $\boldsymbol{W}_t$. The predicted values of $\boldsymbol{Y}_t$ using model i at the forecasting horizon h are denoted as

$$\widehat{\boldsymbol{f}}_{i,t+h} = \boldsymbol{f}_i(\boldsymbol{W}_t, \boldsymbol{W}_{t-1}, \ldots, \boldsymbol{W}_1; \widehat{\boldsymbol{\beta}}_{i,l}), \tag{2.1}$$

where $\boldsymbol{f}_i$ is a measurable-$\mathcal{F}_t$ forecasting function for model i, and $\widehat{\boldsymbol{\beta}}_{i,l}$ is a $(K_i \times N)$ vector of estimated parameters with bootstrap block length l. In-sample forecasts are denoted by the special case $h = 0$ and $\mathbf{W}_t = \mathbf{X}_t$. Each forecasting model i admits a loss function $\boldsymbol{L}_{i,t+h} = \boldsymbol{L}(\boldsymbol{Y}_{t+h}, \widehat{\boldsymbol{f}}_{i,t+h})$ at time t. A popular choice for the loss function includes the quadratic loss $L_{i,t+h}^{(n)} = (Y_{t+h}^{(n)} - \widehat{f}_{i,t+h}^{(n)})^2$, for $n = 1, \ldots, N$, where $Y_{t+h}^{(n)}$ is the observation for the n^{th} dependent variable at time $t + h$, and $\widehat{f}_{i,t+h}^{(n)}$ is the forecast for the n^{th} dependent variable at time $t + h$. Based on a loss function for models i and j, and a σ-field $\mathcal{G}_t$, the null hypothesis of *unconditional equal predictive ability* between models i and j is

$$H_{0,\mathcal{M}} : \mathbb{E}[\boldsymbol{L}(\boldsymbol{Y}_{t+h}, \widehat{\boldsymbol{f}}_{i,t+h}) - \boldsymbol{L}(\boldsymbol{Y}_{t+h}, \widehat{\boldsymbol{f}}_{j,t+h})] = \mathbb{E}[\boldsymbol{d}_{ij,t+h}] = \mathbf{0}_N \quad \forall\, i, j \in \mathcal{M}, \tag{2.2}$$

where $\boldsymbol{d}_{ij,t+h}$ is the *loss differential* between models i and j at time $t+h$, and $\mathcal{M}$ is the set of models under consideration. In this multivariate context, the next section presents the MCS procedure.

2.2 The Model Confidence Set Procedure

The multivariate counterpart of Hansen et al. (2011)'s "set of superior objects" is

$$\mathcal{M}^* \equiv \left\{ i \in \mathcal{M}^0 : \boldsymbol{\mu}_{ij} \leq \mathbf{0}_N \quad \text{for all} \quad j \in \mathcal{M}^0 \right\}. \tag{2.3}$$

The central idea of the MCS procedure is to determine whether a given model i belongs to the set of superior objects. The number of models in $\mathcal{M}$ is m, such that the elements in $\mathcal{M}$ are $i_1, \ldots, i_m$. Following the notation of Hansen et al. (2011), the MCS procedure relies on an equivalence test $\delta_{\mathcal{M}}$ to test $H_{0,\mathcal{M}}$ and an elimination rule $e_{\mathcal{M}}$ to eliminate models in $\mathcal{M}$. The equivalence test takes on values $\delta_{\mathcal{M}} = 0$ if $H_{0,\mathcal{M}}$ is not rejected, and $\delta_{\mathcal{M}} = 1$ if $H_{0,\mathcal{M}}$ is rejected. The elimination rule $e_{\mathcal{M}}$ determines the model removed from $\mathcal{M}$ when $\delta_{\mathcal{M}} = 1$. The output of this algorithm is $\widehat{\mathcal{M}}^*_{1-\alpha}$, the model confidence set. Figure 1 outlines the procedure for determining $\widehat{\mathcal{M}}^*_{1-\alpha}$. The assumptions of Hansen et al. (2011) with regards to asymptotic level and power apply to the multivariate case, and are stated in the Supplementary Material for completeness. The MCS procedure also produces p-values. The p-value $\hat{p}_i$ for model i is defined as the smallest p-value such that model i belongs to the MCS. Thus, a model with $\hat{p}_i = 1$ will be included in the confidence set. This p-value is given by $\hat{p}_{e_{\mathcal{M}_j}} = \max_{i \leq j} P_{H_{0,\mathcal{M}_i}}$, where $P_{H_{0,\mathcal{M}_i}}$ is the p-value that corresponds to the null hypothesis $H_{0,\mathcal{M}_i}$. In a multivariate setting, we can test $H_{0,\mathcal{M}}$ using a sup t and a Hotelling T^2 statistic, as outlined in the next section.

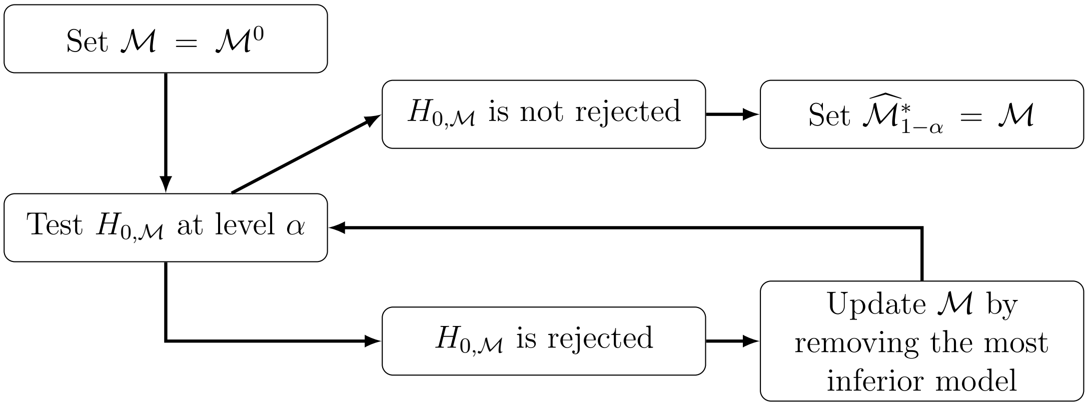

Set $\mathcal{M} = \mathcal{M}^0$
$H_{0,\mathcal{M}}$ is not rejected
Set $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$
Test $H_{0,\mathcal{M}}$ at level α
$H_{0,\mathcal{M}}$ is rejected
Update $\mathcal{M}$ by removing the most inferior model

Figure 1: The model confidence set algorithm. $\mathcal{M}^0$ denotes the set of candidate models. After updating the placeholder set $\mathcal{M}$ as necessary, the procedure stops at the first non-rejection, and the remaining models form the estimated MCS, $\widehat{\mathcal{M}}^*_{1-\alpha}$.

2.3 A Multivariate Test of Unconditional Equal Predictability

To state the central limit theorem that justifies the use of a general quadratic form test statistic, it is necessary to impose two assumptions on the loss differentials:

Assumption 1. $\{\boldsymbol{d}_{ij,t}\}_{i,j\in\mathcal{M}^0}$ *is mixing of size* $-r/(r-2)$ *for* $r > 2$. *Additionally,* $\mathbb{E}\|\boldsymbol{d}_{ij,t}\|^r < \infty$.

Assumption 2. $\{\boldsymbol{d}_{ij,t}\}_{i,j\in\mathcal{M}^0}$ *is covariance stationary.*

Assumption (2) assumes stationarity on the loss differential, but not on the losses themselves.

