Journal of Forecasting
J. Forecast. **24**, 000–000 (2005)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/for.958

Model Uncertainty, Thick Modelling and the Predictability of Stock Returns

MARCO AIOLFI AND CARLO A. FAVERO*
Universita' Bocconi, Italy

ABSTRACT

Recent financial research has provided evidence on the predictability of asset returns. In this paper we consider the results contained in Pesaran and Timmerman (1995), which provided evidence on predictability of excess returns in the US stock market over the sample 1959–1992. We show that the extension of the sample to the nineties weakens considerably the statistical and economic significance of the predictability of stock returns based on earlier data. We propose an extension of their framework, based on the explicit consideration of model uncertainty under rich parameterizations for the predictive models. We propose a novel methodology to deal with model uncertainty based on 'thick' modelling, i.e. on considering a multiplicity of predictive models rather than a single predictive model. We show that portfolio allocations based on a thick modelling strategy systematically outperform thin modelling.

KEY WORDS ■■

1

INTRODUCTION

Recent financial research has provided ample evidence on the predictability of stock returns, identifying a large number of financial and macro variables that appear to predict future stock returns.[1] Even though financial economists and practitioners have agreed upon a restricted set of explanatory variables that could be used to forecast future stock returns, there is no agreement on the use of a single specification. Different attempts have been made to come up with a robust specification.

Pesaran and Timmermann (1995) (henceforth, P&T) consider a time-varying parameterization for the forecasting model to find that the predictive power of various economic factors over stock returns changes through time and tends to vary with the volatility of returns. They apply a 'recursive modelling' approach, according to which at each point in time all the possible forecasting models are estimated and returns are predicted by relying on the best model, chosen on the basis of some given in-sample statistical criterion. The dynamic portfolio allocation, based on the signal generated by a time-varying model for asset returns, is shown to outperform the buy-and-hold strategy over

* Correspondence to: Carlo Favero, IGIER, via Salasco 5, 20124 Milan, Italy. E-mail: carlo.faver@unibocconi.it
[1] See for example Ait-Sahalia and Brandt (2001), Avramov (2002), Bossaert and Hillion (1999), Brandt (1999), Campbell and Shiller (1988a,b), Cochrane (1999), Fama and French (1988), Keim and Stambaugh (1986), Lamont (1998), Lander *et al*. (1997), Lettau and Ludvigson (2001), Pesaran and Timmermann (1995, 2002).

the period 1959–1992. The results obtained for the USA are successfully replicated in a recent paper concentrating on the UK evidence (Pesaran and Timmermann, 2000). Following this line of research, Bossaerts and Hillion (1999) implement different model selection criteria in order to verify the evidence of the predictability in excess returns, discovering that even the best prediction models have no out-of-sample predicting power.

The standard practice of choosing the best specification according to some selection criterion can be labelled as thin modelling because a single forecast is associated with all available specifications. In reality a generic investor faced with a set of different models is not interested in selecting a best model, but to convey all the available information to forecast the $t + 1$ excess return and at the same time have a measure of the risk or uncertainty surrounding this forecast. Only at this point can the investor solve his own asset allocation problem. Since any model will only be an approximation to the generating mechanism and in many economic applications misspecification is inevitable, of substantial consequence and of an intractable nature, the strategy of choosing only the 'best' model (i.e. thin modelling) seems to be rather restrictive. If the economy features a widespread, slowly moving component that is approximated by an average of many variables through time but not by any single economic variable, then models that concentrate on parsimony could be missing it.

Furthermore, if the true process is sufficiently complex, then the reduction strategy can lead to a model ('best' according to some criterion) which is more weakly correlated with the true model than the combination of different models.

In this paper we propose a novel methodology which extends the proposal contained in the original paper by P&T to deal explicitly with model uncertainty. The remainder of the paper is organized as follows. The next section discusses our proposal to deal with model uncertainty under rich parameterization for the predictive models. The third section reassesses the original evidence on the statistical and economic significance of the predictability of stock returns by extending the data set to the nineties and by evaluating comparatively alternative modelling strategies. Then we assess the statistical and economic significance of the predictions through a formal testing procedure and their use in a trading strategy. The last section concludes by providing an assessment of our main findings.

RECURSIVE MODELLING: THIN OR THICK?

Thick modelling

P&T (1995) consider the problem of an investor allocating his portfolio between a safe asset denominated in dollars and US stocks. The decision on portfolio allocation is then completely determined by the forecast of excess returns on US stocks. Their allocation strategy is such that the portfolio is always totally allocated into one asset, which is the safe asset if predicted excess returns are negative, and shares if the predicted excess returns are positive. The authors forecast excess US stock returns by concentrating on an established benchmark set of regressors over which they conduct the search for a 'satisfactory' predictive model. They focus on modelling the decision in real time. To this end they implement a recursive modelling approach, according to which at each point in time, t, a search over a base set of observable k regressors is conducted to make a one-period-ahead forecast. In each period they estimate a set of regressions spanned by all the possible permutations of the k regressors. This gives a total of 2^k different models for excess returns. Models are estimated recursively, so that the data set is expanded by one observation in each period. Therefore, a total of 2^k models are estimated in each period from 1959:12 to 1992:11 to generate a portfolio allocation.

