Implementing Statistical Criteria to Select Return Forecasting Models: What Do We Learn?

Peter Bossaerts
California Institute of Technology

Pierre Hillion
INSEAD

Statistical model selection criteria provide an informed choice of the model with best external (i.e., out-of-sample) validity. Therefore they guard against overfitting ("data snooping"). We implement several model selection criteria in order to verify recent evidence of predictability in excess stock returns and to determine which variables are valuable predictors. We confirm the presence of in-sample predictability in an international stock market dataset, but discover that even the best prediction models have no out-of-sample forecasting power. The failure to detect out-of-sample predictability is not due to lack of power.

1. Introduction

Almost all validation of financial theory is based on historical datasets. Take, for instance, the theory of efficient markets. Loosely speaking, it asserts that securities returns must not be predictable from past information. Numerous studies have attempted to verify this theory, and ample evidence of predictability has been uncovered. This has led many to question the validity of the theory.

Quite reasonably, some have recently questioned the conclusiveness of such findings, pointing to the fact that they are based on repeated reevaluation of the same dataset, or, if not the same, at least datasets that cover similar time periods. For instance, Lo and MacKinlay (1990) argue that

Address corresponcence to Peter Bossaerts, HSS 228-77, California Institute of Technology, Pasadena, CA 91125, or e-mail: pbs@rioja.caltech.edu. P. Bossaerts thanks First Quadrant for financial support through a grant to the California Institute of Technology. First Quadrant also provided the data that were used in this study. The article was revised in part when the first author was at the Center for Economic Research, Tilburg University. P. Hillion thanks the Hong Kong University of Science and Technology for their hospitality while doing part of the research. Comments from Michel Dacorogna, Rob Engle, Joel Hasbrouck, Andy Lo, P. C. B. Phillips, Richard Roll, Mark Taylor, and Ken West, from two anonymous referees, and the editor (Ravi Jagannathan), as well as seminar participants at the Hong Kong University of Science and Technology, University of California San Diego, University of California Santa Barbara, the 1994 NBER Spring Conference on Asset Pricing, the 1994 Western Finance Association Meetings, and the 1995 CEPR/LIFE Conference on International Finance are gratefully acknowledged.

The Review of Financial Studies Summer 1999 Vol. 12, No. 2, pp. 405–428

the "size effect" in tests of the capital asset pricing model (CAPM) may very well be the result of an unconscious, exhaustive search for a portfolio formation criterion with the aim of rejecting the theory.

Repeated visits of the same dataset indeed lead to a problem that statisticians refer to as *model overfitting* [Lo and MacKinlay (1990) called it "data snooping"], that is, the tendency to discover spurious relationships when applying tests that are inspired by evidence from prior visits to the same dataset. There are several ways to address model overfitting. The finance literature has emphasized two approaches. First, one can attempt to collect new data, covering different time periods and/or markets [e.g., Solnik (1993)]. Second, standard test sizes can be adjusted for overfitting tendencies. These adjustments are either based on theoretical approximations such as Bonferroni bounds [Foster, Smith, and Whaley (1997)],[1] or on bootstrapping stationary time series [Sullivan, Timmermann, and White (1997)].

The two routes that the finance literature has taken to deal with model overfitting, however, do present some limitations. New, independent data are available only to a certain extent. And adjustment of standard test sizes merely help in correctly rejecting the simple null hypothesis of no relationship. It will provide little information, however, when, in addition, the empiricist is asked to discriminate between competing models under the alternative of the existence of some relationship.

In contrast, the statistics literature has long promoted *model selection criteria* to guard against overfitting. Of these, Akaike's criterion [Akaike (1974)] is probably the best known. There are many others, however, inspired by different criteria about what constitutes an optimal model (one distinguishes Bayesian and information-theoretic criteria), and with varying degrees of robustness to unit-root nonstationarities in the data.

The purpose of this article is to implement several selection criteria from the statistics literature (including our own, meant to correct some well-known small-sample biases in one of these criteria), based on popularity and on robustness to unit roots in the independent variables. The aim is to verify whether stock index returns in excess of the riskfree rate are indeed predictable, as many have recently concluded [e.g., Fama (1991), Keim and Stambaugh (1986), Campbell (1987), Breen, Glosten, and Jagannathan (1990), Brock, Lakonishok, and LeBaron (1992), Sullivan, Timmermann, and White (1997)].

Our insistence on model selection criteria that are robust to unit-root nonstationarities is motivated by the time-series properties of some candidate predictors, such as price-earnings ratios, dividend yields, lagged index levels, or even short-term interest rates. These variables are either mani-

[1] Foster, Smith, and Whaley (1997) also present simulation-based adjustments.

festly nonstationary, or, if not, their behavior is close enough to unit-root nonstationary for small-sample statistics to be affected.

We study an international sample of excess stock returns and candidate predictors which First Quadrant was kind enough to release to us. The time period nests that of another international study, Solnik (1993). Therefore, we also provide diagnostic tests that compare the two datasets (which are based on different sources).

We discover ample evidence of predictability, confirming the conclusion of studies that were not based on formal model selection criteria. Usually only a few standard predictors are retained, however. Some of these are unit-root nonstationary (e.g., dividend yield). Multiple lagged bond or stock returns are at times included, effectively generating the moving-average predictors that have become popular in professional circles lately [see also Brock, Lakonishok, and LeBaron (1992) and Sullivan, Timmermann, and White (1997)].

Formal model selection criteria guard against overfitting. The ultimate purpose is to obtain the model with the best external validity. In the context of prediction, this means that the retained model should provide good out-of-sample predictability. We test this on our dataset of international stock returns.

Overall, we find no out-of-sample predictability. More specifically, none of the models that the selection criteria chose generates significant predictive power in the 5-year period beyond the initial ("training") sample. This conclusion is based on an SUR test of the slope coefficients in out-of-sample regressions of outcomes onto predictions across the different stock markets. The failure to detect out-of-sample predictability cannot be attributed to lack of power. Schwarz's Bayesian criterion, for instance, discovers predictability in 9 of 14 markets, with an average R^2 of the retained models of 6%. Out of sample, however, none of the retained models generates significant forecasting power. Even with only nine samples of 60 months each, chances that this would occur if 6% were indeed the true R^2 are less than 1 in 333.

The poor external validity of the prediction models that formal model selection criteria chose indicates model nonstationarity: the parameters of the "best" prediction model change over time. It is an open question why this is. One potential explanation is that the "correct" prediction model is actually nonlinear, while our selection criteria chose exclusively among linear models. Still, these criteria pick the *best* linear prediction model: it is surprising that even this best forecaster does not work out of sample.

As an explanation for the findings, however, model nonstationarity lacks economic content. It begs the question as to what generates this nonstationarity. Pesaran and Timmermann (1995) also noticed that prediction performance improves if one switches models over time. They suggest that it reflects learning in the marketplace. Bossaerts (1997) investigates this possibility theoretically. He proves that evidence of predictability will disappear

entirely out of sample if the market learns on the basis of Bayesian updating rules. In other words, Bayesian learning could explain our findings.

The remainder of this article is organized as follows. The next section introduces model selection criteria. Section 3 describes the dataset. Section 4 presents the results. Section 5 discusses the power of the out-of-sample prediction tests. Section 6 concludes. There are three appendixes. They discuss technical issues and list the data sources.

2. Model Selection Criteria

Formal model selection criteria have long been considered in the statistics literature in order to select the "best" model among a set of candidate models. Statisticians realized that there is a tendency to overfit, and hence that the model that has the highest in-sample explanatory power usually does not have the highest external validity (i.e., out-of-sample fit). Several criteria were developed, starting from particular decision criteria, Bayesian or information theoretic.

We decided to pick several model selection criteria in our study of the predictability of excess stock returns. Each has its merit, and many are robust to the presence of unit roots in the candidate predictors. It is not appropriate to discuss here the advantages and shortcomings of the retained selection criteria. Suffice it to mention that all selection criteria contributed uniformly to the main conclusions of this article.

Formally, we use statistical criteria and T observations to select among K linear models that predict the market's excess return, r_t ($t = 1, \ldots, T$). The models differ in terms of the content and dimension of the prediction vector. Let p^k denote the dimension for model k ($k = 1, \ldots, K$). The prediction vector of this model for the tth return r_t, x^k_{t-1}, is obtained by dropping all but p^k elements from the vector of all possible predictors, x_{t-1}. x_{t-1} includes an intercept as one of the predictors, as well as variables such as the short-term Treasury bill yield, etc. (We will be explicit later on.) Letting θ^k denote its coefficient vector, model k can be written as

$$r_t = \theta^{k'} x^k_{t-1} + \epsilon^k_t, \tag{1}$$

with $E[\epsilon^k_t] = 0$, $E[\epsilon^k_t x^k_{t-1}] = 0$.