2.3.1 Hotelling T^2 Statistic

The T^2 statistic provides information about the joint significance of the loss differentials. It can be viewed as a weighted average of the squared means of the loss differentials for all equations. Hotelling-type statistics remain popular to assess multivariate models in this context [see Gibbons et al. (1989), Shanken (1985), and Beaulieu et al. (2023)]. In order to construct a Hotelling T^2 statistic to test the hypothesis that models i and j have equal predictive ability, consider the transformed losses $V_t^n = \iota_\perp' L_t^{(n)}$, where $\iota_\perp'$ is the orthogonal complement of the m-dimensional vector $\iota = (1, \dots, 1)'$. Then, stack the V_t^n vectors into the $N(m-1)$-dimensional vector

$$\boldsymbol{V}_t = \left[V_t^1 \dots V_t^N\right]' = (I_N \otimes \iota_\perp) \left[L_t^{(1)} \dots L_t^{(N)}\right]', \tag{2.4}$$

with expectation $\boldsymbol{\theta} = \mathbb{E}[\boldsymbol{V}_t]$. Since $\iota_\perp' \iota = 0$, the transformed variable $\boldsymbol{V}_t$ has expectation equal to $\mathbf{0}_{N(m-1)}$ under $H_{0,\mathcal{M}}$. Then, under Assumptions (1) and (2), Lemma 2.1 justifies the use of a quadratic form-type statistic. The proof is available in the Supplementary Material.

Lemma 2.1. *Testing $H_{0,\mathcal{M}} : \boldsymbol{\mu_{ij}} = \mathbf{0}_N$ is equivalent to testing $H_{0,\mathcal{M}} : \boldsymbol{\theta} = \mathbf{0}_{N(m-1)}$. Furthermore, for a fixed N, we have:*

$$T^{1/2}(\bar{\boldsymbol{V}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}_{N(m-1)}, \boldsymbol{\Omega}) \tag{2.5}$$

where $\bar{\boldsymbol{V}} = T^{-1} \sum_{t=1}^T \boldsymbol{V}_t$, $\boldsymbol{\Omega} = \lim_{T\to\infty} var(T^{1/2}\bar{\boldsymbol{V}})$.

The covariance matrix $\boldsymbol{\Omega}$ is estimated using a heteroskedasticity and autocorrelation consistent estimator via a moving block bootstrap, following Gonçalves and White (2005). The proposed Hotelling-type T^2 statistic is based on Mariano and Preve (2012)'s statistic for tests for equal predictive ability. I construct the Hotelling T^2 statistics

$$T_{ij}^2 = T(\bar{\boldsymbol{d}}_{ij} - \boldsymbol{\mu}_{ij}^0)' \boldsymbol{\Sigma}_{ij}^{-1} (\bar{\boldsymbol{d}}_{ij} - \boldsymbol{\mu}_{ij}^0), \quad \text{and} \quad T_{i\cdot}^2 = T(\bar{\boldsymbol{d}}_{i\cdot} - \boldsymbol{\mu}_{i\cdot}^0)' \boldsymbol{\Sigma}_{i\cdot}^{-1} (\bar{\boldsymbol{d}}_{i\cdot} - \boldsymbol{\mu}_{i\cdot}^0), \tag{2.6}$$

where $\bar{\boldsymbol{d}}_{ij}$ is the sample counterpart of $\boldsymbol{\mu}_{ij}$, $\bar{\boldsymbol{d}}_{i\cdot}$ is the average of $\bar{\boldsymbol{d}}_{ij}$ over m, and $\boldsymbol{\mu}_{ij}^0$ and $\boldsymbol{\mu}_{i\cdot}^0$ are the values of $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\mu}_{i\cdot}$ under $H_{0,\mathcal{M}}$. The covariance matrices $\boldsymbol{\Sigma}_{ij} = T^{-1}(\boldsymbol{d}_{ij,t} -$

$\bar{\boldsymbol{d}}_{ij})(\boldsymbol{d}_{ij,t} - \bar{\boldsymbol{d}}_{ij})'$ and $\boldsymbol{\Sigma}_{i\cdot} = T^{-1}(\boldsymbol{d}_{i\cdot,t} - \bar{\boldsymbol{d}}_{i\cdot})(\boldsymbol{d}_{i\cdot,t} - \bar{\boldsymbol{d}}_{i\cdot})'$ are computed with a moving block bootstrap. The resulting statistic for the elimination rule $e_{\mathcal{M}}$ can be written as

$$T_{\max,\mathcal{M}} \equiv \max_{i,j\in\mathcal{M}} T^2_{i\cdot}. \tag{2.7}$$

2.3.2 Supremum t Statistic

The supremum t statistic is based on selecting the largest t statistic among the statistics computed from the individual estimation of each equation in the multivariate system.[3] The supremum statistic has the advantage of two attractive properties: (i) it does not require the inversion of a possibly high-dimensional covariance matrix; and (ii) by estimating the system equation-by-equation, it allows extensions to the case where $N > T$. These advantages may come at the expense of power compared to the T^2 statistic, particularly the residuals of the regressions are only weakly correlated, as is shown later in simulations. For each equation $n \in \{1, \ldots, N\}$, I use a t-statistic to test each of the sub-null hypotheses

$$H^{ij}_{0,\mathcal{M}} : \boldsymbol{\mu}_{ij} \in \mathcal{H}_0 \quad \text{for all } i,j \in \mathcal{M} \qquad \text{and} \qquad H^{ij}_{A,\mathcal{M}} : \boldsymbol{\mu}_{ij} \in \mathcal{H}_A \quad \text{for all } i,j \in \mathcal{M}, \tag{2.8}$$

where $\mathcal{H}_0$ is the set compatible with the null hypothesis $H_{0,\mathcal{M}}$, and $\mathcal{H}_A$ is the set compatible with the alternative hypothesis $H_{A,\mathcal{M}}$. When the distribution of the statistic can be simulated, bootstrapping the critical values avoids costly level adjustments when dealing with combined statistics, while maintaining the dependence between the statistics; see Dufour et al. (2015) and Harvey and Liu (2021) for assumptions under which the bootstrap is asymptotically valid. The supremum t statistics take the following form:

$$t^n_{\mathcal{M},\sup} = \sup_{i,j} t^n_{ij} = \sup_{i,j}\left[\bar{d^n_{ij}}/\sqrt{\mathrm{var}(\bar{d^n_{ij}})}\right] \qquad \text{and} \qquad t_{\mathcal{M},\sup} = \sup_n t^n_{\mathcal{M},\sup}, \tag{2.9}$$

[3] Other criteria exist to combine p-values, see Tippett et al. (1931); Fisher (1932); Pearson (1933), and Bergamelli et al. (2019); Dufour et al. (2015); Spreng and Urga (2022) for applications.

for $n = 1, \ldots, N$, where $\mathrm{var}(\bar{d}_{ij}^n)$ is estimated via bootstrap. The statistic $t^n_{\mathcal{M},\sup}$ is used to compute the t statistic equation-by-equation, and the statistic $t_{\mathcal{M},\sup}$ is the elimination rule $e_{\mathcal{M}}$.

2.4 Bootstrap Implementation and Validity

Once the MCS p-values are computed, the elimination rules are used to exclude the inferior models. The block bootstrap procedure for multivariate loss functions is detailed in the Supplementary Material. To show the validity of the bootstrap procedure, consider the $(Nm \times 1)$ vector of multivariate sample losses

$$\bar{\boldsymbol{Z}} = (\bar{\boldsymbol{d}}_{1\cdot}, \ldots, \bar{\boldsymbol{d}}_{m\cdot})', \tag{2.10}$$

which is expressed as a linear transformation of $\bar{\boldsymbol{V}}$: $\bar{\boldsymbol{Z}} = \boldsymbol{G}'\bar{\boldsymbol{V}}$, where $\boldsymbol{G}$ is a fixed $((m-1)N \times Nm)$ matrix. Consequently, the result follows in Lemma 2.2. The proof is stated in the Supplementary Material.

Lemma 2.2. *Under assumption 1, for a fixed N, we have:*

$$T^{1/2}(\bar{\boldsymbol{Z}} - \boldsymbol{\psi}) \xrightarrow{d} N_{Nm}(\boldsymbol{0}_{Nm}, \boldsymbol{\Omega}_{\boldsymbol{Z}}) \tag{2.11}$$

where $\boldsymbol{\psi} = \mathbb{E}[\bar{\boldsymbol{Z}}]$ and $\boldsymbol{\Omega}_{\boldsymbol{Z}} = \lim_{T\to\infty} var(T^{1/2}\bar{\boldsymbol{Z}})$.