P&T estimate all the possible specifications of the following forecasting equation:

$$(x_{t+1} - r_{t+1}) = \beta_i' \mathbf{X}_{t,i} + \varepsilon_{t+1,i} \quad (1)$$

where $x_t + 1$ are the monthly returns on the S&P500 index and r_{t+1} are the monthly returns on the US dollar denominated safe asset (1-month T-bill), $\mathbf{X}_{t,i}$ is the set of regressors, observable at time t, included in the ith specification ($i = 1, \ldots, 2^k$) for the excess return. The relevant regressors are chosen from a benchmark set containing the dividend yield DY_t, the price–earnings ratio PE_t, the 1-month T-bill rate $I1_t$ and its lag $I1_{t-1}$, the 12-month T-bill rate $I12_t$ and its lag $I12_{t-1}$, the year-on-year lagged rate of inflation π_{t-1}, the year-on-year lagged change in industrial output ΔIP_{t-1}, and the year-on-year lagged growth rate in the narrow money stock ΔM_{t-1}. A constant is always included and all variables based on macroeconomic indicators are measured by 12-month moving averages to decrease the impact of historical data revisions on the results.[2]

At each sample point the investor computes OLS estimates of the unknown parameters for all possible models, chooses one forecast for excess returns given the predictions of $2^k = 512$ models, and maps this forecast into a portfolio allocation by choosing shares if the forecast is positive and the safe asset if the forecast is negative. P&T select in each period only one forecast, i.e. the one generated by the best model selected on the basis of a specified selection criteria which weights goodness-of-fit against parsimony of the specification (such as adjusted R^2, BIC, Akaike, Schwarz). We follow Granger (2003) and label this approach 'thin' modelling in that the forecast for excess returns and consequently the performance of the asset allocation are described over time by a thin line.

The specification procedure mimics a situation in which variables for predicting returns are chosen in each period from a pool of potentially relevant regressors according to the behaviour often observed in financial markets of attributing different emphasis to the same variables in different periods. Obviously, keeping track of the selected variables helps the reflection on the economic significance of the 'best' regression.

The main limitation of thin modelling is that model, or specification, uncertainty is not considered. In each period the information coming from the discarded $2^k - 1$ models is ignored for the forecasting and portfolio allocation exercise.

This choice seems to be particularly strong in the light of the results obtained by Bayesian research, which stresses the importance of estimation risk for portfolio allocation.[3] A natural way to interpret model uncertainty is to refrain from the assumption of the existence of a 'true' model and attach instead probabilities to different possible models. This approach has been labelled 'Bayesian model averaging'.[4] Bayesian methodology reveals the existence of in-sample and out-of-sample predictability of stock returns, even when commonly adopted model selection criteria fail to demonstrate out-of-sample predictability.

The main difficulty with the application of Bayesian model averaging to problems like ours lies with the specification of prior distributions for parameters in all 2^k models of interest. Recently, Doppelhofer *et al*. (2000) have proposed an approach labelled 'Bayesian averaging of classical estimates' (BACE), which overcomes the need for specifying priors by combining the averaging of esti-

[2] See our data appendix for further details.
[3] See, for example, Barberis (2000), Kandel and Stambaugh (1996).
[4] For recent surveys of the literature about Bayesian model selection and Bayesian model averaging see respectively Chipman *et al*. (2001) and Hoeting *et al*. (1999). Avramov (2002) provides an interesting application.

mates across models, a Bayesian concept, with classical OLS estimation, interpretable in the Bayesian camp as coming from the assumption of diffuse, non-informative, priors.

In practice, BACE averages parameters across all models by weighting them proportionally to the logarithm of the likelihood function corrected for the degrees of freedom, using then a criterion similar to the Schwarz model selection criterion. It is important to note that the consideration of model uncertainty in our context generates potential for averaging at two different levels: averaging across the different predicted excess returns and averaging across the different portfolio choices driven by the excess returns.

There is also a vast literature[5] about forecast combination showing that combining in general works.

All forecasting models can be interpreted as a parsimonious representation of a general unrestricted model (GUM). Such approximations are obtained through the reduction process, which shrinks the GUM towards the local DGP (LDGP).[6] White has shown that if the LDGP is contained in the GUM, then asymptotically the reduction process converges to the LDGP. However, there is the possibility that the LDGP is only partially contained in the GUM or completely outside the GUM. In this case the reduction procedure will converge asymptotically to a model that is closest to the true model, according to some distance function. As pointed out by Granger and Jeon (2003), there are good reasons for thinking that the thin modelling approach may not be a good strategy because a remarkable amount of information is lost. There are also a few recent results (Stock and Watson, 1999; Giacomini and White, 2003) suggesting that some important features of the data, as measured in terms of forecast ability, can be lost in the reduction process. In fact, if the true DGP is quite complex, then the reduction process can lead to a model ('best' model) which contains less of the true model than the combination of different models. As pointed out by Granger (2003), it seems the economy might contain a widespread, slowly moving component that is approximated by an average of many variables through time but not by any single, economic variable, like a slow swing in the economy. If so, models that concentrate on parsimony could be missing this component.

This simple insight motivates the pragmatic idea of forecast combination, in which forecasts based on different models are the basic object of analysis. Forecast combination can be viewed as a key link between the short-run, real-time forecast production process, and the longer-run, ongoing process of model development. Furthermore, in a large study of structural instability, Stock and Watson (1996) report that a majority of macroeconomic time series models undergo structural change, suggesting another argument for not relying on a single forecasting model. Finally, another advantage of this approach is that a process, potentially non-linear, is linearized by looking at the linear specifications as Taylor expansions around different points.

The explicit consideration of estimation risks naturally generates 'thick' modelling, where both the prediction of models and the performance of the portfolio allocations over time are described by a thick line to take account of the multiplicity of models estimated. The thickness of the line is a direct reflection of the estimation risk.

Pesaran and Timmermann show that thin modelling allows us to outperform the buy-and-hold strategy. Re-evaluating their results from a thick modelling perspective raises immediately one ques-

[5] An incomplete list includes Chan *et al*. (1999), Clemen (1989), Diebold and Pauly (1987), Elliott and Timmermann (2002), Giacomini and White (2002), Granger (2002), Clements and Hendry (2001), Marcellino (2002), Stock and Watson (2001, 2003).

[6] An overview of the literature, and the developments leading to general-to-specific (Gets) modelling in particular, is provided by Campos *et al*. (2003).

tion: '*why choose just one model to forecast excess returns?*'. In the next section we reassess the evidence in P&T by using three different testing procedures of the performance of various forecasting models. We provide an empirical evaluation of the comparative performance of thin and thick modelling and address the issue of how to convey all the available information into a trading rule.