In the first model, with $k = 1$, we included only the intercept. (Hence, $p^1 = 1$.) This way, selection criteria are allowed to decide in favor of *no predictability*, beyond a constant. The latter is usually interpreted as a (fixed) risk premium. This option is important. Indeed, the original goal of this study was to verify whether the evidence of return predictability would still emerge if examined with formal selection criteria.

Each selection criterion chooses among the K possible model specifications. We will use the notation k^* to denote the preferred model. Seven

model selection criteria were employed: the adjusted R^2, Akaike's information criterion [AIC; Akaike (1974)], Schwarz's criterion [a Bayesian information criterion, BIC; Schwarz (1978)], the Fisher information criterion [FIC; Wei (1992)], the posterior information criterion [PIC; Phillips and Ploberger (1996)], Rissanen's predictive least squares criterion [PLS; Rissanen (1986a)],[2] and our adjustment to correct well-known biases of the latter, PLS-MDC.

Appendix A provides formal definitions of each of these criteria. The adjusted R^2, AIC, and BIC were chosen on the basis of their popularity; FIC, PIC, PLS, and PLS-MDC were chosen because of their robustness in the face of unit-root nonstationarities.

PLS-MDC is new and hence needs to be motivated further. It is based on a technique to estimate the dimension of the state vector in Markov models, referred to as Markov dimension criterion (MDC). MDC chooses the dimension of the state vector by investigating the out-of-sample mean square prediction error of rolling regressions that are run on the basis of various subsets of past information.

In conjunction with PLS, MDC provides a correction for small-sample biases. PLS chooses models on the basis of the out-of-sample mean square prediction error of *one* rolling regression that uses *all* past observations. Rissanen (1986a) suggested this selection criterion, but observed that it is biased in small samples in favor of picking the model with the least possible variables [see also the evidence in Wei (1992)]. The underfitting is due to the noise introduced by the error in the predictions of early observations in the sample. These predictions are unreliable because they are based on very few prior observations (remember that model estimates in PLS are computed only from prior observations). The fewer parameters to be estimated, however, the lower the prediction noise of those early observations. Because of the lower noise level, PLS tends to prefer models with fewer parameters, that is, with less explanatory variables.

In PLS-MDC, we consider the performance of the same models where parameter estimates are not only based on all previous observations, but on different subsamples as well, where we drop observations that reach a certain age. In other words, while PLS is based on expanding-window estimation, PLS-MDC also considers estimates based on windows of fixed size. PLS-MDC effectively penalizes models where excluded variables are still heavily correlated with future prediction errors, indicating that the prediction vector was chosen to be too small. Since a formal discussion of PLS-MDC distracts from the main points of this article, it is delegated to an appendix. The interested reader can consult Appendix B.

[2] PLS is based on Rissanen's earlier idea of minimum descriptive length [Rissanen (1986b)]; see also Kavalieris (1989).

As mentioned before, the purpose of statistical model selection criteria is to avoid overfitting. The model specification that fits the data best (minimum in-sample forecast error) is not necessarily chosen. In contrast, the retained model will have maximum external validity. In our context of return prediction, this means that the preferred model will have best out-of-sample forecasting performance.

We decided to verify the external validity of formal model selection whenever $k^* > 1$. All models with index (k) larger than one contain at least one nontrivial forecasting variable. When one of them is chosen, the model selection criterion clearly supports predictability. To assess the external validity of such a conclusion, we ran a test on a sample that postdated the sample on which we based the model choice. To avoid confusion we will refer to the original sample on which model choice was based as the *training sample*. The sample that was used to check for external validity will be called the *testing sample*. The latter has size n, and its elements are indexed $t = T + 1, \ldots, T + n$.

External validation is investigated by projecting the market's excess return, r_t, onto our forecast, z_{t-1}. z_{t-1} is obtained from model k^*, as follows:

$$z_{t-1} = \hat{\theta}_{t-1}^{k^*\prime} x_{t-1}^{k^*}, \tag{2}$$

where $\hat{\theta}_{t-1}^{k^*}$ is the OLS estimate of θ^{k^*}, based on the pairs $(r_\tau, x_{\tau-1}^{k^*})$, observed over $\tau = 1, \ldots, t-1$. We estimate the slope coefficient in

$$r_t = \alpha + \beta z_{t-1} + v_t \tag{3}$$

from the observations indexed $t = T + 1, \ldots, T + n$. We use the OLS estimate of β to compute a standard t-ratio, and refer to the standard normal distribution to determine p-levels. External validity is confirmed if p-levels are low, say, below 0.05.

Notice that we did not adjust the t-ratio for error in the estimation of the parameters of the prediction model. Asymptotically, such an adjustment is not necessary, because we force the precision of the parameter estimates to increase as we advance through the testing sample. See also West (1996). Unfortunately, small-sample corrections are not available and would be complicated by the unit-root nature of some of the predictors.

3. The Data

We investigated the predictability of the 1-month local-currency excess stock return for 14 countries.[3] Our forecasts of the market's excess return is based on (a subset of) the following predictors: a January dummy; the

[3] Results can be expected to be different if excess stock returns are converted to a common currency, in which case returns from speculation in the foreign exchange market determine the outcome as well. See, for example, Ferson and Harvey (1993).

monthly excess stock return, lagged once and twice; a monthly bond excess return, lagged once and twice;[4] the yield-to-maturity on a representative Treasury bond; the stock market's price level; the yield-to-maturity on a 3-month Treasury bill (also used to compute excess stock and bond returns); the stock market's dividend yield; and the stock market's price-earnings ratio. None of these predictors use information that would not have been available at the moment that future excess stock returns were predicted. Appendix C provides a list of the data sources.

Table 1 displays descriptive statistics of the monthly excess stock returns and the predictors. The lengths of the samples ($T + n$) run from a minimum of 122 observations (Norway) to a maximum of 471 observations (United States). Returns are expressed in percent per month; yields in percent per year. Table 1 also lists the beginning of the estimation sample for each country ($t = 1$). Our choice was based on data availability, and hence was purely incidental.

To get an idea of the amount of predictability that obtains in this dataset using in-sample statistical analysis, Table 2 lists the results from a regression of monthly excess stock returns (now *not* expressed in percentage terms) onto (i) an intercept, (ii) a January dummy, (iii) the stock dividend yield, (iv) the short-term interest rate, and (v) the long-term bond yield. Solnik (1993) also reports these estimates, for a similar dataset, but a shorter time period. The results of regressions like the ones reported in Table 2 match those of Solnik when run over his subperiod only (1/71-8/90). There is one exception though: our bond yield has more predictive power than his.

A closer look at Table 2 reveals the patterns that Solnik discussed at length: predictability is not uniform across stock markets [R^2s are between 2% (Germany) and 9.8% (U.K.)]; predictors vary. The well-known negative correlation between excess stock returns and contemporaneous interest rates is significant (at the 1% level) only for Belgium, France, and the United States.

An unfortunate choice of subperiod, however, may give one the wrong impression of the amount of predictability. We tried one possible subsample and generated impressive evidence of predictability. Table 2 reports the results from estimation of the same regression over the subperiod 1/71–8/80 (10 years shorter than Solnik's sampling period). They are listed under Period II. The R^2s run from a low of 4.4% (Canada) to a high of 36.4% (Belgium) or higher![5] Overall, the significance of the coefficients in the regression is overwhelming.

[4] This bond return is actually computed from the yield on a representative Treasury bond and an estimate of the duration. Hence it cannot strictly be considered to be a bond (excess) return. Nevertheless, its computation does not require future information. Therefore it is a proper predictor.

[5] Some high R^2s may be the result of lack of observations. The highest R^2, 40% (Spain), is based on a sampling period that runs from 1/78 to 8/80 only. Belgium's high R^2, however, is based on a full sample.