The covariance matrix $\boldsymbol{\Omega}_{\boldsymbol{Z}}$ is estimable via bootstrap. Next, I present a result regarding the asymptotic distribution of the $T_{\max,\mathcal{M}}$ statistic, ensuring the asymptotic validity of the bootstrap procedure. Let $\boldsymbol{\omega}_i^2$ denote the i^{th} diagonal block of $\boldsymbol{\Omega}_{\boldsymbol{Z}}$, and $\hat{\boldsymbol{\omega}}^2_{i,T} \equiv \widehat{\mathrm{var}}\left(T^{1/2}\bar{\boldsymbol{d}}_{i\cdot}\right) = T\widehat{\mathrm{var}}\left(\bar{\boldsymbol{d}}_{i\cdot}\right) \xrightarrow{p} \boldsymbol{\omega}_i^2$, and let $D \equiv \mathrm{diag}\left(\boldsymbol{\omega}_1^2, \ldots, \boldsymbol{\omega}_m^2\right)$. Consider the Nm-dimensional random variable $\boldsymbol{\xi}$, distributed according to $N_{Nm}(\boldsymbol{0}_{Nm}, \boldsymbol{\varrho})$, where $\boldsymbol{\varrho} = D^{-1/2}\boldsymbol{\Omega}_{\boldsymbol{Z}}D^{-1/2}$; as well as the distribution $\boldsymbol{F}_{\boldsymbol{\varrho}}$ of the statistic $\max_i \boldsymbol{\xi}_i'\boldsymbol{\xi}_i$. Theorem 2.3 shows that the limiting distribution of the $T_{\max,\mathcal{M}}$ statistic is $\boldsymbol{F}_{\boldsymbol{\varrho}}$. The proof can be found in the Supplementary Material.

Theorem 2.3. *Suppose that assumptions 1 and 2 hold. Then, for the statistic (2.7), we have* $T_{\max,\mathcal{M}} \xrightarrow{d} \boldsymbol{F}_{\boldsymbol{\varrho}}$ *under the null hypothesis* $H_{0,\mathcal{M}}$*, and* $T_{\max,\mathcal{M}} \xrightarrow{p} \infty$ *under the alternative hypothesis* $H_{A,\mathcal{M}}$.

Given the result from Theorem 2.3, the moving-block bootstrap can be implemented. Under these conditions, the bootstrap implementation follows as in Hansen et al. (2011); as the distribution of the $T_{\max,\mathcal{M}}$ statistic is the same as its bootstrapped equivalent $T_{b^*,\max}$. Details are available in the Supplementary Material. In the next section, two simulation designs showcase the procedure's properties.

3 Simulation Study

In this section, I consider two simulation designs, similar to Designs $I.B$ and II of Hansen et al. (2011), with the adaption that the vectors of losses within a model can admit some degree of dependence for Design $I.B$, and added empirical relevance for Design II.

3.1 Design I: Dependent Losses

In this design, I generate random dependent losses by drawing realizations from a multivariate normal distribution, which is parameterized by a covariance matrix with Kronecker product structure, to allow for both between- and within-model correlation. The block bootstrap length is set to $l = 2$, the number of bootstrap iterations to $B =$ 1,000, and the number of simulation replications to $C =$ 2,500.

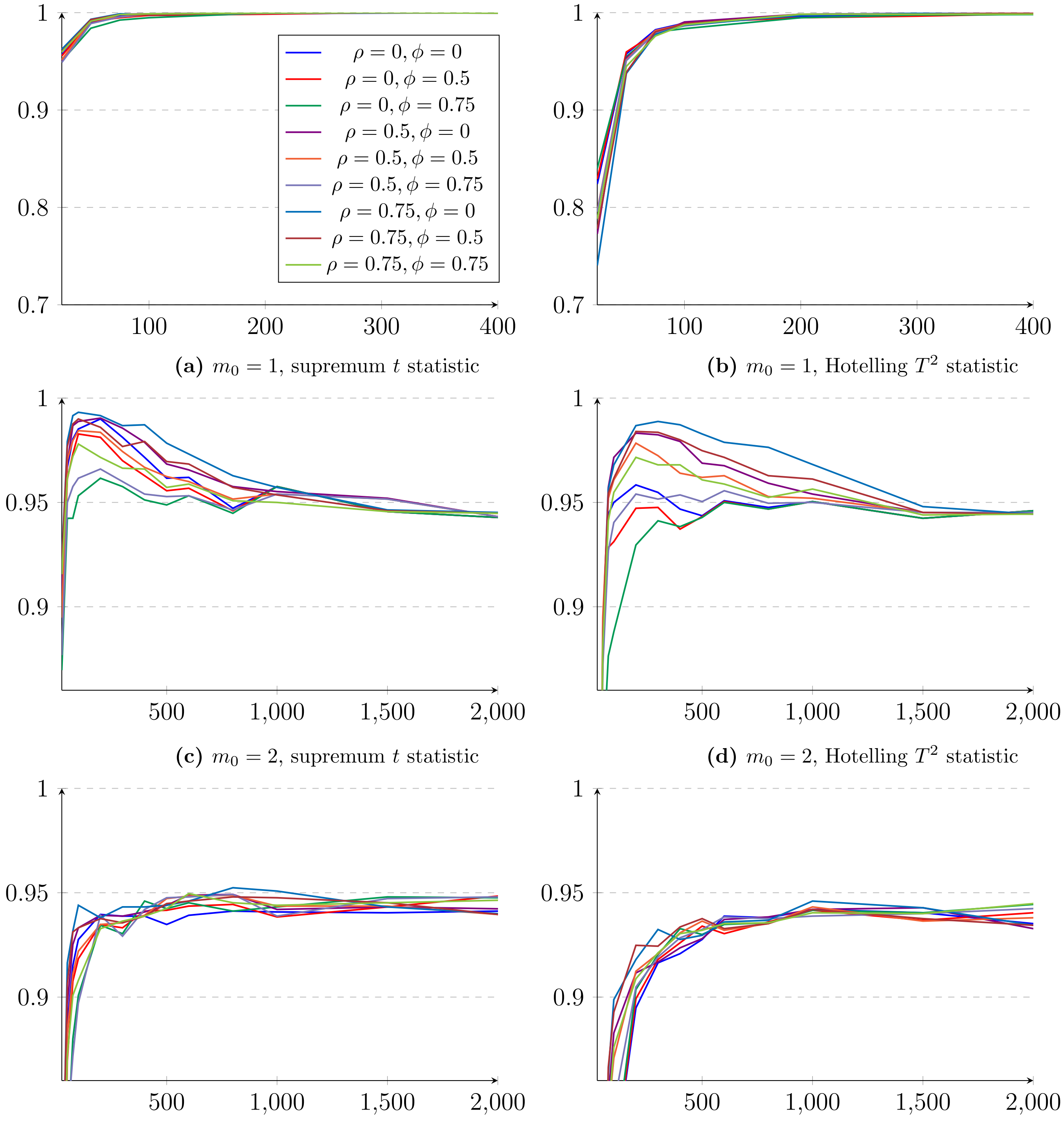
Let m_0 denote the number of best models from a set of m candidate models. This design uses a $(T \times Nm)$ matrix of losses $\boldsymbol{L} = [\boldsymbol{L}_{i_1}, \ldots, \boldsymbol{L}_{i_m}]$ drawn from $N_{Nm}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Each dependent variable in a model admits a loss with mean $\boldsymbol{\theta}$, parameterized as follows:

$$\boldsymbol{\theta} = \begin{cases} \mathbf{0}_N & \text{if the model belongs to the MCS,} \\ \boldsymbol{\iota}_N & \text{otherwise, where each element of } \boldsymbol{\iota}_N \text{ is } \iota = 1/(m - m_0). \end{cases} \tag{3.1}$$