A FIRST LOOK AT THE EMPIRICAL EVIDENCE

We start be replicating[7] the exercise in P&T by using the same data set and by extending their original sample to 2001, keeping track of all the forecasts produced by taking into account the $2^k - 1$ combinations of regressors in a predictive model for US excess returns (the time series of this variable is reported in Figure 1). We do so by looking at the within-sample econometric performance, at the out-of-sample forecasting performance and at the performance of the portfolio allocation.

Figure 2 allows us to analyse the within-sample econometric performance by reporting the $\overline{R}^2$ for 2^k models estimated recursively. The difference in the selection criterion across different models is small, and almost negligible for models ranked close to each other.

We assess the forecasting performance of different models by using three types of tests: the Pesaran–Timmermann (1995) sign test, the Diebold–Mariano (1995) test and the White (2000) reality check. All tests and their implementations are fully described in an appendix. The P&T sign test is an out-of-sample test of predictive ability, based on the proportion of times that the sign of a given variable is correctly predicted by the sign of some predictor. The Diebold–Mariano (1995) test is

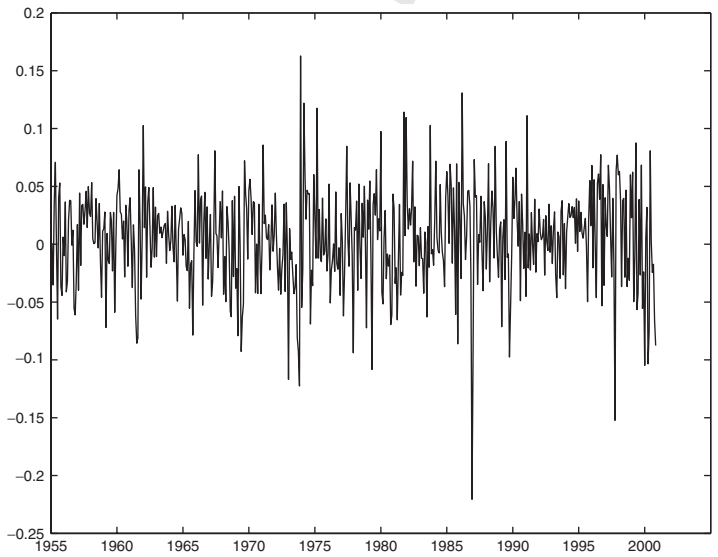

Figure 1. Excess return on S&P500. Sample 1955–2001

[7] In fact, we replicate the allocation results in the case of no transaction costs. Transaction costs do not affect the portfolio choice in the original exercise, therefore they do not affect the mapping from the forecasts to the portfolio allocation, which is the main concern of our paper.

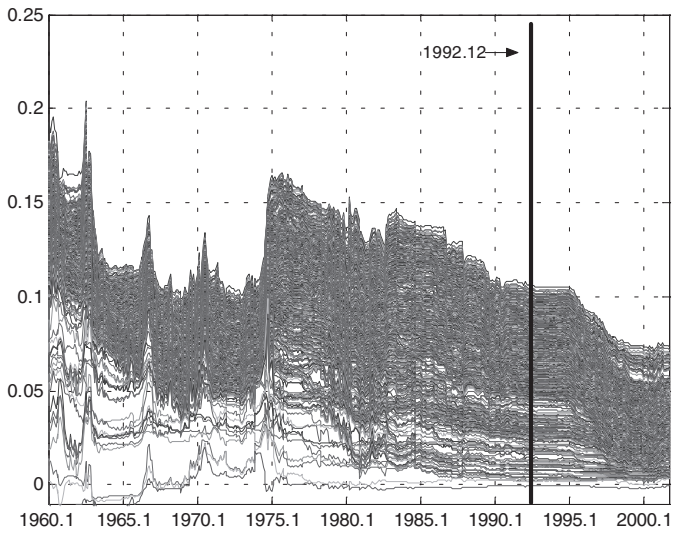

Figure 2. The figure reports the panel of the time-varying adjusted R^2 for the 2^k available models estimated recursively. The first observation refers to the smallest sample (1954.1–1959.12), the last observation refers to the full sample (1954.1–2001.8). The vertical line in 1992.12 shows the results for the P&T sample

testing the null of a zero population mean loss differential between two forecasts. We use this test to evaluate the forecasting performance of thin modelling against several thick modelling alternatives. Finally, we implement the bootstrap reality check by White (2000), based on the consistent critical values given by Hansen (2001), to test the null that our benchmark (thin) model performs better than other available forecasting (thick) models. Importantly, this testing procedure allows us to take care of the possibility of data-snooping. We report the outcomes of the tests applied to the recursive modelling proposed by P&T in Table I.

We consider the whole sample 1959–2001 and we also split it into four decades. We compare the thin modelling, labelled as best (in terms of its adjusted R^2) with several thick modelling alternatives. We label *Top x%*, the forecast obtained by averaging over the top *x%* models, ranked according to their adjusted R^2. The line labelled *All* contains the results of averaging across all 2^k models. We then label *Median*, the forecast obtained by considering the median of the empirical distribution of the within-sample performance. Lastly, we consider in the line *Dist* a synthetic measure of the skewness of this empirical distribution; in this case the selected prediction is that indicated by the majority of the models considered, independently from their ranking in terms of the within-sample performance. In general all tests show that it is possible to improve on the performance of the best model in terms of R^2 by using the information contained in the $2^k - 1$ models dominated (in many cases marginally) in terms of R^2. The sign test for the full sample shows that the thin modelling is always dominated by some thick modelling alternative. When different decades are considered, we observe that the percentage of correctly predicted signs is always significant for thick modelling in the three decades 1960–1970, 1970–1980 and 1980–1990, while the thin modelling alternative does not deliver a statistically significant value in the decade from 1980 to 1990. Interestingly, the decade 1990–2000 is an exception in that none of the strategies adopted delivers a statistically significant predictive performance. The evidence of the P&T tests is confirmed by the Diebold and Mariano