Table 1
Monthly averages, entire sample period (ending 5/95)

Country	$t=1$	$T+n$	r_t	$r_{b,t-1}$	$Y_{b,t-1}$	$r_{f,t-1}$	$Y_{s,t-1}$	pe_{t-1}
Australia	9/69	309	0.13	0.02	10.71	9.71	4.28	13.95
			(6.82)	(2.37)	(2.80)	(3.87)	(1.14)	(4.69)
Belgium	4/69	313	0.47	0.04	9.23	8.81	8.29	13.13
			(4.76)	(1.44)	(1.94)	(2.03)	(3.24)	(4.07)
Canada	12/69	305	0.12	0.08	9.81	8.84	3.63	19.36
			(5.02)	(2.51)	(2.18)	(3.43)	(0.75)	(16.08)
France	9/71	284	0.33	0.05	10.01	9.63	4.73	17.99
			(6.22)	(2.04)	(2.49)	(2.81)	(1.71)	(21.45)
Germany	3/69	314	0.18	0.09	7.76	6.50	4.00	15.09
			(5.23)	(1.97)	(1.26)	(2.85)	(1.02)	(9.89)
Italy	1/70	304	−0.20	−0.04	12.45	12.55	2.74	42.90
			(7.04)	(1.86)	(3.63)	(4.33)	(0.79)	(73.44)
Japan	2/78	207	0.27	0.22	6.31	5.64	1.14	37.96
			(5.50)	(3.03)	(1.69)	(2.18)	(0.56)	(20.78)
Netherlands	4/69	313	0.38	0.11	8.14	6.50	5.44	9.25
			(4.90)	(2.58)	(1.41)	(2.47)	(1.38)	(3.84)
Norway	3/85	122	0.20	0.18	10.74	11.14	2.27	27.31
			(8.16)	(1.86)	(2.45)	(3.20)	(0.76)	(46.23)
Spain	1/78	208	0.16	−0.08	13.09	14.31	7.53	14.80
			(6.25)	(3.32)	(2.37)	(5.02)	(3.77)	(22.13)
Sweden	2/70	303	0.59	0.00	10.46	9.47	3.57	12.55
			(6.22)	(2.48)	(2.30)	(3.44)	(1.46)	(6.67)
Switzerland	5/69	312	0.40	0.17	5.15	3.06	3.33	12.49
			(4.91)	(1.02)	(1.06)	(3.03)	(1.03)	(3.09)
UK	2/69	315	0.33	0.05	10.77	9.88	4.96	11.86
			(6.34)	(3.11)	(2.00)	(2.94)	(1.17)	(4.06)
US	2/56	471	0.37	−0.03	7.23	5.72	3.67	14.96
			(4.16)	(2.90)	(2.80)	(2.87)	(0.80)	(4.36)

($t=1$) indicates the initial month of the sample; $T+n$ denotes the sample size, covering both the estimation and testing sample; r_t is the excess return on the stock index over month t (in percentage points); $r_{b,t-1}$ is the excess bond return over month $t-1$ (in percentage points); $Y_{b,t-1}$ is the bond yield as observed at the end of month $t-1$ (annualized and in percentage points); $r_{f,t-1}$ is the short-term yield as observed at the end of month $t-1$ (annualized and in percentage points); $Y_{s,t-1}$ denotes the stock dividend yield as of the end of month $t-1$ (annualized and in percentage points); pe_{t-1} denotes the stock price-earnings ratio at the end of month $t-1$. Standard deviation in parentheses.

Of course, one may wonder whether this evidence in favor of predictability is caused by our picking particular predictors, mainly inspired by previous work (Solnik's), and hence potentially affected by data snooping biases. Formal selection criteria are meant to address this issue, and we will discuss results from their application, reported in Table 3.

The discrepancy between the regression results of Period I (entire sample) and Period II (1/71–8/80) indicates the presence of model nonstationarity. The sign of the regression coefficients is almost always the same across the two periods, but the magnitude often differs dramatically (with Period II generating the highest values). Pesaran and Timmermann (1995) also document an increase in predictability of U.S. stock returns in the 1970s.

4. Empirical Results

We first discuss the nature of the prediction models that each of the formal selection criteria picked; in particular, whether they implied return

Table 2
OLS regressions of monthly excess stock returns on a subset of predictors, various sample periods

Country	Predictors										R^2	
	Intercept		January dummy		Dividend yield		Short-term yield		Bond Yield			
	I	II	I	II	I	I	I	II	I	II	I	II
Australia	−0.044	−0.080	0.021	0.047	0.0093	0.0139	−0.0030	−0.0177	0.0031	0.0165	.033	.125
	(0.022)	(0.037)	(0.016)	(0.022)	(0.0042)	(0.0065)	(0.0025)	(0.0069)	(0.0034)	(0.0081)		
Belgium	−0.001	0.038	0.033	0.053	0.0023	0.0098	−0.0211	−0.0531	0.0184	0.0365	.086	.364
	(0.014)	(0.022)	(0.011)	(0.011)	(0.0016)	(0.0023)	(0.0057)	(0.0101)	(0.0058)	(0.0100)		
Canada	−0.009	−0.063	0.015	0.035	0.0072	0.0055	−0.0032	−0.0046	0.0012	0.0089	.038	.044
	(0.023)	(0.060)	(0.012)	(0.018)	(0.0048)	(0.0079)	(0.0020)	(0.0049)	(0.0031)	(0.0104)		
France	0.002	−0.031	0.022	0.041	0.0028	0.0165	−0.0072	−0.0068	0.0056	−0.0004	.039	.137
	(0.019)	(0.044)	(0.015)	(0.021)	(0.0031)	(0.0065)	(0.0029)	(0.0039)	(0.0036)	(0.0081)		
Germany	0.002	−0.027	0.0057	0.043	0.0059	0.0074	−0.0021	−0.0025	−0.0015	0.0008	.020	.123
	(0.021)	(0.030)	(0.012)	(0.013)	(0.0037)	(0.0047)	(0.0014)	(0.0014)	(0.0032)	(0.0031)		
Italy	−0.032	−0.007	0.046	0.039	0.0042	−0.0064	−0.0008	−0.0057	0.0021	0.0075	.042	.082
	(0.025)	(0.030)	(0.016)	(0.021)	(0.0064)	(0.0097)	(0.0022)	(0.0025)	(0.0024)	(0.0036)		
Japan	0.034	0.245	0.015	−0.006	0.0063	−0.1417	−0.0006	−0.0086	−0.0046	0.0142	.023	.315
	(0.019)	(0.132)	(0.014)	(0.015)	(0.0118)	(0.0629)	(0.0030)	(0.0028)	(0.0057)	(0.0054)		
Netherlands	0.006	−0.012	0.031	0.054	0.0125	0.0150	−0.0016	−0.0031	−0.0081	−0.0080	.090	.212
	(0.019)	(0.034)	(0.011)	(0.014)	(0.0033)	(0.0043)	(0.0016)	(0.0019)	(0.0037)	(0.0046)		
Norway	0.131	NO	0.029	NO	0.0146	NO	0.0034	NO	−0.0165	NO	.030	NO
	(0.141)		(0.040)		(0.0169)		(0.0129)		(0.0179)			
Spain	−0.002	−0.005	0.024	−0.001	−0.0070	−0.0345	−0.0019	−0.0020	0.0029	0.0319	.047	.400
	(0.038)	(0.076)	(0.018)	(0.032)	(0.0021)	(0.0110)	(0.0010)	(0.0008)	(0.0035)	(0.0103)		
Sweden	−0.054	0.016	0.031	0.043	0.0029	0.0293	−0.0014	−0.0054	0.0058	−0.0138	.051	.180
	(0.023)	(0.026)	(0.013)	(0.013)	(0.0027)	(0.0098)	(0.0018)	(0.0023)	(0.0029)	(0.0055)		
Switzerland	−40.002	−0.014	0.021	0.063	0.0090	0.0144	0.0017	−0.0006	−0.0065	−0.0094	.029	.159
	(0.018)	(0.024)	(0.011)	(0.015)	(0.0050)	(0.0084)	(0.0015)	(0.0034)	(0.0039)	(0.0054)		
UK	0.006	−0.012	0.040	0.051	0.0233	0.0196	−0.0022	−0.0058	−0.0090	−0.0033	.098	.123
	(0.026)	(0.052)	(0.015)	(0.025)	(0.0059)	(0.0082)	(0.0018)	(0.0033)	(0.0043)	(0.0081)		
US	−0.028	−0.037	0.008	0.019	0.0125	0.0205	−0.0068	−0.0093	0.0033	0.0016	.074	.172
	(0.010)	(0.025)	(0.007)	(0.014)	(0.0034)	(0.0078)	(0.0015)	(0.0028)	(0.0016)	(0.0066)		

The results under "I" refer to the complete sample. Those under "II" refer to the sample starting in 1/71 or later (in the absence of observations) and running until 8/80. OLS estimates of the coefficients are displayed. Standard errors in parentheses. The predictors "Dividend yield," "Short-term yield" and "Bond yield" are annualized and expressed in percentage terms. Unlike in Table 1, the dependent variable, the monthly excess stock return is not expressed in percentage points. NO = no observations in period II.

predictability. Subsequently, we report tests of the external validity of the selected models (whenever they did imply return predictability in-sample). The estimation sample runs until 5/90. The testing sample covers the remaining period: 6/90–5/95 (hence, $n = 60$).

4.1 Retained Forecasting Models

The seven model selection criteria were allowed to pick a suitable prediction model based on the 10 predictors listed in Table 1, as well as a constant (the intercept in the linear prediction model). As mentioned before, predictability is decided against whenever only the constant is retained. The parameters in the prediction models were estimated with OLS.

Table 3 displays the composition of the linear prediction model preferred by each of the selection criteria. A zero indicates that the corresponding variable was discarded, a one indicates that the variable was kept. The following facts stand out when looking at this table.