The covariance matrix is set as $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\phi \otimes \boldsymbol{\Sigma}_\rho$, where $\boldsymbol{\Sigma}$ is parameterized so that the covariance matrix between the losses of models i and j, $\boldsymbol{L}_i$ and $\boldsymbol{L}_j$, is $\boldsymbol{\Sigma}_\phi$; and that for a given model i, the covariance between each $(T \times 1)$ loss vector l_i^n for any given dependent variable n, is $\boldsymbol{\Sigma}_\rho$. The $(n,q)^{th}$ and $(i,j)^{th}$ elements of $\boldsymbol{\Sigma}_\phi$ and $\boldsymbol{\Sigma}_\rho$ are defined as $\boldsymbol{\Sigma}_\phi(n,q) = \phi^{|n-q|}$ and $\boldsymbol{\Sigma}_\rho(i,j) = \rho^{|i-j|}$, for $n,q = 1,\dots,N$ and $i,j = 1,\dots,m$, respectively. $\boldsymbol{\Sigma}_\phi$ and $\boldsymbol{\Sigma}_\rho$ are of dimension $(N \times N)$ and $(m \times m)$, respectively. The results of the simulation for $m_0 = 1$, 2 and 5 best models, $m = 10$ candidate models, and $N = 5$ dependent variables are presented in Figures 2 and 3. Additional simulation results for $N = 10$ and $N = 50$ dependent variables are available in the Supplementary Material.[4] The top, middle, and bottom panels show the results for $m_0 = 1$, 2 and 5 best models, respectively. Figure 2 plots the frequency at which the best model is selected by the MCS procedure. This frequency reflects the ability of the procedure to include the best model(s), and is interpreted as the size property of the procedure. Figure 3 plots the average cardinality (the number of elements in the set) of the MCS. This property illustrates the ability of the procedure to eliminate the inferior models.

Overall, the procedure behaves well and delivers the expected coverage probability and number of selected models. The top panels in Figure 2 verify Corollary 1 of Hansen et al. (2011) stated in the Supplementary Material for both considered statistics, which implies that if the cardinality of the true MCS $\mathcal{M}^*$ is one, then the coverage probability $P(\mathcal{M}^* = \widehat{\mathcal{M}}^*_{1-\alpha})$ of the MCS is one in the limit. This result is achieved for sample sizes

[4] In the case where N is large relative to T, the covariance matrix is near-singular and matrix inversion is inaccurate. Those cases are omitted for the Hotelling statistic.



(a) $m_0 = 1$, supremum t statistic

(b) $m_0 = 1$, Hotelling T^2 statistic

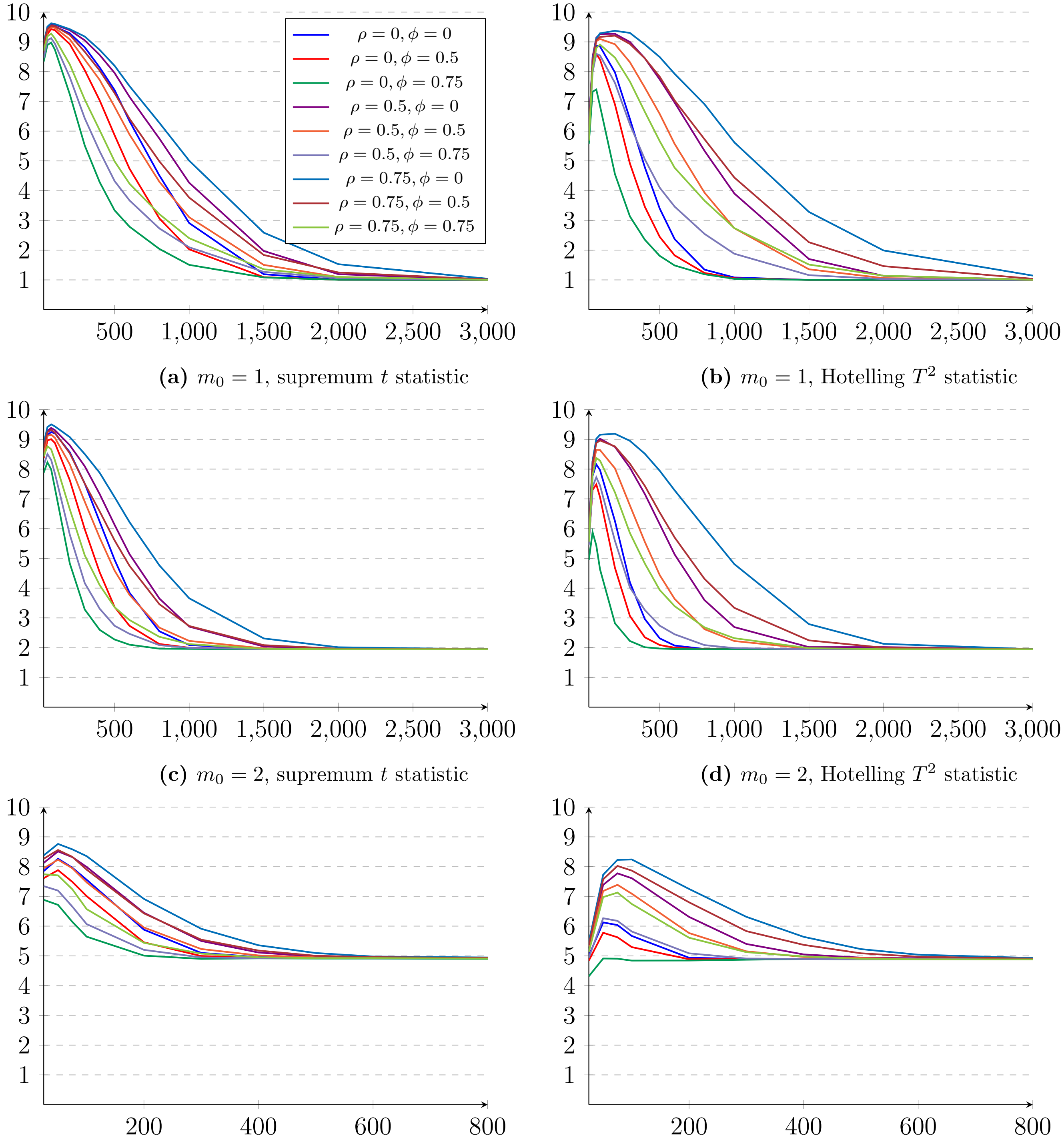
(c) $m_0 = 2$, supremum t statistic

(d) $m_0 = 2$, Hotelling T^2 statistic

(e) $m_0 = 5$, supremum t statistic

(f) $m_0 = 5$, Hotelling T^2 statistic

Figure 2: Simulation design for the supremum t and Hotelling T^2 statistics with dependent losses, $m = 10$ candidate models, $m_0 = 1$, 2 and 5 best models, $N = 5$ dependent variables, and $\alpha = 0.05$. The left panel shows the results for the supremum t statistic, and the right panel shows the results for the Hotelling T^2 statistic. The top, middle, and bottom panels show the results for $1, 2$ and 5 best models, respectively. In each panel, the vertical axis shows the frequency at which the best model is included in the estimated model confidence set $\widehat{\mathcal{M}}^*_{1-\alpha}$, and the horizontal axis shows the sample size.



(a) $m_0 = 1$, supremum t statistic

(b) $m_0 = 1$, Hotelling T^2 statistic

(c) $m_0 = 2$, supremum t statistic

(d) $m_0 = 2$, Hotelling T^2 statistic

(e) $m_0 = 5$, supremum t statistic

(f) $m_0 = 5$, Hotelling T^2 statistic

Figure 3: Simulation design for the supremum t and Hotelling T^2 statistics with dependent losses, $m = 10$ candidate models, $m_0 = 1$, 2 and 5 best models, $N = 5$ dependent variables, and $\alpha = 0.05$. The left panel shows the results for the supremum t statistic, and the right panel shows the results for the Hotelling T^2 statistic. The top, middle, and bottom panels show the results for $1, 2$ and 5 best models, respectively. In each panel, the vertical axis shows the average cardinality of the estimated model confidence set $\widehat{\mathcal{M}}^*_{1-\alpha}$, and the horizontal axis shows the sample size.

greater than 600 for the sup t statistic, and greater than 1,000 for the T^2 statistic. In conjunction with the top panels of Figure 2, the top panels of Figure 3 show that not only the procedure selects the best model in the MCS asymptotically, but only the best model, with probability one. For parameterizations where there exists more than one best model (the middle and bottom panels in Figures 2 and 3), the frequency at which the best models are included in the MCS rapidly reaches the desired $(1 - \alpha)$ coverage probability, and even exceeds that threshold for small sample sizes when $m_0 = 2$. The MCS is an asymptotic procedure, and is slightly conservative for small sample sizes in this case. When there are five best models, inclusion frequency with the sup t statistic reaches 95% coverage after 1,000 observations - faster than with the T^2 statistic. This happens at the expense of power, especially for low values of the within-model correlation parameter ρ. For $\rho = 0$, the T^2 statistic selects close to five models for sample sizes as low as 200, where the sup t statistic requires at least $T = 500$. When the within-model correlation parameter ρ is close to zero, the off-diagonal elements of Σ_{ij} are small, and in reverse, the off-diagonal elements of Σ_{ij}^{-1} are large, and the statistic is large. Consequently, when ρ is closer to zero, the statistic rejects more, and the number of included models is lower.