Table I. Forecasting performance of thin versus thick modelling. The results are based on recursive least squares estimation with the constant term as the only focal variable. The Pesaran–Timmermann marker-timing test (PT) is the percentage of times that the sign of the realized excess returns is correctly predicted by the forecast combination strategy reported by rows. The Diebold and Mariano (DM) test statistic is used to test the null of equal predictive ability between thin and different versions of thick modelling. The White bootstrap reality check (RC) is used to test the null that the in-sample best model performs better than all the other available forecasting models. **,* indicate significance at the 1% and 5% levels, respectively. For RC we report the *p*-value

	PT	DM	RC	PT	DM	RC
	Panel A: 1960–1970			Panel B: 1970–1980		
Best	0.57			0.62**		
Top 1%	0.57	–1.20	0.00	0.62**	–0.73	0.00
Top 5%	0.56	–0.82	0.00	0.63**	–0.20	0.00
Top 10%	0.56	–1.08	0.00	0.63**	–0.24	0.00
Top 20%	0.56	–0.85	0.00	0.61**	–0.65	0.00
Top 30%	0.57	–1.03	0.01	0.63**	–0.58	0.01
Top 40%	0.58*	–1.04	0.03	0.60*	–0.83	0.03
Top 50%	0.59*	–1.13	0.03	0.60*	–0.99	0.04
Top 60%	0.58*	–1.19	0.06	0.60*	–0.98	0.06
Top 70%	0.58*	–1.14	0.07	0.61**	–1.08	0.07
Top 80%	0.58*	–1.02	0.10	0.60*	–10.7	0.10
Top 90%	0.58*	–0.96	0.13	0.59*	–1.00	0.12
All	0.57	–0.98	0.16	0.58*	–0.88	0.13
Median	0.57		0.14	0.60*		0.13
Dist	0.57		0.00	0.60*		0.00
	Panel C: 1980–1990			Panel D: 1990–2000		
Best	0.57			0.48		
Top 1%	0.57	1.11	0.00	0.49	0.33	0.12
Top 5%	0.58	–0.77	0.00	0.46	0.84	0.31
Top 10%	0.59	–1.31	0.00	0.46	1.51	0.39
Top 20%	0.60*	–1.28	0.00	0.47	1.81	0.42
Top 30%	0.62*	–1.43	0.02	0.46	1.85	0.42
Top 40%	0.64**	–1.34	0.03	0.47	1.68	0.41
Top 50%	0.64**	–1.33	0.05	0.49	1.44	0.41
Top 60%	0.64**	–1.32	0.06	0.48	1.11	0.40
Top 70%	0.64**	–1.31	0.07	0.48	0.89	0.39
Top 80%	0.63**	–1.29	0.08	0.48	0.62	0.39
Top 90%	0.62**	–1.22	0.09	0.47	0.26	0.41
All	0.62*	–1.16	0.11	0.47	–0.22	0.41
Median	0.62*		0.10	0.45		0.41
Dist	0.62*		0.00	0.45		0.00

Table I. *Continued*

	PT	DM	RC
	Panel E: 1960–2001		
Best	0.56*		
Top 1%	0.56*	–1.67	0.00
Top 5%	0.55*	–5.21**	0.00
Top 10%	0.55*	–5.35**	0.00
Top 20%	0.55*	–6.21**	0.00
Top 30%	0.56**	–6.37**	0.00
Top 40%	0.57**	–6.57**	0.00
Top 50%	0.57**	–6.46**	0.01
Top 60%	0.57**	–6.24**	0.01
Top 70%	0.57**	–6.02**	0.01
Top 80%	0.57**	–5.79**	0.01
Top 90%	0.56**	–5.57**	0.02
All	0.56**	–5.09**	0.03
Median	0.55**		0.02
Dist	0.55**		0.00

tests. All the observed values for the statistics implemented on the full sample are negative and significant, showing that the null of equal predictive ability of thin and thick modelling is rejected, at the 1% level, independently from the adopted thick modelling specification. Such evidence is considerably weakened when the sample is split into decades. Finally, the reported *p*-values for the White reality check show that the null that all the alternative thick modelling strategies are not better than the thin model is consistently rejected when the full sample is considered. Splitting the sample into decades weakens the results only for the period 1990–2000.

The results of the forecasting performance are confirmed by the performance of the portfolio allocation. We report in Figure 3 the cumulative end-of-period wealth delivered by the portfolios associated with all 512 possible models, ranked in terms of their $\overline{R}^2$. Following P&T, portfolios are always totally allocated into one asset, which is the safe asset if predicted excess returns are negative, and shares if the predicted excess returns are positive. We add as a benchmark the final wealth given by the buy-and-hold strategy. Figure 3 shows that in general the value of the end-of-period wealth is not a decreasing function of the $\overline{R}^2$, and that the buy-and-hold strategy is in general dominated, again with the notable exception of the decade 1990–2000, where the buy-and-hold strategy gives the highest wealth.

To sum up, our evidence suggests that thick modelling dominates thin modelling but also that the evidence for excess return predictability is considerably weaker in the period 1990–2000.[8] In fact, over this sample, the adjusted R^2 of all models decreases substantially, the sign tests for predictive performance are not significant any more, and the econometric performance-based portfolio allocation generates lower wealth than the buy-and-hold strategy.

[8] This is also observed by Paye and Timmermann (2002).