- The adjusted R^2 almost invariably selects a model with more predictors than other criteria. This overfitting occurs despite the (admittedly ad hoc) adjustment for the loss of degrees of freedom when regressors are added. In the case of the United States, for instance, the adjusted R^2

Table 3
Model choice over the estimation sample (to 5/90), various selection criteria

Country	SSC	d_J	r_{t-1}	r_{t-2}	$r_{b,t-1}$	$r_{b,t-2}$	P_{t-1}	$Y_{b,t-1}$	$r_{f,t-1}$	$Y_{s,t-1}$	pe_{t-1}
Australia	$\overline{R}^2$	1	0	0	0	0	0	0	0	1	0
	AIC	0	0	0	0	0	0	0	0	1	0
	BIC	0	0	0	0	0	0	0	0	0	0
	FIC	0	0	0	1	1	0	0	0	0	0
	PIC	0	0	1	1	1	0	0	0	0	0
	PLS	1	0	0	0	0	0	0	0	0	0
	PLS-MDC	0	0	0	0	0	0	0	0	0	1
Belgium	$\overline{R}^2$	1	1	0	0	0	1	1	1	1	1
	AIC	1	1	0	0	0	0	1	1	1	1
	BIC	1	1	0	0	0	0	1	1	0	0
	FIC	1	1	1	1	1	0	0	0	0	0
	PIC	1	1	1	1	1	0	1	1	0	0
	PLS	1	1	0	0	0	0	0	0	0	1
	PLS-MDC	1	1	0	0	0	0	1	1	0	0
Canada	$\overline{R}^2$	1	1	1	0	0	1	0	0	1	1
	AIC	0	0	1	0	1	0	0	1	0	0
	BIC	0	0	0	0	0	0	0	0	0	0
	FIC	0	0	1	1	1	0	0	0	0	0
	PIC	0	1	1	1	1	0	0	0	0	0
	PLS	1	0	0	0	0	0	0	0	0	0
	PLS-MDC	0	1	0	0	0	0	0	0	1	0
France	$\overline{R}^2$	1	0	0	1	1	1	0	1	1	0
	AIC	1	0	0	1	0	1	0	1	1	0
	BIC	0	0	0	1	0	0	0	0	0	0
	FIC	0	0	0	1	1	0	0	0	0	0
	PIC	0	1	1	1	1	0	0	0	0	0
	PLS	0	0	0	1	0	0	0	0	0	0
	PLS-MDC	0	1	0	0	0	0	0	0	0	0
Germany	$\overline{R}^2$	0	0	0	1	0	1	0	1	0	1
	AIC	0	0	0	0	0	0	0	1	0	1
	BIC	0	0	0	0	0	0	0	0	0	0
	FIC	0	0	0	1	1	0	0	0	0	0
	PIC	0	1	1	1	1	0	0	0	0	0
	PLS	0	0	0	1	0	0	0	0	0	0
	PLS-MDC	0	1	0	0	0	0	0	0	0	0
Italy	$\overline{R}^2$	1	1	0	0	0	0	1	0	1	0
	AIC	1	1	0	0	0	0	0	0	0	0
	BIC	1	1	0	0	0	0	0	0	0	0
	FIC	1	1	1	1	1	0	0	0	0	0
	PIC	1	1	1	1	1	0	0	0	0	0
	PLS	1	1	0	0	0	0	0	0	0	0
	PLS-MDC	1	1	0	0	1	0	0	0	0	0
Japan	$\overline{R}^2$	0	0	1	0	0	1	1	0	1	1
	AIC	0	0	1	0	0	1	1	0	1	1
	BIC	0	0	0	0	0	1	0	0	1	0
	FIC	0	1	1	1	1	0	0	0	0	0
	PIC	0	1	1	1	1	0	0	0	0	0
	PLS	0	0	0	1	1	0	0	0	0	0
	PLS-MDC	0	1	0	0	0	0	0	0	0	0
Netherlands	$\overline{R}^2$	1	1	0	1	0	1	0	1	1	0
	AIC	1	0	0	1	0	1	0	1	1	0
	BIC	1	0	0	0	0	1	1	0	1	0
	FIC	1	1	0	1	1	0	0	0	0	0

Table 3
(continued)

Country	SSC	d_J	r_{t-1}	r_{t-2}	$r_{b,t-1}$	$r_{b,t-2}$	P_{t-1}	$Y_{b,t-1}$	$r_{f,t-1}$	$Y_{s,t-1}$	pe_{t-1}
Netherlands	PIC	1	1	1	1	1	0	0	0	0	0
cont'd.	PLS	1	0	0	0	0	0	1	0	1	0
	PLS-MDC	1	0	0	0	0	0	1	0	1	0
Norway	$\overline{R}^2$	0	1	0	0	1	1	1	1	0	0
	AIC	0	1	0	0	0	0	0	0	0	0
	BIC	0	0	0	0	0	0	0	0	0	0
	FIC	0	1	1	1	1	0	0	0	0	0
	PIC	0	1	1	1	1	0	0	0	0	0
	PLS	0	0	0	0	1	0	0	0	0	0
	PLS-MDC	1	1	0	0	1	1	1	0	0	0
Spain	$\overline{R}^2$	1	1	0	0	0	1	0	1	1	1
	AIC	1	1	0	0	0	1	0	1	1	1
	BIC	0	0	0	0	0	0	0	0	0	0
	FIC	0	1	1	1	1	0	0	0	0	0
	PIC	0	1	1	1	1	0	0	0	0	0
	PLS	0	0	0	0	0	0	0	1	0	0
	PLS-MDC	0	0	0	0	0	0	0	1	0	0
Sweden	$\overline{R}^2$	1	1	0	0	1	0	1	0	0	1
	AIC	1	1	0	0	0	0	1	0	0	0
	BIC	1	0	0	0	0	0	0	0	0	0
	FIC	1	1	0	1	1	0	0	0	0	0
	PIC	1	1	1	1	1	0	0	0	0	0
	PLS	1	1	0	0	1	0	0	0	0	0
	PLS-MDC	1	0	0	0	0	0	0	0	0	1
Switzerland	$\overline{R}^2$	1	0	1	1	1	1	1	0	1	1
	AIC	1	0	1	0	0	0	0	0	1	1
	BIC	0	0	0	0	0	0	0	0	1	1
	FIC	0	0	1	1	1	0	0	0	0	0
	PIC	1	1	1	1	1	0	0	0	1	1
	PLS	0	0	0	0	0	0	0	1	1	1
	PLS-MDC	0	0	0	0	0	0	0	1	1	1
UK	$\overline{R}^2$	1	1	1	1	1	1	0	1	1	1
	AIC	1	1	0	0	0	0	1	0	1	0
	BIC	1	0	0	0	0	0	1	0	1	0
	FIC	1	0	1	1	1	0	0	0	1	0
	PIC	1	1	1	1	1	0	1	0	1	0
	PLS	1	0	0	0	0	0	0	0	0	0
	PLS-MDC	0	1	0	0	0	0	0	0	0	0
US	$\overline{R}^2$	1	0	0	1	0	1	0	1	1	1
	AIC	0	0	0	1	0	1	0	1	1	0
	BIC	0	0	0	0	0	1	0	1	1	0
	FIC	0	0	0	1	1	0	0	1	1	0
	PIC	0	1	1	1	1	0	0	1	1	0
	PLS	0	0	0	0	0	0	0	1	1	0
	PLS-MDC	0	0	0	0	0	0	0	1	1	0

d_J denotes the January dummy; P_{t-1} is the stock price level at the end of month $t-1$. "SSC" is short for statistical selection criterion. Results for the following SSC are displayed: $\overline{R}^2$ (adjusted R^2), AIC (Akaike information criterion), BIC (Schwarz's information criterion), FIC (Fisher information criterion), PIC (posterior criterion), PLS (predictive least squares criterion), and PLS-MDC (predictive least squares criterion adjusted with Markov dimension criterion). See text for additional information. The entry "0" indicates that the corresponding predictor was dropped by the selection criterion. The entry "1" indicates that the predictor was retained.

retains six predictors, whereas PLS and PLS-MDC limit the forecasting model to two variables.

- Some popular predictors are only occasionally selected. For instance, in 9 of 14 cases, BIC and FIC drop the January dummy.[6]
- The model selection criteria often retain predictors that are suspected to be unit-root nonstationary, such as the dividend yield, the price-earnings ratio, or even the initial stock price level. Model selection criteria such as FIC, PIC, PLS, and PLS-MDC are robust to unit-root nonstationarities, so this finding should not bother us.
- For a given country, the model selection criteria agree to a certain extent on the variables to be retained as predictors. The contemporaneous yield on short-term Treasury bills, for instance, is dropped by *all* selection criteria in 4 of 14 cases. It is retained by all selection criteria only for the United States.
- FIC and PIC almost universally select lagged bond and stock excess returns. These are often chosen in pairs, so that FIC and PIC effectively construct the moving average predictors that have been popular in professional circles lately.
- BIC decides against predictability in five cases (Australia, Canada, Germany, Norway, and Spain). This means that all predictors are dropped (only the intercept is retained in the prediction model).
- PLS almost invariably picks a small model when compared to other criteria, except BIC. PLS-MDC adjusts the choice by changing and/or adding predictors.