Figure 3 captures the power properties of the MCS procedure. All else equal, the average number of selected models decreases for greater values of the between-model correlation parameter ϕ. This additional power reflects the information captured by ϕ, making it easier for the procedure to reject incorrect models. However, greater values of the within-model correlation parameter ρ increase the average number of selected models, making it harder to reject incorrect models, in contrast with the results of Hansen et al. (2011). This pattern remains for $m_0 = 1$, 2 and 5 best models and holds true for both statistics. For the size property, there is no persistent pattern with respect to the different values of the correlation parameters.

3.2 Design II: Regression Models

This simulation design is based on empirically calibrated asset pricing factor models of the form:

$$R_{n,t} - r_t^f = \alpha_n + f_{1,t}\beta_{1,n} + \ldots + f_{K,t}\beta_{K,n} + \epsilon_{n,t}, \qquad \text{for } t = 1, \ldots, T, \text{ and } n = 1, \ldots, N. \quad (3.2)$$

The dependent variables $Y_{n,t} = R_{n,t} - r_t^f$ and the regressors are the excess returns of the portfolios and factors described in Section 4. I consider three data-generating processes (DGP) constructed from the regressors of the Fama and French (2015) five-factor model. The first DGP consists of a constant term and the market premium factor MKT_t. The second DGP is the Fama and French (1993) three-factor model, which is composed of a constant term, the market premium factor MKT_t, the size factor SMB_t, and the value factor HML_t. The third DGP is the model which includes all the regressors in the Fama and French (1993) three-factor model, along with the profitability RMW_t and the investment CMA_t factors, *i.e.* the Fama and French (2015) five-factor model. I consider a sample with 492 monthly observations, from July 1972 to June 2013. I estimate different combinations of the regressors of the Fama and French (2015) model, starting from a model containing a constant term only, and progressively adding regressors, and additional combinations which exclude the market factor MKT_t. In total, I consider 10 different combinations which will be the set of candidate models $\mathcal{M}^0$. The simulation is conducted as follows:

1. Estimate the parameters $\alpha_n, \beta_{1,n}, \ldots, \beta_{K,n}$ and the covariance matrix $\mathbf{\Sigma}$ for each candidate model $i \in \mathcal{M}^0$.

2. Compute the fitted values $\widehat{Y}_{n,t}$ for each candidate model $i \in \mathcal{M}^0$.

3. Draw errors $U_{n,t}$ from a multivariate normal distribution with covariance $\widehat{\mathbf{\Sigma}}$ for each candidate model $i \in \mathcal{M}^0$.

4. Compute the simulated fitted values $\widehat{Y}^{s}_{n,t} = \widehat{\alpha}_n + f_{1,t}\widehat{\beta}_{1,n} + \ldots + f_{K,t}\widehat{\beta}_{K,n} + U_{n,t}$ for each candidate model $i \in \mathcal{M}^0$.

5. Compute $L_{n,t} = (\widehat{Y}^{DGP}_{n,t} - \widehat{Y}^{s}_{n,t})^2$, the quadratic loss between the DGP's fitted values and the other candidates' simulated values.

6. Repeat steps (1) to (5) C times.

7. Compute the frequency at which each model is selected in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$.

The frequencies for the sup t and the Hotelling T^2 statistics are presented in Table 1. Shading indicates the DGP. The block bootstrap length is $l = 12$, and additional results for $l = 3$ and 24 are available in the Supplementary Material. It appears costly to omit the market factor: in all cases, the models that exclude the market factor are never selected in the MCS, even when other regressors are added. Inclusion frequencies align between the two statistics in this respect. The size and value factors show some redundancy, and are included at a lower frequency than the other factors, even when the DGP is a model that includes them, like the Fama and French (2015) model. Moreover, the inclusion frequencies tend to decrease as the number of test portfolios increases for the Hotelling statistic, whereas the frequencies computed using the supremum t statistic are largely unaffected. Results pertaining to the power properties of the procedure are in line with that of Design I, in that the Hotelling T^2 statistic always includes less models in the MCS than the supremum t statistic. Also, a larger number of test portfolios helps in eliminating inferior models.

Table 1: Frequency at which each candidate model is included in the 95% MCS.

		sup t statistic			Hotelling T^2 statistic		
Portfolios	Candidate Models	DGP 1	DGP 2	DGP 3	DGP 1	DGP 2	DGP 3
R_{12IND}	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_1$	0.9996	0.9992	0.9988	0.9964	0.9976	0.9972
	$1, X_1, X_2$	0.9856	0.0000	0.0000	0.9356	0.0000	0.0000
	$1, X_1, X_2, X_3$	0.9852	0.9876	0.0000	0.9248	0.9532	0.0000
	$1, X_1, X_2, X_3, X_4$	0.9796	0.9784	0.7868	0.8748	0.8996	0.6568
	$1, X_1, X_2, X_3, X_4, X_5$	0.9788	0.9756	0.9636	0.8848	0.9032	0.9196
	$1, X_2$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3, X_4$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3, X_4, X_5$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Average #MCS	4.9288	3.9408	2.7492	4.6164	3.7536	2.5736
$R_{25ME/BEME}$	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_1$	0.9996	1.0000	1.0000	0.9432	0.9936	0.9972
	$1, X_1, X_2$	0.9968	0.0000	0.0000	0.7508	0.0000	0.0000
	$1, X_1, X_2, X_3$	0.9852	0.9844	0.0872	0.5176	0.7536	0.0004
	$1, X_1, X_2, X_3, X_4$	0.9884	0.9684	0.9288	0.4816	0.6008	0.5392
	$1, X_1, X_2, X_3, X_4, X_5$	0.9868	0.9716	0.9860	0.4736	0.5812	0.8600
	$1, X_2$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3, X_4$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3, X_4, X_5$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Average #MCS	4.9568	3.9244	3.0020	3.1668	2.9292	2.3968
R_{49IND}	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_1$	0.9964	0.9936	0.9912	0.6060	0.6912	0.7748
	$1, X_1, X_2$	0.9724	0.0016	0.0000	0.2556	0.0016	0.0000
	$1, X_1, X_2, X_3$	0.9784	0.8988	0.0188	0.2420	0.3276	0.0064
	$1, X_1, X_2, X_3, X_4$	0.9768	0.9080	0.8364	0.2564	0.3236	0.3412
	$1, X_1, X_2, X_3, X_4, X_5$	0.9788	0.8992	0.8620	0.2404	0.3188	0.5348
	$1, X_2$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3, X_4$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$1, X_2, X_3, X_4, X_5$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Average #MCS	4.9028	3.7012	2.7084	1.6004	1.6628	1.6572

Notes: #MCS denotes the cardinality of the estimated model confidence set $\widehat{\mathcal{M}}^*_{1-\alpha}$. The block bootstrap length is $l = 12$. Shading denotes the DGP for each experiment.

In the next section, I propose to test a large number of asset pricing factors models that have received support in the literature using the MCS procedure.

4 Model Confidence Sets for Asset Pricing Models

In this section, I apply the MCS procedure to a set of multivariate asset pricing factor models. This analysis highlights three key empirical facts. First, there exists striking differences between the in-sample (IS) and out-of-sample (OOS) results. The IS MCS always contains more models than the OOS MCS. Second, the IS tests based on the Hotelling statistic eliminate more inferior models than the sup t ones. Third, OOS results depend largely on the considered test portfolios.