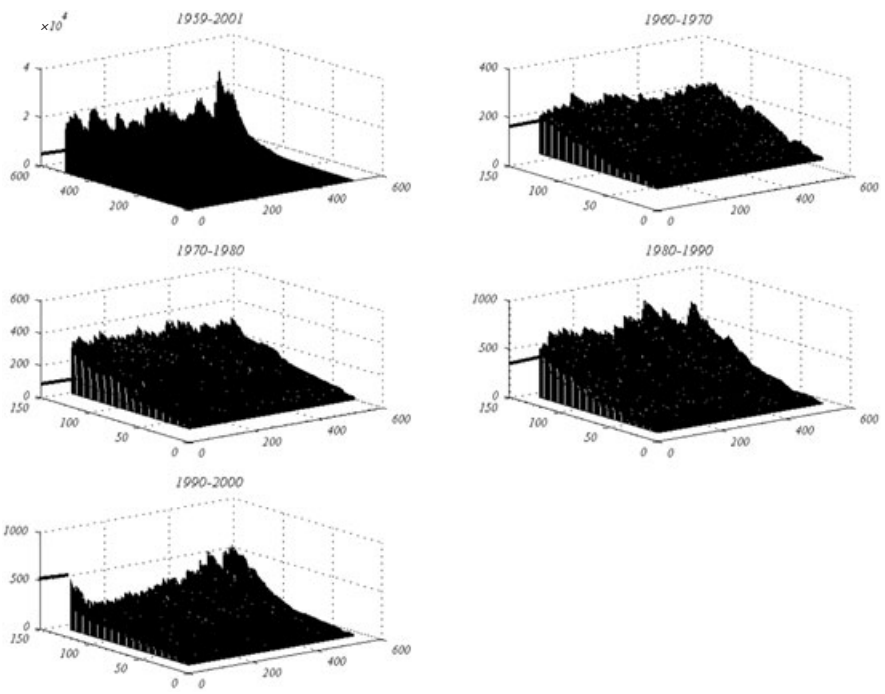

Figure 3. Cumulative wealth obtained from all possible portfolio allocations. Allocations are associated with models ranked according to their adjusted R^2. The thick line pins down the final wealth delivered by the buy-and-hold strategy

In the next section we shall evaluate refinements in the specification and the modelling selection strategy in the spirit of thick modelling.

OUR PROPOSAL FOR THICK MODELLING

In the light of the evidence reported in the previous section we propose extensions of the original methodology both at the stage of model specification and of portfolio allocation.

The empirical evidence reported in the previous section shows clearly that the ranking of models in terms of their within-sample performance does not match at all the ranking of models in terms of their *ex post* forecasting power. This empirical evidence points clearly against BACE using within-sample criteria to weight models. Consistent with this evidence, we opted for the selection method proposed by Granger and Jeon (2003) of using a '. . . *procedure [which] emphasizes the purpose of the task at hand rather than just using a simple statistical pooling* . . .'. Our task at hand is asset allocation.

Model specification

At the stage of model specification we consider two issues: the importance of balanced regressions and the optimal choice of the window of observations for estimation purposes.

A regression is balanced when the order of integration of the regressors matches that of the dependent variables. Excess returns are stationary, but not all variables are candidate to explain that stationarity. To achieve a balanced regression in this case, cointegration among the included non-stationary variables is needed. As shown by Sims *et al*. (1990) the appropriate stationary linear combinations of non-stationary variables will be naturally selected by the dynamic regression, when all non-stationary variables potentially included in a cointegrating relation are included in the model. Therefore, when model selection criteria are applied, one must make sure that such criteria do not lead us to exclude any component of the cointegrating vector from the regression. Following Pesaran and Timmermann (2001) we divide variables into focal, labelled A_t and secondary focal, labelled B_t. Focal variables are always included in all models, while the variables in B_t are subject to the selection process. We take these variables as those defining the long-run equilibria for the stock market. Following the lead of traditional analysis[9] and recent studies (Lander *et al*., 1997), we have chosen to construct an equilibrium for the stock market by concentrating on a linear relation between the long-term interest rates, R_t, and the logarithm of the earning price ratio, *ep*. Also, recent empirical analysis (see Zhou, 1996) finds that stock market movements are closely related to shifts in the slope of the term structure. Such results might be explained by a correlation between the risk premia on long-term bonds and the risk premium on stocks. Therefore, we consider the term spread as a potentially important cointegrating relation. On the basis of this consideration we include in the set of focal variables the yield to maturity on 10-year government bonds (a variable which was not included in the original set of regressors in P&T), the log of the earning price ratio and the interest rate on 12-month Treasury bills, to ensure that the selected model is balanced and includes the two relevant cointegrating vectors. We do not impose any restrictions on the coefficients of the focal variables.[10]

The second important issue at the stage of model selection is the choice of the window of observations for estimation (i.e. for how long a predictive relationship stays in effect).[11] The question of stability is equally important since the expected economic value from having discovered a good historical forecasting model is much smaller if there is a high likelihood of the model breaking down subsequently.

---

[9] '. . . Theoretical analysis suggests that both the dividend yield and the earnings yield on common stocks should be strongly affected by changes in the long-term interest rates. It is assumed that many investors are constantly making a choice between stock and bond purchases; as the yield on bonds advances, they would be expected to demand a correspondingly higher return on stocks, and conversely as bond yields decline . . .'. (*Graham and Dodd Security Analysis*, 4th edition, 1962, p. 510)

The above statement suggests that either the dividend yield or the earnings yield on common stocks could be used.

[10] We have assessed the choice of our focal variable by estimating recursively a VAR including the yield to maturity of 10-year government bonds, the log of the earning–price ratio and the interest rate on 12-month Treasury bills. The null of no cointegration is always rejected when the Johansen (1995) procedure is implemented by allowing for an intercept in the cointegrating vectors. We choose not to impose any restriction on the number of cointegrating vectors and on cointegrating parameters as they are not constant over time (a full set of empirical results is available upon request).

[11] Recent empirical studies cast doubt upon the assumed stability in return forecasting models. An incomplete list includes Ang and Bekaert (2001), Lettau and Ludvigson (2001), Paye and Timmermann (2002).