Overall, however, the predictability that remains after the application of formal selection criteria confirms the evidence of predictability in earlier studies. In other words, *the predictability that was uncovered in previous work is clearly not caused by overfitting.*

Notice that this verdict is uniform across selection criteria. Since each selection criterion starts from a different decision-theoretic framework, it is comforting to observe such an agreement. In other words, our conclusion is not based on the application of a specific, haphazardly chosen criterion.

4.2 External Validation

Formal model selection criteria try to determine the model with the best external validity. To verify whether they indeed pick models with external validity in the context of stock market returns, we tested for out-of-sample forecasting power by projecting the excess returns in our testing sample (6/90–5/95) onto the forecasts from each "optimal" model.

[6] It is well known that the January effect is less pronounced for large stock. As the excess stock return used in this article refers to a subset of large and liquid companies, it should not be too surprising that the January effect often fails to emerge.

Table 4 displays the results. Exploiting the information in the cross section of stock returns, we used SUR regression to improve the estimate of the t-ratios. For comparison, other popular out-of-sample statistics [the adjusted R^2 and root mean square prediction error of the out-of-sample regressions, that is, of Equation (3)] are provided as well. In order to put the root mean square prediction error of the out-of-sample regressions [see Equation (3)] into perspective, we also report the root mean square excess stock return over the estimation and testing samples (in brackets).

Table 4 generates the following initial reactions about out-of-sample predictability.

- For models based on the adjusted R^2, an enormous disparity emerges between in-sample and out-of-sample adjusted R^2. For the U.S. stock market, for instance, the in-sample adjusted R^2 is 8.3%. It drops to 0.5% out of sample.[7]
- The in-sample adjusted R^2s of models selected on the basis of BIC, FIC, PIC, PLS, and PLS-MDC are a somewhat more reliable indicator of the out-of-sample goodness-of-fit. For FIC and PIC, for instance, the out-of-sample adjusted R^2 is at least as high as the in-sample adjusted R^2 in 6 of 14 cases.[8]
- A comparison of the out-of-sample root mean square prediction error and the root mean square excess return reveals an across-the-board dismal performance of the models selected by the formal criteria. As a matter of fact, the decision by BIC not to include predictors at all in many cases is corroborated by this finding.
- Our model selection criterion, PLS-MDC, overfits as much as FIC or PIC.

All this indicates poor out-of-sample predictability. A formal test of external validity of the retained models, however, can be obtained from the t-ratios in Table 4. Except for Japan, we find that none of the t-ratios of the slope coefficient in the SUR regression of out-of-sample outcomes onto predictions are significant![9] Consequently, *we fail to find that models chosen on the basis of formal selection criteria have any external validity*.

Table 5 summarizes the tests by reporting F-tests of the *joint* significance of the slope coefficients in an SUR regression of out-of-sample excess stock

[7] The poor out-of-sample record in terms of R^2 is actually a bit exaggerated, because the out-of-sample use of this measure supposes that the forecaster knew beforehand in what direction the selected model would be biased.

[8] When BIC decides not to include any predictors at all, the in-sample and out-of-sample adjusted R^2s are set equal to zero.

[9] When running simple OLS (as opposed to SUR), the t-statistics are significant (at the 5% level) only for four stock markets: France (BIC, FIC, PIC, and PLS), Japan (BIC), The Netherlands (BIC, PLS, and PLS-MDC), and the United Kingdom (AIC and PIC).

Table 4
Testing-sample (6/90–5/95) performance of prediction models selected by several formal statistical criteria

Country	SSC	$\overline{R}^2$		t	RMS	
		In	Out	Out	In	Out
Australia					(7.29)	(4.40)
	$\overline{R}^2$	.019	.001	−0.08	7.18	4.50
	AIC	.017	.000	0.39	7.20	4.45
	BIC	—	—	—	7.29	4.40
	FIC	−.005	.017	−1.36	7.27	4.43
	PIC	−.005	.018	−1.91	7.25	4.44
	PLS	.003	.001	−0.66	7.40	4.41
	PLS-MDC	.003	.016	1.44	7.40	4.39
Belgium					(4.87)	(4.28)
	$\overline{R}^2$	.107	.005	0.35	4.53	4.51
	AIC	.106	.006	0.27	4.54	4.49
	BIC	.089	.000	0.53	4.60	4.70
	FIC	.046	.023	0.67	4.70	4.29
	PIC	.085	.000	0.48	4.58	4.70
	PLS	.070	.011	0.04	4.82	4.30
	PLS-MDC	.089	.000	0.11	4.85	4.70
Canada					(5.36)	(3.27)
	$\overline{R}^2$	.042	.004	−0.20	5.19	3.29
	AIC	.041	.002	0.02	5.21	3.33
	BIC	—	—	—	5.36	3.27
	FIC	.033	.034	0.08	5.23	3.37
	PIC	.029	.034	0.11	5.23	3.37
	PLS	.002	.000	−0.79	5.47	3.27
	PLS-MDC	−.001	.032	−0.88	5.51	3.30
France					(6.42)	(5.41)
	$\overline{R}^2$	.047	.013	−1.35	6.17	5.55
	AIC	.047	.011	−1.54	6.19	5.53
	BIC	.024	.058	−0.52	6.32	5.27
	FIC	.022	.058	−0.18	6.31	5.26
	PIC	.016	.072	0.44	6.38	5.22
	PLS	.024	.058	−0.46	6.49	5.27
	PLS-MDC	.004	.005	−0.82	6.51	5.39
Germany					(5.13)	(5.65)
	$\overline{R}^2$	.015	.000	−1.06	5.04	5.76
	AIC	.012	.001	−1.14	5.07	5.73
	BIC	—	—	—	5.13	5.65
	FIC	.005	.006	−0.38	5.09	5.60
	PIC	.000	.001	−0.98	5.08	5.64
	PLS	.006	.022	0.02	5.21	5.58
	PLS-MDC	.001	.003	−1.55	5.26	5.66
Italy					(6.96)	(7.40)
	$\overline{R}^2$	.051	.017	0.91	6.72	7.32
	AIC	.049	.022	0.82	6.75	7.27
	BIC	.049	.022	0.65	6.75	7.27
	FIC	.045	.032	1.31	6.72	7.23
	PIC	.045	.032	1.25	6.72	7.23
	PLS	.049	.022	0.83	7.03	7.27
	PLS-MDC	.048	.026	0.70	7.09	7.27
Japan					(4.59)	(7.20)
	$\overline{R}^2$	.124	.018	−0.01	4.19	7.26
	AIC	.123	.020	0.28	4.21	7.24
	BIC	.083	.062	1.41	4.35	7.19
	FIC	.036	.020	−2.40**	4.43	7.51
	PIC	.036	.020	−2.37**	4.43	7.51
	PLS	.040	.002	−0.81	4.60	7.30
	PLS-MDC	.002	.068	−3.08**	4.67	7.42

Table 4
(continued)

Country	SSC	$\overline{R}^2$		t	RMS	
		In	Out	Out	In	Out
Netherlands					(5.12)	(3.84)
	$\overline{R}^2$	.097	.022	0.45	4.80	3.84
	AIC	.095	.021	0.32	4.82	3.84
	BIC	.076	.055	0.65	4.89	3.88
	FIC	.052	.032	1.30	4.94	3.76
	PIC	.048	.031	1.28	4.94	3.76
	PLS	.076	.055	0.66	5.11	3.88
	PLS-MDC	.076	.055	1.36	5.11	3.88
Norway					(8.21)	(8.08)
	$\overline{R}^2$	.065	.001	0.30	7.55	8.51
	AIC	.026	.012	0.01	7.97	8.31
	BIC	—	—	—	8.21	8.08
	FIC	.000	.021	0.37	7.88	8.53
	PIC	.000	.021	0.33	7.88	8.53
	PLS	.001	.015	0.09	8.65	8.27
	PLS-MDC	.054	.002	−0.37	9.68	8.55
Spain					(6.11)	(6.60)
	$\overline{R}^2$	.103	.001	−0.09	5.65	8.66
	AIC	.103	.001	−0.07	5.65	8.66
	BIC	—	—	—	6.11	6.60
	FIC	.007	.008	0.26	5.99	6.60
	PIC	.007	.008	0.14	5.99	6.60
	PLS	.023	.004	−0.45	6.21	6.67
	PLS-MDC	.023	.004	−0.95	6.21	6.67
Sweden					(5.72)	(7.94)
	$\overline{R}^2$	.048	.001	−0.26	5.51	8.11
	AIC	.046	.013	0.23	5.55	7.88
	BIC	.020	.025	1.03	5.64	7.80
	FIC	.026	.017	0.05	5.59	7.87
	PIC	.022	.016	−0.08	5.59	7.88
	PLS	.033	.036	0.61	5.64	7.75
	PLS-MDC	.019	.004	0.45	5.68	7.90
Switzerland					(4.96)	(4.80)
	$\overline{R}^2$	.063	.001	0.17	4.71	5.02
	AIC	.057	.002	−0.49	4.77	4.96
	BIC	.037	.000	−0.27	4.83	4.94
	FIC	.010	.001	−0.12	4.89	4.79
	PIC	.055	.000	−0.05	4.74	4.99
	PLS	.034	.007	−0.77	4.91	4.95
	PLS-MDC	.034	.007	−0.61	4.91	4.95
UK					(6.72)	(4.43)
	$\overline{R}^2$	.100	.037	0.16	6.25	4.78
	AIC	.095	.048	0.00	6.33	4.51
	BIC	.082	.029	−0.72	6.39	4.52
	FIC	.073	.022	0.74	6.40	4.45
	PIC	.091	.056	0.80	6.31	4.46
	PLS	.027	.000	−1.32	6.84	4.54
	PLS-MDC	.004	.011	0.35	6.87	4.36