Recently, numerous approaches have been developed to compare asset pricing models; namely using Sharpe ratio-based statistics [Fama and French (2018), Kan et al. (2019), Barillas et al. (2020)], machine learning methods [Feng et al. (2020), Gu et al. (2020), Kozak et al. (2020)], mispricing distance measures [Gospodinov et al. (2013), Gagliardini and Ronchetti (2020), Zhang et al. (2021)], and Bayesian methods [Barillas and Shanken (2018), Bryzgalova et al. (2023)].[5] While these newly introduced techniques may be sensitive to (*i*) distributional assumptions, (*ii*) hyperparameter tuning, and (*iii*) factor tradability assumptions, the MCS procedure provides a flexible framework for the evaluation of asset pricing models, which can accommodate tradable or non-tradable factors for a variety of loss functions, and aims to control coverage, *i.e.*, the probability of including the true unknown set of superior models. In contrast to tests of superior predictive ability [see Hansen (2005), Giacomini and White (2006), and Li et al. (2022)], the MCS procedure uses an equal predictive ability test, implying that choosing a benchmark model to compare against is not required. Moreover, the MCS procedure remains agnostic to the modeling process, and can be viewed as a model-free approach: the researcher can receive series of candidate predictions, and compute the MCS to obtain the set of modeling approaches

[5] See Weigand (2019) and Giglio et al. (2022) for a review of machine learning applications in empirical asset pricing, and Barillas and Shanken (2017), Hou et al. (2018), He et al. (2022) and Pukthuanthong et al. (2018) for factor models comparison.

with equal predictive ability.

The considered factor models depict the relationship between expected returns of a portfolio R_n over the risk-free rate r^f, and the risk factors f:

$$E[R_n - r^f] = \beta_n E[f], \tag{4.1}$$

where β_n is a vector of factor loadings (or sensitivities) for portfolio n defined as $\beta_n = [\beta_{1,n}\, \beta_{2,n} \ldots \beta_{K,n}]$, and f is a vector of factors defined as $f = [f_1\, f_2 \ldots f_K]'$. The portfolio returns are commonly referred to as the *test portfolios*. The expected returns model states that the expected excess returns on the test portfolios are proportional to the expected returns on the factors: investors must be compensated for their exposure to risk factors based on the corresponding loadings β_n. The loadings can be estimated in the following regression:

$$R_{n,t} - r^f_t = \alpha_n + f_{1,t}\beta_{1,n} + \ldots + f_{K,t}\beta_{K,n} + \epsilon_{n,t}, \qquad \text{for } t = 1, \ldots, T, \text{ and } n = 1, \ldots, N, \tag{4.2}$$

where $R_{n,t}$, r^f_t, and $f_{k,t}$ are the time-t counterpart of the variables in equation (4.1), and $\epsilon_{n,t}$ is the error term associated with portfolio n at time t. When dealing with traded factors, the loadings are interpreted as portfolio weights. Since the factors are identical across the N equations and I allow for correlations across portfolios, estimating the multivariate regressions in equation (4.2) is equivalent to estimating a system of seemingly unrelated equations (SURE), which is performed via ordinary least squares (OLS). For the test portfolios, I use the Fama and French research portfolios available on Professor French's website.[6] The return series are 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 the market

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

equity (size) and the book-to-market value ratio ($N = 25$), and the industry ($N = 12$ and $N = 49$), as suggested by Lewellen et al. (2010). The portfolios formed on size and book-to-market are the intersection of five portfolios formed on size and five portfolios formed on book-to-market (B/M). The industry portfolios are sorted according to the industry that the issuing firm falls under using the Compustat Standard Industrial Classification (SIC) codes for the previous fiscal year, or the Center for Research in Security Prices (CRSP) code if the latter is unavailable. Summary statistics and industry classifications for the test portfolios are presented in the Supplementary Material. I consider the time-t quadratic loss:

$$\boldsymbol{L}_{i,t} = [(R_{1,t} - \widehat{R}^{i}_{1,t})^2, \ldots, (R_{N,t} - \widehat{R}^{i}_{N,t})^2], \tag{4.3}$$

where $\widehat{R}^{i}_{n,t}$ is the return of portfolio n at time t, estimated by model i. To compute each statistic, I use a moving-block bootstrap with block length $l = 12$. The bootstrap procedure is detailed in the Supplementary Material.

I consider the 14 asset pricing models presented in Table 2, over a time period from July 1972 to June 2013, totaling 492 observations.[7] Detailed factor descriptions are available in the Supplementary Material. As seen in the simulations, this sample size is often sufficient to achieve $(1 - \alpha)$ probability coverage.

[7] The integration of a benchmark model through cross-validation would be a useful question to explore and a worthy research objective.

Table 2: Candidate factor models.

Category	Example risk factor	Reference	Abbreviation
Market	Market risk premium	Sharpe (1964)	CAPM
Accounting	Size premium	Fama and French (1993)	FF3
Accounting	Value premium	Asness and Frazzini (2013)	AF
Accounting	Profitability premium	Novy-Marx (2013)	NM
Accounting	Investment premium	Fama and French (2015)	FF5
Accounting	Quality premium	Asness et al. (2019)	AFP
Liquidity	Aggregate market liquidity	Pástor and Stambaugh (2003)	PS
Liquidity	Zero-volume trading days	Liu (2006)	LIU
Momentum	Momentum	Carhart (1997)	CAR
Momentum	Modified value premium	Asness et al. (2013)	AMP
Momentum	Trend	Han et al. (2016)	HZZ
Intermediary	Intermediary capital ratio	He et al. (2017)	HKM
Mispricing	Firm management mispricing	Stambaugh and Yuan (2017)	SY
Behavioral	Post earnings anomalies	Daniel et al. (2020)	DHS

Tables 3 and 4 show the results of the MCS procedure for the sup t and the T^2 statistics, for in- and out-of-sample tests. A MCS p-value greater than 5% indicates inclusion in the 95% model confidence set, which is denoted by the grey shading.

For in-sample tests, superior models are generally easier to distinguish with the T^2 statistic than with the supremum t statistic. When using the sup t statistic, seven candidate models remain with the industry portfolios, and five remain with the size-sorted portfolios. In contrast, when using the T^2 statistic, five models remain for the 12-industry portfolios, and only one candidate model remains for both the size-sorted and the 49-industry portfolios: the Fama and French (2015) five-factor model. The "last" remaining model - the model with the largest MCS p-value - is always the same, regardless of the statistic: the Fama and French (2015) five-factor model, for all types of portfolios considered.

The use of the 25 size- and B/M-sorted portfolios as test portfolios has been the subject of scrutiny, due to their strong correlation with the size and the value factors; see Lewellen et al. (2010). Following the recommendations in the literature, I also include 12 and 49 industry-sorted portfolios in my analysis. The results are surprising. First, the MCS under

Table 3: MCS p-values for the candidate factor models, using the quadratic loss function, for in-sample and out-of-sample tests, at the monthly frequency from July 1972 to June 2013. For out-of-sample tests, the horizon h is 12, 24, and 60 months. R_{12IND}, and $R_{ME/BEME}$ denote the portfolio returns for 12 industry-sorted portfolios, and 25 size- and book-to-market-sorted portfolios, respectively. Shading denotes the inclusion in the 95% MCS. The block bootstrap length is $l = 12$ months. The number of bootstrap iterations is set to B = 1,000.