In the absence of breaks in the DGP the usual method for estimation and forecasting is to use an expanding window. In this case, by augmenting an already selected sample period with new observations, more efficient estimates of the same fixed coefficients are obtained by using more information as it becomes available. However, if the parameters of the regression model are not believed to be constant over time, a rolling window of observations with a fixed size is frequently used. When a rolling window is used, the natural issue is the choice of its size. This problem has already been observed by Pesaran and Timmermann (2002), who provide an extensive analysis of model instability, structural breaks and the choice of window observations. In line with their analysis we deal with the problem of window selection by starting from an expanding window, every time a new observation is available we run a backward CUSUM and CUSUM squared test to detect instability in the intercept and/or in the variance. We then keep expanding the window only when the null of no structural break is not rejected. Consider a sample of T observations and the following model:

$$y_{t,T} = \beta^{i'} x^i_{t,T} + \mu_{t,T} \quad i = 1, \ldots, 2^k$$

where $y_{t,T} = (y_t, y_t, y_{t+2}, \ldots, y_T)$ and $x^i_{t,T} = (x^i_t, x^i_{t+1}, x^i_{t+2}, \ldots, x^i_T)$ where $T - t + 1$ is the optimal window and T the last available observation. Recall that we are interested in forecasting y_{T+1} given $x_{T+1}, \hat{\beta}^{i'}$. The problem of the optimal choice of t given model i can be solved by running a CUSUM test with the order of the observations reversed in time starting from the mth observation and going back to the first observation available (we refer to this procedure as ROC). Critical values by Brown *et al*. (1975) can be used to decide if a break has occurred. Unlike the Bai–Perron method, the ROC method does not consistently estimate the breakpoint.[12] On the other hand, the simpler look-back approach only requires detecting a single break and may succeed in determining the most recent breakpoint in a manner better suited for forecasting. Once a structural break (either in the mean or in the variance) has been detected, we have found the appropriate t. Clearly the appropriate t can be the first observation in the sample (in this case we have an expanding window) or any number between 1 and m (flexible rolling window). This procedure allows us to optimally select the observation window[13] for each of the 2^k different models estimated at time t.

In terms of model selection we now have several methodologies available: the original P&T recursive estimation (based on an expanding window of observations) with no division of variables into focal and semi-focal, the rolling estimation (based on a fixed window of 60 observations) with no division of variables into focal and semi-focal, the balanced recursive estimation, in which variables are divided into focal and non-focal, to make sure that cointegrating relationship(s) are always included in the specification, and a flexible estimation, in which the optimal size for the estimation window is chosen for all possible samples. We consider two versions of the flexible estimation that differ by the division of variables into focal and semi-focal.

Asset allocation

To analyse how the value of the investor's portfolio evolves through time, we first introduce some notations. Let W_t be the funds available to the investor at the end of period t, α^S_t the number of shares

[12] As pointed out by Pesaran and Timmermann (2002), ironically this may well benefit the ROC method in the context of forecasting since it can be optimal to include pre-break data in the estimation of a forecasting model. Although doing so leads to biased predictions, it also reduces the parameter estimation uncertainty.

[13] We impose that the shortest observation window automatically selected cannot be smaller than 2 or 3 times the dimension of the parameters' vector. So also the minimum observation window is a function of regressors included in each of 2^k different models.

held at the end of period t, r_t^s the rate of return on S&P500 and r_t^b the rate of return on safe assets in period t, S_t and B_t the investor's position in stock and safe assets at the end of period t, respectively. At a particular point in time, t, the budget constraint of the investor is given by:

$$W_t = (1 + r_{t-1}^s)S_{t-1} + (1 + r_{t-1}^b)B_{t-1}$$

P&T propose an allocation strategy such that the portfolio is always totally allocated into one asset, which is the safe asset if predicted excess returns are negative, and shares if the predicted excess returns are positive. We consider three alternative ways of implementing thick modelling when allocating portfolios. Given the 2^k forecasts for excess returns in each period, define α_t^S and $\alpha_t^B = (1 - \alpha_t^S)$ to be respectively the weight on stocks and the safe asset (short-term bills), let $\{y_i\}_{i=1}^{2k}$ be the full set of excess return forecasts obtained in the previous step, and let $n = \omega' 2^k$, where $\omega = [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$ is the set of weights, in terms of the percentage of the model ordered according to their adjusted R^2, chosen to build up the appropriate trimmed means of the available forecasts. Then we propose the following allocation criteria.

Distribution thick modelling. We look at the empirical distribution of the forecasts to apply the following criterion:

$$\alpha_{\omega_i}^S = \begin{cases} 1 & \text{if } \left[\dfrac{\sum_{i=1}^{n_{\omega_j}} (y_i > 0)}{n_{\omega_i}} \right] > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

where $n_{\omega j}\,(y_i > 0)$ is the number of models giving a positive prediction for excess returns within the jth class of the trimming grid (for example $n_{\omega_2}\,(y_i > 0)$ is the number of models in the best 5% of the ranking in terms of their $\overline{R}^2$ predicting a positive excess return). In practice if more than 50% of the considered models predict an upturn (downturn) of the market, we put all the wealth in the stock market (safe asset).

Meta thick modelling. We use the same criterion as above to derive a less aggressive portfolio allocation, in which corner solutions are the exception rather than the rule:

$$\alpha_{\omega_i}^S = \left[\frac{\sum_{i=1}^{n_{\omega i}} (y_i > 0)}{n_{\omega_i}} \right]$$

Kernel thick modelling. We compute the weighted average of predictions $\bar{y}$ (with weights based on the relative adjusted R^2, through a triangular kernel function that penalizes deviations from the best model in terms of R^2 and the bandwidth determined by the number of observations) and then we apply this rule:

$$\alpha_{\omega_i}^S = \begin{cases} 1 & \text{if } \bar{y} > 0 \\ 0 & \text{otherwise} \end{cases}$$

EMPIRICAL RESULTS

Our empirical results are reported in Tables II–IV and Figures 3–5.