Table 4
(continued)

Country	SSC	$\overline{R}^2$		t	RMS	
		In	Out	Out	In	Out
US					(4.27)	(3.34)
	$\overline{R}^2$	.083	.005	−0.13	4.06	3.46
	AIC	.082	.007	−0.24	4.07	3.41
	BIC	.077	.003	−0.45	4.08	3.45
	FIC	.065	.004	0.53	4.11	3.25
	PIC	.063	.024	−0.07	4.10	3.28
	PLS	.057	.041	1.14	4.25	3.28
	PLS-MDC	.057	.041	0.90	4.25	3.28

"In" refers to the estimation sample (running until 5/90); "Out" refers to the testing sample (6/90–5/95). $\overline{R}^2$ denotes the adjusted R^2. t denotes the t-ratio of the slope coefficient in an SUR regression of excess stock returns onto our predictions; RMS denotes root mean square error (×100) [in parentheses: root mean square of monthly excess returns (×100)]. The out-of-sample results are based on rolling (i.e., expanding-window) regressions."SSC" is short for statistical selection criterion. Results for the following SSC are displayed: $\overline{R}^2$ (adjusted R^2), AIC (Akaike information criterion), BIC (Schwarz's information criterion), FIC (Fisher information criterion), PIC (posterior criterion), PLS (Predictive Least Squares Criterion), and PLS-MDC (predictive least squares criterion adjusted with Markov dimension criterion). See text for additional information. One and two asterisks denote significance at the 5% and 1% level, respectively (two-tailed).

Table 5
Testing-sample (6/90–5/95) performance of prediction models selected by several formal statistical criteria: summary

Criterion	F	p-value
$\overline{R}^2$	0.36	.99
AIC	0.42	.97
BIC	0.60	.80
FIC	0.94	.52
PIC	1.09	.37
PLS	0.69	.31
PLS-MDC	1.65	.06

F-tests measure the joint significance of the slope coefficients in regressions of excess stock returns onto our predictions for all countries. They are based on SUR estimation of the slope coefficients. See Table 4 (column 5) for t-tests on indivual countries.

returns onto our predictions. The results corroborate the findings based on the individual t-ratios.

In conclusion, Tables 4 and 5 provide broad evidence against any out-of-sample forecasting power of models chosen on the basis of selection criteria that are nevertheless supposed to have corrected for overfitting. Again, the evidence is uniform across selection criteria: the conclusions are not caused by an accidental choice of a selection criterion that happens to imply the opposite.

5. Power

We failed to find out-of-sample forecasting power even if the models are chosen by established selection criteria that are constructed to provide ex-

ternal validity rather than maximum in-sample fit. To gauge the significance of these findings, a discussion of the power of the tests is in order.

For BIC, for instance, the out-of-sample tests reported in Tables 4 and 5 are based on nine samples of 60 observations (months) each. The t-tests of the slope coefficients of outcomes onto out-of-sample predictions fail to reject the null of no predictability in all nine samples. The joint F-test further corroborates this lack of predictability. One may wonder whether these outcomes are expected in view of the relatively low in-sample R^2 generated by the retained models. What are the chances of finding nine insignificant slope coefficients in samples of 60 observations when the true R^2 is 6%? This is the average R^2 that BIC-based models generated across the nine markets. See Table 4. (We ignore the difference between the *adjusted* R^2, which is reported in Table 4, and the standard R^2.)

Figure 1 provides the answer. The left panel displays the power function of 5% t-tests on the slope coefficient in OLS projections among normally distributed variables. The solid line is based on an analytical approximation of the actual power function, and the circles represent estimates based on 500 replications of samples of size 60.[10] For an R^2 of 6%, the power is 0.48, indicating that there is only a one in two chance of rejecting the (false) null of no predictability.

The right panel of Figure 1 translates the power function into the probability of finding *no* rejections in a cross section of nine (independent) samples, as a function of the R^2. In other words, it provides an idea of the probability of obtaining the results that were reported in Tables 4 and 5 if there were actually predictability. The chance of obtaining no rejections of lack of predictability if in fact the true R^2 is 6% is only 0.003, that is, 1 in 333. Hence we cannot attribute our findings to lack of power.

The results for the other criteria are analogous. They sometimes generate lower R^2s, but support predictability in more than nine markets. PIC, for instance, generates an average R^2 of only 3.5%, yet supports predictability in each market. The probability of observing no rejections in the 14 samples of size 60 is 0.007, that is, about 1 in 140.

6. Conclusion

We applied formal statistical model selection criteria to determine the best model to predict excess stock returns in an international dataset. The choice set consisted of linear models with varying numbers of predictors. This included one with only an intercept. Selection of the latter would clearly

[10] The analytical approximation is based on suggestions advocated in Anderson (1984) for sample sizes larger than 25. See page 123 of his book. The Monte Carlo analysis is based on bivariate normal random draws with variance one. The actual value of the variances does not matter because the t-test is scale-free: it depends on neither the variance of the regressor nor that of the regressand, but instead on the R^2 only.

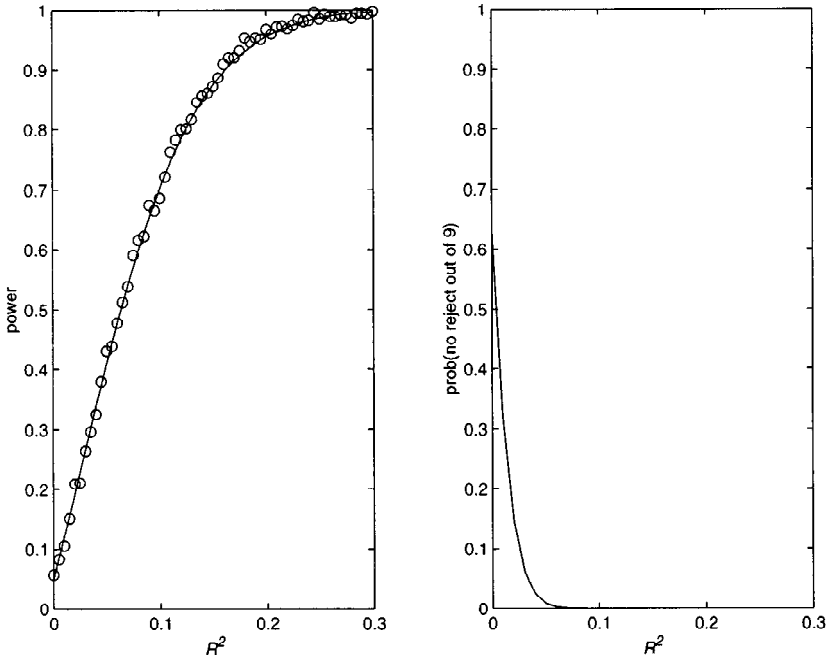

Figure 1
Left panel: Power of the 5% *t*-test of the slope coefficient of OLS projections among two normally distributed random variables, as a function of R^2. Sample size: 60. The solid line depicts an analytical approximation based on the suggestions in Anderson (1984, p. 123). The circles are based on a Monte Carlo study with 500 replications each. *Right Panel*: Probability of observing no rejections in nine independent trials when the probability of rejection is as indicated by the power function in the left panel.

indicate strong evidence against predictability, even if all other models will always have a better in-sample fit.

We only occasionally observed cases where a selection criterion decides against predictability. Overall we find ample evidence of predictability, confirming the recent wave of evidence that stock excess returns can succesfully be predicted. In other words, such evidence does not appear to have been caused by model overfitting.