		In-sample		Out-of-sample					
		07/1972 - 06/2013		$h = 12$		$h = 24$		$h = 60$	
	Model	sup t	T^2	sup t	T^2	sup t	T^2	sup t	T^2
R_{12IND}	CAPM	0.0340	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	FF3	0.1620	0.4090	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	CAR	0.1620	0.0640	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	PS	0.1620	0.4090	0.0000	0.0000	1.0000	1.0000	0.0050	0.0000
	FF5	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0050	0.0000
	HKM	0.0340	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	AF	0.1620	0.0020	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
	NM	0.0340	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	HZZ	0.0340	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	SY	0.1620	0.0020	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	LIU	0.0460	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	DHS	0.0460	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	AMP	0.0460	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	AFP	0.1620	0.0640	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
$R_{ME/BEME}$	CAPM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	FF3	0.4960	0.0040	0.0000	0.0000	0.0000	0.0000	0.0150	0.0000
	CAR	0.4960	0.0040	0.0000	0.0000	0.0000	0.0000	0.0150	0.0000
	PS	0.4960	0.0040	0.0000	0.0000	0.0000	0.0000	0.0050	0.0000
	FF5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	HKM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AF	0.0080	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	NM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	HZZ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	SY	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LIU	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	DHS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AMP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AFP	0.4960	0.0040	0.0000	0.0000	0.0000	0.0000	0.0150	0.0000

Table 4: MCS p-values for the candidate factor models, using the quadratic loss function, for in-sample and out-of-sample tests, at the monthly frequency from July 1972 to June 2013. For out-of-sample tests, the horizon h is 12, 24, and 60 months. R_{49IND} denotes the portfolio returns for 49 industry-sorted portfolios. Shading denotes the inclusion in the 95% MCS. The block bootstrap length is $l = 12$ months. The number of bootstrap iterations is set to $B = 1{,}000$.

		In-sample		Out-of-sample					
		07/1972 - 06/2013		$h = 12$		$h = 24$		$h = 60$	
	Model	sup t	T^2	sup t	T^2	sup t	T^2	sup t	T^2
R_{49IND}	CAPM	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	FF3	0.3600	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000
	CAR	0.3600	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	PS	0.3600	0.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.0000
	FF5	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	HKM	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AF	0.3230	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	NM	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	HZZ	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	SY	0.3600	0.0000	0.0000	0.0000	0.0000	0.0000	0.0150	0.0000
	LIU	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	DHS	0.0050	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AMP	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AFP	0.3600	0.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000

the size and B/M portfolios and the industry portfolios are extremely similar with the sup t statistic: the MCS always includes the Fama and French (1993), Carhart (1997), Pástor and Stambaugh (2003), Fama and French (2015), Asness et al. (2019) models. In addition, when using the industry portfolios as test portfolios, the MCS also includes the Asness and Frazzini (2013) and Stambaugh and Yuan (2017) models. The recent work by Barillas and Shanken (2017, 2018) argues that with the assumption of factor tradability, the test assets become irrelevant to the analysis. This assumption, however, is not imposed in the MCS framework.

These two models are the only models that include the size factor without the B/M factor. Note that the Asness and Frazzini (2013) contains a modified B/M factor, albeit it does not explain the variation in the size and B/M returns sufficiently well to warrant inclusion in the MCS. Second, only models containing the size factor are included in the MCS when using industry portfolios. Even models that exclude the market factor (Pástor and Stambaugh (2003)) or the B/M factor are selected by the MCS procedure, as long as they contain the size factor. Third, the groundbreaking capital asset pricing model (CAPM) of Sharpe (1964) is never included in the MCS. While empirical evidence on the usefulness of the market factor is mixed, the MCS results complement the recent approach of Harvey and Liu (2021) in the following sense. Using a bootstrap approach to control for multiple testing, they find that the theoretically-grounded market factor is the single most dominant factor, using individual stocks instead of portfolios as test assets. This paper comes to a similar conclusion, albeit when the market factor is combined with other pertinent risk factors. With the exception of the Pástor and Stambaugh (2003), all considered candidate models that are included in the MCS contain the market factor, suggesting that both the market factor and additional factors are relevant, nevertheless the market factor alone is insufficient.

The picture becomes clearer in out-of-sample results: the MCS contains only one model for both statistics. The selected model depends (i) on the test portfolios; and (ii) on the forecasting horizon, but both statistics always select the same model.

With respect to the choice of the sup t or the T^2 statistic, each statistic offers different insights. While both of them account for correlation via the bootstrap procedure, as shown in Harvey and Liu (2021), the sup t statistic selects the portfolio whose t statistic rejects the null of equal predictability the most, whereas the T^2 statistic averages out and weighs the squared means, according to the elements in the covariance matrix Σ_{ij}. For instance, for the Asness and Frazzini (2013) and Stambaugh and Yuan (2017) models, the largest statistics across portfolios are sufficiently large to reject $H_{0,\mathcal{M}}$ with 12-industry portfolios, but fail to reject $H_{0,\mathcal{M}}$ when considering all the portfolios jointly, leading to both models being excluded from the MCS when using the T^2 statistic. Different model inclusions in the MCS can be explained by power differences for varying within- and between-model correlations, as seen in the simulation results. When the within-model correlation ρ is small, the test using the T^2 statistic is more powerful than with the sup t statistic.

Temporal instabilities in the estimated coefficients are a well-documented fact in the context of beta pricing models. To address these concerns, I divide the original sample period into 10-year sub-periods so that the variation in the coefficients is small enough to offer short-term stability.[8] The results of in-sample tests for the four sub-periods are available in the Supplementary Material. Using the sup t statistic, the Fama and French (2015) model dominates the 1992 to 2002 period for the industry portfolios. The results for other time periods are a testimony of the temporal instabilities, as no single model stands out. Moreover, for the period surrounding the 2007-2008 financial crisis, it becomes difficult to distinguish the models: between five and 14 models are selected in the MCS. Using the Hotelling T^2 statistic, the results are clear-cut. At most one model remains for every period, and the Fama and French (2015) model emerges as the winner in most time periods and most portfolio types. Notably, the Carhart (1997) and the Stambaugh and Yuan (2017) are also occasionally selected. The behavioral model of Daniel et al. (2020) is always excluded, although this exclusion does not imply a rejection of all behavior-based models. The MCS results do not necessarily favour more parsimonious models, despite evidence to the contrary in recent work; see Bryzgalova et al. (2023).

5 Conclusion

This paper provides a multivariate extension of the model confidence set procedure originally proposed by Hansen et al. (2011) for univariate models, and proposes two statistics to test equal predictive ability: a supremum-type t statistic and a Hotelling-type T^2 statistic. Both statistics summarize the information contained in the systems of equations to test equal predictive ability. The extensive simulation study showcases the asymptotic size

[8] See Fama and MacBeth (1973), Roll and Ross (1980), and Gibbons (1982) on the use of short time periods and Gagliardini et al. (2016) for an alternative approach via a time-varying parameter model.

and power properties of the procedure. The procedure is adequately sized, even in small samples. In many cases, the desired coverage probability is achieved in samples as small as $T = 500$, and the procedure can eliminate all inferior models around $T = 800$ when there are a large number of good models. Simulations also show that one of the key properties of the MCS procedure, namely that the estimated MCS converges to the true MCS in probability when the latter is a singleton, holds true in the multivariate case.

The empirical analysis answers several outstanding questions with regards to the factor proliferation problems encountered in asset pricing. Namely, how do models featuring recently discovered factors compare, and does a particular model stand out? I apply the MCS procedure to a set of 14 candidate models and I find that the prominent Fama and French (2015) five-factor model is never included in the MCS for out-of-sample tests with the 49 industry portfolios. This finding is persistent for both the supremum t and the Hotelling T^2 statistics. For in-sample predictions, the candidate models are often indistinguishable in their capacity to explain expected returns under the supremum statistic. To address concerns relating to temporal instabilities, the sample is divided into 10-year periods, and the MCS is performed in-sample over four different periods. Although the selected models change often, the Fama and French (2015) model is the only model selected for the 1992 to 2002 period for the industry portfolios.

A confidence set approach provides valuable insights and has significant strengths. First, the candidate models do not need to follow a certain structure, *e.g.* with respect to nesting or factor tradibility; second, a baseline model is not required; and third, the model confidence set procedure allows models to be viewed as statistically equal. In the context of beta pricing models, the model confidence set procedure also allows us to establish the significance of models, as opposed to the marginal contributions of new factors.

References

Asness, C. and A. Frazzini (2013): "The devil in HMLs details," *The Journal of Portfolio Management*, 39, 49–68.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2019): "Quality minus junk," *Review of Accounting Studies*, 24, 34–112.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013): "Value and momentum everywhere," *The Journal of Finance*, 68, 929–985.