Table II. Pesaran–Timmermann markt-timing test of thin and thick modelling excess return forecasts. Each panel reports the proportion of times that in a given sample the sign of realized excess returns is correctly predicted by the sign of alternative thin and thick modelling one-step-ahead forecasts generated by five different estimation strategies. **·* indicate significant evidence of market timing at the 5% and 10% levels, respectively. Top *x*% is the combination of the trimmed mean of the best *x*% forecasting models, Med is the combination scheme based on the median, and Dist is the combination scheme based on the majority rule applied to all the available forecasting models. REC, ROLL, OW denote recursive estimation, rolling estimation with fixed window length, optimal estimation window, respectively. The numbers in square brackets shown the number of focal variables considered. [1] is just the eonstant, while [4] denotes the following set of regressors: constant, log of the price–earnings ratio, yield-to-maturity on long-term bonds, yield on 12-month Treasury bills

	REC [1]	ROLL [1]	REC [4]	OW [1]	OW [4]	REC [1]	ROLL [1]	REC [4]	OW [1]	OW [4]
	Panel A: 1960–1970					Panel B: 1970–1980				
Best	0.57	0.55	0.53	0.50	0.54	0.62**	0.51	0.57	0.57	0.56
Top 1%	0.57	0.54	0.52	0.53	0.52	0.62**	0.52	0.58*	0.56	0.55
Top 5%	0.56	0.55	0.53	0.53	0.53	0.63**	0.52	0.57	0.57	0.57
Top 10%	0.56	0.57	0.53	0.51	0.53	0.63**	0.53	0.57	0.57	0.59*
Top 20%	0.56	0.57	0.53	0.54	0.54	0.61**	0.49	0.57	0.59*	0.54
Top 30%	0.57	0.55	0.55	0.53	0.54	0.63**	0.51	0.57	0.59*	0.55
Top 40%	0.58*	0.56	0.56	0.57	0.54	0.60*	0.53	0.54	0.62**	0.54
Top 50%	0.59*	0.56	0.55	0.57	0.54	0.60*	0.53	0.55	0.60*	0.54
Top 60%	0.58*	0.56	0.54	0.56	0.53	0.60*	0.54	0.56	0.61**	0.53
Top 70%	0.58*	0.57	0.57	0.57	0.52	0.61**	0.57*	0.57	0.61**	0.54
Top 80%	0.58*	0.57	0.53	0.56	0.53	0.60*	0.55	0.54	0.60*	0.54
Top 90%	0.58*	0.57	0.53	0.55	0.54	0.59*	0.56	0.54	0.60*	0.55
All	0.57	0.56	0.54	0.55	0.55	0.58*	0.56	0.55	0.59*	0.55
Median	0.57	0.53	0.55	0.57	0.55	0.60*	0.57	0.54	0.61**	0.53
Dist	0.57	0.53	0.55	0.57	0.55	0.60*	0.57	0.54	0.61**	0.53
	Panel C: 1980–1990					Panel D: 1990–2000				
Best	0.57	0.57*	0.59	0.57*	0.53	0.48	0.49	0.50	0.48	0.46
Top 1%	0.57	0.58*	0.60*	0.56*	0.54	0.49	0.51	0.51	0.47	0.47
Top 5%	0.58	0.59*	0.59	0.59*	0.57	0.46	0.52	0.50	0.45	0.49
Top 10%	0.59	0.60*	0.61*	0.60*	0.59*	0.46	0.53	0.50	0.44	0.47
Top 20%	0.60*	0.59*	0.59	0.61**	0.57	0.47	0.51	0.47	0.50	0.46
Top 30%	0.62*	0.60**	0.61*	0.60**	0.57	0.46	0.53	0.48	0.51	0.48
Top 40%	0.64**	0.61**	0.62*	0.60**	0.59*	0.47	0.54	0.48	0.47	0.46*
Top 50%	0.64**	0.63**	0.62*	0.60**	0.59*	0.49	0.53	0.48	0.49	0.49
Top 60%	0.64**	0.61**	0.60*	0.57*	0.58	0.48	0.55	0.47	0.51	0.52
Top 70%	0.64**	0.63**	0.60*	0.60**	0.59*	0.48	0.57	0.45	0.52	0.52
Top 80%	0.63**	0.63**	0.60*	0.63**	0.57	0.48	0.57	0.45	0.49	0.53
Top 90%	0.62**	0.64**	0.58	0.66**	0.59*	0.47	0.58	0.47	0.57	0.55
All	0.62*	0.63**	0.59*	0.66**	0.60	0.47	0.57	0.48	0.57	0.57
Median	0.62*	0.65**	0.55	0.63**	0.57	0.45	0.59	0.51	0.56	0.58
Dist	0.62*	0.65**	0.55	0.63**	0.57	0.45	0.59	0.51	0.56	0.58

Table II. *Continued*

	REC [1]	ROLL [1]	REC [4]	OW [1]	OW [4]
			Panel E: 1960–2001		
Best	0.56*	0.54	0.55*	0.53*	0.53
Top 1%	0.56*	0.55*	0.55*	0.53*	0.52
Top 5%	0.55*	0.55*	0.55*	0.54*	0.54**
Top 10%	0.55*	0.56*	0.55*	0.53	0.55**
Top 20%	0.55*	0.54	0.54*	0.56**	0.53
Top 30%	0.56**	0.54*	0.55**	0.55*	0.53*
Top 40%	0.57**	0.55*	0.55**	0.56**	0.53
Top 50%	0.57**	0.55*	0.55**	0.56**	0.54
Top 60%	0.57**	0.55*	0.54*	0.55*	0.54
Top 70%	0.57**	0.57**	0.54*	0.56**	0.54
Top 80%	0.57**	0.57**	0.53	0.56**	0.54
Top 90%	0.56**	0.57**	0.53	0.57**	0.55*
All	0.56**	0.56*	0.54*	0.58**	0.56**
Median	0.55**	0.57**	0.53	0.57**	0.56*
Dist	0.55**	0.57**	0.53	0.57**	0.56*

In Tables II–IV we evaluate the forecasting performance of all methodologies by using our three testing procedures.

In Table II we report the results of the Pesaran–Timmermann market-timing test of thin and thick modelling excess return forecasts, in Table III we report the results of the Diebold–Mariano test of equal predictive ability between thin and thick modelling excess return forecasts, and finally in Table IV we report the results for White's reality check to test the null that thin modelling-based forecasts outperform thick modelling-based forecasts.