At the same time, however, we find that the out-of-sample forecasting power of retained prediction models is nil. Not a single model could be validated externally, upending the very goal of the model selection criteria we used. This clearly demonstrates model nonstationarity in excess stock returns. It is not clear why model nonstationarity would be present, however. While restrictions in the choice set (only linear forecasting models could be selected) may in part be blamed (indeed, the "true" return generating model may be nonlinear), it is surprising that no linear model generates any out-

of-sample forecasting power. Indeed, the models were chosen with criteria that pick only the "best!" As mentioned in the introduction, however, recent theoretical asset pricing results do suggest that our evidence may be related to learning in the marketplace.

Similar evidence appears in a preliminary report on a study of the performance of trading rules on U.S. stock and futures data [Sullivan, Timmermann, and White (1997)]. They searched for the trading rule that did best in-sample (in terms of average returns or Sharpe ratio). Using a stationary bootstrapping procedure, they confirmed that outperformance of the selected trading rule was not caused by model overfitting. Out-of-sample, however, the best trading rule did not outperform. Notice, however, that the trading rule with the best in-sample performance does not necessarily have the best external validity. In contrast, our study was designed to pick the forecasting model that would have maximum external validity. Therefore, it is all the more powerful and striking that we find that models fail out-of-sample even if they are chosen exclusively for their external promise.

Appendix A. Formal Model Selection Criteria

Several model selection criteria are used to choose among K linear models. Equation (1) displays a typical model. The choice is based on a sample of T observations, the estimation sample: $t = 1, \ldots, T$.

The adjusted R^2 criterion is well known. To define the other criteria, let

$$X_{t,l}^{k'} = [x_{\max(0,t-l)}^{k}, \ldots, x_{t-1}^{k}],$$

and

$$Y_{t,l}' = [r_{\max(1,t-l+1)}, \ldots, r_t].$$

Also,

$$A_{t,l}^{k} = X_{t,l}^{k'} X_{t,l}^{k}.$$

The least squares regression coefficient equals

$$\hat{\theta}_{t,l}^{k} = (A_{t,l}^{k})^{-1} X_{t,l}^{k'} Y_{t,l} \tag{4}$$

(this expression is only well-defined for $t \geq m$ and $l \geq m$, where m is such that $A_{m,m}^{k}$ is invertible). Now define the sum of squared errors of model k in the subsample from observation $t - l + 1$ to t, computed from the least squares estimate of the slope coefficients in the same subsample:

$$SSE_{t,l}^{k} = (Y_{t,l} - X_{t,l}^{k} \hat{\theta}_{t,l}^{k})'(Y_{t,l} - X_{t,l}^{k} \hat{\theta}_{t,l}^{k}).$$

This sum of squared regression errors is the workhorse of traditional information criteria.

The following information criteria can be defined in terms of $SSE_{t,l}^{k}$. First, Akaike's information criterion:

$$AIC(k) = T \log \frac{SSE_{T,T}^{k}}{T} + 2p^{k}. \tag{5}$$

Second, Schwarz's criterion:

$$BIC(k) = T \log \frac{SSE^k_{T,T}}{T} + p^k \log T. \tag{6}$$

For FIC and PIC, define model K to be the largest model (i.e., the one with the largest number of regressors). This model will be used as the benchmark.

$$FIC(k) = SSE^k_{T,T} \frac{T}{T - p^k} + \frac{SSE^K_{T,T}}{T - p^K} \log \left(\frac{|A^k_{T,T}|}{\frac{SSE^k_{T,T}}{T-p^k}} \right); \tag{7}$$

$$PIC(k) = SSE^K_{T,T} \left(\frac{SSE^k_{T,T}}{SSE^K_{T,T}} - 1 \right) + \frac{SSE^K_{T,T}}{T - p^K} \log \left(\frac{|A^k_{T,T}|}{\frac{SSE^K_{T,T}}{T-p^K}} \right). \tag{8}$$

In each case, the model is chosen that minimizes the criterion function.

To define the PLS and our extension, PLS-MDC, specify the average sum of squares differently:

$$\mu^l_T(k) = \frac{1}{T - b} \sum_{t=b}^{T} (r_t - \hat{\theta}^{k'}_{t-1,l} x^k_{t-1})^2,$$

where b is the first integer such that $\hat{\theta}^k_{b-1,b-1}$ is uniquely defined [see Equation (4)]. PLS then picks the model (k) that minimizes $\mu^T_T(k)$. Let

$$\mathcal{K} = \{k : T = \arg\min\{\mu^l_T(k)\}\}$$

[$\mu^l_T(k)$ is minimized over $l \in \{b, b+1, \ldots, \bar{l}, T\}$, with $\bar{l} < \infty$]. PLS-MDC picks the model that minimizes $\mu^T_T(k)$ for k *restricted to lie in* $\mathcal{K}$.[11] For a formal analysis, see Appendix B.

Appendix B. PLS-MDC

Rolling regressions can be used to determine the dimension of the state vector in a Markov model. We will first demonstrate why, then we will explain how this generates a decision rule that can improve on the ability of PLS to pick the right prediction model.

To be as general as possible, we allow the state vector to include several lags of the variable to be predicted, as well as lags of a limited set of other conditioning variables. Also, the regression equation can be general. Specifically, it can be nonlinear, in contrast to the model in the main part of the article.

Consider a time series $\{r_t; t = 1, 2, \ldots\}$ taking values in R. Posit that it is Markov relative to a state vector x_t. This state vector may include $r_{t-1}, \ldots, r_{t-q}$ $(q < \infty)$ and various lags of other conditioning variables. Let f denote the regression function:

$$r_t = f(x_{t-1}) + \epsilon_t, \tag{9}$$

[11] The optimal model according to PLS-MDC can be determined by the following algorithm. (i) Determine k^* which minimizes $\mu^T_T(k)$. (ii) Check whether $\mu^l_T(k^*) \geq \mu^T_T(k^*)$, all l. If so, stop. (iii) Otherwise, check whether $\mu^l_T(k') \geq \mu^T_T(k')$, where k' is the best alternative to k^*. (iv) Continue until $\mu^l_T(k) \geq \mu^T_T(k)$, all l.

where ϵ_t satisfies: $E[\epsilon_t|x_{t-1}] = 0$. The Markov property implies that

$$E[\epsilon_t|x_{t-1}, x_{t-2}, \ldots] = E[\epsilon_t|x_{t-1}] = 0.$$

We want to construct a criterion by which to determine whether the latter is correct. In other words, we wish to develop a criterion to test whether additional lags of elements in x_t should be included in the state vector. We will focus on mean square prediction errors. This essentially means that we posit a quadratic loss function. It also justifies our focusing on the regression function only.

Construct sequences of estimates of f, $\{f_t^l; t = m, \ldots T\}$ from subsets of past information (T denotes the total sample size). The integer l indexes the subsets. The observations $r_t, \ldots, r_{\max(1,t-l+1)}$ and $x_{t-1}, \ldots, x_{\max(0,t-l)}$ constitute subset l at t. The ls are chosen from a set $\{m, m+1, \ldots, \bar{l}, T\}$ (m is the minimal sample size for which the regression estimates are uniquely defined), with $\bar{l} < \infty$. In fact, the sequences $\{f_t^l; t = m, \ldots, T\}$ form rolling regressions. Generate, subsequently, one-step-ahead forecasts. Then calculate mean square prediction errors for each sequence.

Let μ_T^l denote the mean square prediction error from the rolling regression in which all but the most recent l observations are thrown away. This means

$$\mu_T^l = \frac{1}{T-m} \sum_{t=m+1}^{T} (r_t - f_{t-1}^l(x_{t-1}))^2.$$

This mean square prediction error is an estimate of the out-of-sample forecasting performance of the model. Each error is computed on the basis of estimates that use only past information.

Assume the following.

Assumption 1.

$$\frac{1}{T-m} \sum_{t=m+1}^{T} (r_t - f(x_{t-1}))^2 \to C$$

in probability ($C < \infty$).

Assumption 2. *f_T^T converges to f in mean square.*

Assumption 1 is almost superfluous: convergence of the estimator of the regression function (Assumption 2) could hardly be obtained without convergence of the mean square regression error (Assumption 1). Notice that these assumptions put minimal restrictions on the stochastic properties of the state vector. In particular, the state vector could be unit-root nonstationary. If the regression function f is linear, we essentially require r_t and x_{t-1} to be cointegrated. Assumption 1 would immediately follow from finiteness of the second moment of the regression error. Assumption 2 would hold for the ordinary least squares estimator of f. Our criterion shares this generality (applicability to unit-root nonstationary processes) with the FIC and PIC criterion [see Wei (1992) and Phillips and Ploberger (1996)].

It is also important to emphasize that Assumptions 1 and 2 survive in many situations where the regression error is heteroskedastic. In this respect, our criterion contrasts with the FIC and PIC criterion, which were developed under particular regression error models.