Barillas, F., R. Kan, C. Robotti, and J. Shanken (2020): "Model comparison with sharpe ratios," *Journal of Financial and Quantitative Analysis*, 55, 1840–1874.

Barillas, F. and J. Shanken (2017): "Which alpha?" *The Review of Financial Studies*, 30, 1316–1338.

——— (2018): "Comparing asset pricing models," *The Journal of Finance*, 73, 715–754.

Beaulieu, M.-C., J.-M. Dufour, L. Khalaf, and O. Melin (2023): "Identification-robust beta pricing, spanning, mimicking portfolios, and the benchmark neutrality of catastrophe bonds," *Journal of Econometrics*, 236, 105464.

Bergamelli, M., A. Bianchi, L. Khalaf, and G. Urga (2019): "Combining p-values to test for multiple structural breaks in cointegrated regressions," *Journal of Econometrics*, 211, 461–482.

Bryzgalova, S., J. Huang, and C. Julliard (2023): "Bayesian solutions for the factor zoo: We just ran two quadrillion models," *The Journal of Finance*, 78, 487–557.

Carhart, M. M. (1997): "On persistence in mutual fund performance," *The Journal of Finance*, 52, 57–82.

Daniel, K., D. Hirshleifer, and L. Sun (2020): “Short-and long-horizon behavioral factors,” *The Review of Financial Studies*, 33, 1673–1736.

Dufour, J.-M., L. Khalaf, and M. Voia (2015): “Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability,” *Communications in Statistics-Simulation and Computation*, 44, 2329–2347.

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.

——— (2018): “Choosing factors,” *Journal of Financial Economics*, 128, 234–252.

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 (2020): “Taming the factor zoo: A test of new factors,” *The Journal of Finance*, 75, 1327–1370.

Fisher, R. (1932): “Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925,” .

Gagliardini, P., E. Ossola, and O. Scaillet (2016): “Time-varying risk premium in large cross-sectional equity data sets,” *Econometrica*, 84, 985–1046.

Gagliardini, P. and D. Ronchetti (2020): “Comparing asset pricing models by the conditional Hansen-Jagannathan distance,” *Journal of Financial Econometrics*, 18, 333–394.

Giacomini, R. and H. White (2006): “Tests of conditional predictive ability,” *Econometrica*, 74, 1545–1578.

Gibbons, M. R. (1982): "Multivariate tests of financial models: A new approach," *Journal of Financial Economics*, 10, 3–27.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989): "A test of the efficiency of a given portfolio," *Econometrica*, 1121–1152.

Giglio, S., B. Kelly, and D. Xiu (2022): "Factor models, machine learning, and asset pricing," *Annual Review of Financial Economics*, 14, 337–368.

Giglio, S., Y. Liao, and D. Xiu (2021): "Thousands of alpha tests," *The Review of Financial Studies*, 34, 3456–3496.

Gonçalves, S. and H. White (2005): "Bootstrap standard error estimates for linear regression," *Journal of the American Statistical Association*, 100, 970–979.

Gospodinov, N., R. Kan, and C. Robotti (2013): "Chi-squared tests for evaluation and comparison of asset pricing models," *Journal of Econometrics*, 173, 108–125.

Gospodinov, N. and C. Robotti (2021): "Common pricing across asset classes: Empirical evidence revisited," *Journal of Financial Economics*, 140, 292–324.

Grønborg, N. S., A. Lunde, A. Timmermann, and R. Wermers (2021): "Picking funds with confidence," *Journal of Financial Economics*, 139, 1–28.

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

Han, Y., G. Zhou, and Y. Zhu (2016): "A trend factor: Any economic gains from using information over investment horizons?" *Journal of Financial Economics*, 122, 352–375.

Hansen, P. R. (2005): "A test for superior predictive ability," *Journal of Business & Economic Statistics*, 23, 365–380.

Hansen, P. R., A. Lunde, and J. M. Nason (2003): "Choosing the best volatility models: the model confidence set approach," *Oxford Bulletin of Economics and Statistics*, 65, 839–861.

——— (2011): "The model confidence set," *Econometrica*, 79, 453–497.

Harvey, C. R. (2017): "Presidential address: The scientific outlook in financial economics," *The Journal of Finance*, 72, 1399–1440.

Harvey, C. R. and Y. Liu (2019): "A census of the factor zoo," *Available at SSRN 3341728.*

——— (2020): "False (and missed) discoveries in financial economics," *The Journal of Finance*, 75, 2503–2553.

——— (2021): "Lucky factors," *Journal of Financial Economics.*

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

He, A., D. Huang, J. Li, and G. Zhou (2022): "Shrinking factor dimension: A reduced-rank approach," *Management science.*

He, Z., B. Kelly, and A. Manela (2017): "Intermediary asset pricing: New evidence from many asset classes," *Journal of Financial Economics*, 126, 1–35.

Hou, K., H. Mo, C. Xue, and L. Zhang (2018): "Which factors?" *Review of Finance*, 23, 1–35.

Kan, R., X. Wang, and X. Zheng (2019): "In-sample and out-of-sample sharpe ratios of multi-factor asset pricing models," *Available at SSRN 3454628.*

Kozak, S., S. Nagel, and S. Santosh (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135, 271–292.

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

Li, J., Z. Liao, and R. Quaedvlieg (2022): "Conditional superior predictive ability," *The Review of Economic Studies*, 89, 843–875.

List, J. A., A. M. Shaikh, and Y. Xu (2019): "Multiple hypothesis testing in experimental economics," *Experimental Economics*, 22, 773–793.

Liu, L. Y., A. J. Patton, and K. Sheppard (2015): "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes," *Journal of Econometrics*, 187, 293–311.

Liu, W. (2006): "A liquidity-augmented capital asset pricing model," *Journal of Financial Economics*, 82, 631–671.

Mariano, R. S. and D. Preve (2012): "Statistical tests for multiple forecast comparison," *Journal of Econometrics*, 169, 123–130.

Montiel Olea, J. L. and M. Plagborg-Møller (2019): "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, 34, 1–17.

Novy-Marx, R. (2013): "The other side of value: The gross profitability premium," *Journal of Financial Economics*, 108, 1–28.

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

Pearson, K. (1933): "On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random," *Biometrika*, 25, 379–410.

Pukthuanthong, K., R. Roll, and A. Subrahmanyam (2018): "A protocol for factor identification," *The Review of Financial Studies.*

Roll, R. and S. A. Ross (1980): "An empirical investigation of the arbitrage pricing theory," *The Journal of Finance*, 35, 1073–1103.

Romano, J. P. and E. Lehmann (2005): "Testing statistical hypotheses," .

Romano, J. P. and M. Wolf (2016): "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing," *Statistics & Probability Letters*, 113, 38–40.

Shanken, J. (1985): "Multivariate tests of the zero-beta CAPM," *Journal of financial economics*, 14, 327–348.

Sharpe, W. F. (1964): "Capital asset prices: A theory of market equilibrium under conditions of risk," *The Journal of Finance*, 19, 425–442.

Spreng, L. and G. Urga (2022): "Combining p-values for Multivariate Predictive Ability Testing," *Journal of Business & Economic Statistics*, 1–13.

Stambaugh, R. F. and Y. Yuan (2017): "Mispricing factors," *The Review of Financial Studies*, 30, 1270–1315.

Tippett, L. H. C. et al. (1931): "The methods of statistics." .

Varneskov, R. T. (2016): "Flat-top realized kernel estimation of quadratic covariation with nonsynchronous and noisy asset prices," *Journal of Business & Economic Statistics*, 34, 1–22.

Wang, L., X. Han, and X. Tong (2023): "Skilled Mutual Fund Selection: False Discovery Control Under Dependence," *Journal of Business & Economic Statistics*, 41, 578–592.

Weigand, A. (2019): "Machine learning in empirical asset pricing," *Financial Markets and Portfolio Management*, 33, 93–104.

Zhang, X., Y. Liu, K. Wu, and B. Maillet (2021): "Tradable or nontradable factors—what does the Hansen–Jagannathan distance tell us?" *International Review of Economics & Finance*, 71, 853–879.