In order to understand our model selection criterion, we first prove the following theorem. It states that μ_T^l is minimized, asymptotically, for $l = T$.

Theorem 1. *Let $l \in \{m, m+1, \ldots, \bar{l}\}$, where $\bar{l} < \infty$. Then, under Assumptions 1 and 2, for all l:*

$$\lim_{T\to\infty} P\{\mu_T^l - \mu_T^T > 0\} = 1.$$

(P is the probability measure that defines the probability space in which the processes x_t and r_t live.)

Proof.

$$\begin{aligned}
\mu_T^l &= \frac{1}{T-m}\sum_{t=m+1}^{T}(r_t - f_{t-1}^l(x_{t-1}))^2 \\
&= \frac{1}{T-m}\sum_{t=m+1}^{T}(r_t - f(x_{t-1}))^2 + \frac{1}{T-m}\sum_{t=m+1}^{T}(f_{t-1}^l(x_{t-1}) - f(x_{t-1}))^2 \\
&\quad - 2\frac{1}{T-m}\sum_{t=m+1}^{T}(r_t - f(x_{t-1}))(f_{t-1}^l(x_{t-1}) - f(x_{t-1})).
\end{aligned}$$

Hence,

$$\begin{aligned}
\mu_T^l - \mu_T^T &= \frac{1}{T-m}\sum_{t=m+1}^{T}(f_{t-1}^l(x_{t-1}) - f(x_{t-1}))^2 \\
&\quad - \frac{1}{T-m}\sum_{t=m+1}^{T}(f_{t-1}^T(x_{t-1}) - f(x_{t-1}))^2 \\
&\quad - 2\frac{1}{T-m}\sum_{t=m+1}^{T}(r_t - f(x_{t-1}))(f_{t-1}^l(x_{t-1}) - f(x_{t-1})) \\
&\quad + 2\frac{1}{T-m}\sum_{t=m+1}^{T}(r_t - f(x_{t-1}))(f_{t-1}^T(x_{t-1}) - f(x_{t-1})). \qquad (10)
\end{aligned}$$

Because of the Markov nature of the time series and Assumption 1, the third term will converge to zero in probability. Assumption 2 and Cesàro summability implies that the second and fourth terms converge to zero in probability as well. Finally, the probability that the first term is strictly positive is nonzero if $l \leq \bar{l} < \infty$. ■

The crucial part of the proof of Theorem 1 occurs in the third term of Equation (10):

$$-2\frac{1}{T-m}\sum_{t=m+1}^{T}(r_t - f(x_{t-1}))(f_{t-1}^l(x_{t-1}) - f(x_{t-1})). \qquad (11)$$

This expression is the average cross-product between the regression error (ϵ_t) and the error in estimating the regression function. Since $f(x_{t-1})$ depends on x_{t-1}, and $f_{t-1}^l(x_{t-1})$ depends on $r_{t-1}, \ldots, r_{\max(1,t-l)}$ and $x_{t-2}, \ldots, x_{\max(0,t-l-1)}$, the error in estimating the

regression function, $f^l_{t-1}(x_{t-1}) - f(x_{t-1})$, is a function *only* of past information. Because of the Markov assumption, the regression error is orthogonal to past information. Hence Equation (11), which is an estimate of the cross-moment between the regression error and a particular function of past information, converges to zero.

This discussion should reveal an important fact: the correctness of Theorem 1 depends crucially on the Markov assumption. In other words, it hinges on our including the correct number of lags of r_t and the right additional conditioning variables in the state vector. If our process is Markov only after including *additional* lags of elements in the state vector and/or of r_t, the regression error ϵ_t will not be orthogonal to past information. Since the estimation error $[f^l_{t-1}(x_{t-1}) - f(x_{t-1})]$ is a function of $r_{t-1}, \ldots, r_{\max(1,t-l)}$ and $x_{t-2}, \ldots, x_{\max(1,t-l-1)}$, its cross-moment with the regression error will be nonzero. Because $l < \infty$, the estimation error will *not* disappear asymptotically. Since the expression in Equation (11) is the sample version of this cross-moment, it will be nonzero for large T, and, provided the cross-moment is large enough, $\mu^l_T - \mu^T_T \leq 0$. Consequently the falseness of the Markov assumption may be revealed in large samples by inspection of the mean square prediction error of rolling regressions based on subsamples of past observations.

We immediately obtain a selection criterion. We shall call it the Markov dimension criterion (MDC), because it is designed to determine the appropriate dimension of the state vector.

MDC. *Accept the hypothesis that* $\{r_t; t = 1, 2, \ldots\}$ *is Markov relative to the state vector* $\{x_{t-1}; t = 1, \ldots\}$ *if* $T = \arg\min\{\mu^l_T\}$, *otherwise reject.*

From the derivation of the MDC, it should be clear that it may not always reveal the falseness of x_{t-1} as the state vector, whereas it will always uncover the correctness of such a model (though only asymptotically). When the correlation between the regression error and information beyond lag 1 is too weak, the sample cross-moment between the regression error and estimation error [Equation (11)] may not be able to offset the size of the estimation error [the first term in Equation (10)]. If so, the mean square prediction error remains minimized at $l = T$, and the MDC would fail to reject.

How does MDC help PLS in determining the right prediction model? As mentioned in the text, PLS tends to pick prediction vectors (x_{t-1}) that are too small. In each application, MDC can check whether this is indeed so, starting from the optimal prediction vector according to PLS, and adding predictors if required by the MDC decision rule. The resulting model selection criterion is referred to as PLS-MDC.

Further discussion of PLS-MDC can be found in Bossaerts and Hillion (1993).

Appendix C. Data Sources

Stock index: S&P500 (USA; Reuters); MSCI (others; Morgan Stanley).

Bond index: Lehman Brothers long T-bond (USA); Reuters (Australia, Belgium, Canada, France, Germany, Netherlands, Norway); MSCI (Denmark, Italy, Japan, Spain, Switzerland, UK).

Three-month cash yield: T-bill (USA; *Wall Street Journal*); Reuters (Australia, Belgium, Canada, Denmark, France, Germany, Japan, Netherlands, Norway, Switzerland); *The Economist* (Italy, Spain, UK).

Stock dividend yield: MSCI (all).

Price-earnings ratio: MSCI (all).

References

Akaike, H., 1974, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Anderson, T. W., 1984, *An Introduction to Multivariate Statistical Analysis*, New York, John Wiley & Sons.

Bossaerts, P., 1997, "A Theorem on (Certain Kinds of) Out-of-Sample Prediction Tests in Finance," working paper, California Institute of Technology.

Bossaerts, P., and P. Hillion, 1993, "Implementing Statistical Criteria to Select Return Forecasting Models: What Do We Learn?," working paper, California Institute of Technology.

Breen, W., L. R. Glosten, and R. Jagannathan, 1990, "Predictable Variations in Stock Index Returns," *Journal of Finance*, 44, 1177–1189.

Brock, W., J. Lakonishok, and B. LeBaron, 1992, "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns," *Journal of Finance*, 47, 1731–1764.

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

Fama, E. F., 1991, "Efficient Capital Markets: II," *Journal of Finance*, 46, 1575–1618.

Ferson, W., and C. Harvey, 1993, "The Risk and Predictability of International Equity Returns," *Review of Financial Studies*, 6, 527–566.

Foster, F. D., T. Smith, and R. E. Whaley, 1997, "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2," *Journal of Finance*, 52, 591–607.

Kavalieris, L., 1989, "The Estimation of the Order of an Autoregression Using Recursive Residuals and Cross-Validation," *Journal of Time Series Analysis*, 10, 271–281.

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

Lo, A. and A. C. MacKinlay, 1990, "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies*, 3, 431–467.

Pesaran, M. H., and A. Timmermann, 1995, "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance*, 50, 1201–1228.

Phillips, P. C. B., and W. Ploberger, 1996, "Posterior Odds Testing for a Unit Root with Data-Based Model Selection," *Econometrica*, 64, 381–412.

Rissanen, J., 1986a, "Order Estimation by Accumulated Prediction Errors," *Journal of Applied Probability*, 23A, 55–61.

Rissanen, J., 1986b, "Stochastic Modeling and Complexity," *Annals of Statistics*, 14, 1080–1100.

Schwarz, G., 1978, "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 416–464.

Solnik, B., 1993, "The Performance of International Asset Allocation Strategies Using Conditioning Information," *Journal of Empirical Finance*, 1, 33–56.

Sullivan, R., A. Timmermann, and H. White, 1997, "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap," working paper, University of California, San Diego.

Wei, C., 1992, "On Predictive Least Squares Principles," *Annals of Statistics*, 20, 1–42.

West, K., 1996, "Asymptotic Inference About Prediction Ability," *Econometrica*, 64, 1067–1084.