Overall, all three tests suggest that the flexible estimation delivers the best results. The most remarkable improvements occur when the Diebold–Mariano and White's reality check are implemented over the decade 1990–2000. The P&T sign test confirms the results of the other two tests but also signals that the null that any chosen predictor has no power in predicting excess returns over the decade 1990–2000 cannot be rejected.

On the basis of this evidence we proceed to evaluate the performance of asset allocation based on thin and thick modelling, considering the buy-and-hold strategy as a benchmark.

Figures 4 and 5 evaluate the performance of different portfolio allocation criteria, by comparing the end-of-period cumulative wealth associated with the recursive estimation and the rolling estimation with optimally chosen window and focal regressors with the cumulative wealth associated with a simple buy-and-hold strategy.[14] Each figure considers an estimation criterion and reports the performance of portfolio allocations for the thin modelling approach and different types of thick modelling along with the buy-and-hold strategy. We report, for the full sample and for the four decades, the end-of-period wealth associated with a beginning-of-period wealth of 100.

With very few exceptions thick modelling dominates thin modelling. Moreover, the more articulated model specification procedures deliver better results than the simple recursive criterion. The best performance is achieved when the distribution thick modelling is applied to the best 20% of

[14] Evaluation has also been conducted in terms of period returns and Sharpe ratios; results are available upon request.

Table III. Diebold–Mariano test of equal predictive ability between thin and thick modelling excess return forecasts. Each panel reports the proportion of times that in a given sample the sign of realized excess returns is correctly predicted by the sign of alternative thin and thick modelling one-step-ahead forecasts generated by five different estimation strategies. **·* indicate significant evidence of market timing at the 5% and 10% levels, respectively. Top *x*% is the combination of the trimmed mean of the best *x*% forecasting models. REC, ROLL, OW denote recursive estimation, rolling estimation with fixed window length, optimal estimation window, respectively. The numbers in square brackets show the number of focal variables considered. [1] is just the constant, while [4] denotes the following set of regressors: constant, log of the price–earnings ratio, yield-to-maturity on long-term bonds, yield on 12-month Treasury bills

	REC [1]	ROLL [1]	REC [4]	OW [1]	OW [4]	REC [1]	ROLL [1]	REC [4]	OW [1]	OW [4]
	Panel A: 1960–1970					Panel B: 1970–1980				
Top 1%	–1.19	–0.29	0.04	–1.88	0.01	–0.73	–1.66	0.36	–0.92	0.19
Top 5%	–0.82	–1.18	–0.51	–2.66*	–0.92	–0.20	–1.97	0.35	–3.05**	–1.19
Top 10%	–1.08	–1.65	–0.84	–2.49	–1.08	–0.24	–2.58	0.48	–3.37**	–1.63
Top 20%	–0.84	–2.00	–1.13	–2.57*	–1.27	–0.65	–3.30*	0.25	–2.93*	–1.70
Top 30%	–1.02	–2.23*	–1.48	–2.66	–1.37	–0.57	–3.73**	–0.27	–2.80**	–1.69
Top 40%	–1.04	–2.36*	–1.65	–2.67*	–1.48	–0.83	–3.80**	–0.07	–2.79*	–1.62
Top 50%	–1.13	–2.41*	–1.65	–2.65*	–1.54	–0.98	–3.87**	0.40	–2.79*	–1.75
Top 60%	–1.18	–2.46*	–1.61	–2.62	–1.58	–0.98	–3.81**	0.49	–2.78*	–1.81
Top 70%	–1.13	–2.51**	–1.52	–2.58*	–1.65	–1.08	–3.71**	0.51	–2.78*	–1.87
Top 80%	–1.02	–2.51*	–1.44	–2.55*	–1.67	–1.06	–3.66**	0.57	–2.78**	–1.84
Top 90%	–0.96	–2.47*	–1.39	–2.54*	–1.73	–1.00	–3.59**	0.60	–2.72*	–1.79
All	–0.97	–2.45*	–1.38	–2.58*	–1.72	–0.88	–3.52**	0.69	–2.66**	–1.73
	Panel C: 1980–1990					Panel D: 1990–2000				
Top 1%	1.10	–1.04	0.50	–0.60	–0.61	0.33	0.45	0.92	–0.02	–2.31
Top 5%	–0.76	–2.20*	0.53	–2.21*	–0.17	0.84	–1.26	–1.22	–1.21	–2.88*
Top 10%	–1.30	–2.91**	–0.26	–2.28*	–0.26	1.51	–2.10	–0.86	–1.70	–3.29*
Top 20%	–1.28	–3.32**	–0.96	–2.24*	–0.49	1.80	–2.75**	–0.87	–2.04*	–3.20**
Top 30%	–1.42	–3.47*	–0.69	–2.21*	–0.93	1.84	–2.93**	–1.26	–2.15*	–3.70**
Top 40%	–1.34	–3.93**	–0.61	–2.27*	–1.57	1.67	–3.03**	–1.40	–2.32*	–3.98**
Top 50%	–1.33	–4.11**	–0.50	–2.32*	–2.10*	1.44	–3.07*	–1.51	–2.40*	–4.00**
Top 60%	–1.32	–4.25**	–0.45	–2.30*	–2.29*	1.11	–3.12**	–1.57	–2.40*	–4.02**
Top 70%	–1.30	–4.26**	–0.35	–2.31	–2.44*	0.88	–3.21**	–1.61	–2.39*	–4.02**
Top 80%	–1.29	–4.18**	–0.32	–2.29*	–2.38*	0.62	–3.28**	–1.62	–2.41*	–3.99**
Top 90%	–1.21	–4.06**	–0.37	–2.29*	–2.38*	0.26	–3.29**	–1.63	–2.42*	–3.95**
All	–1.16	–3.96**	–0.43	–2.26*	–2.39*	–0.22	–3.29**	–1.61	–2.40*	–3.92**
	Panel E: 1960–2001									
Top 1%	–1.67	–1.42	–0.26	0.29	–0.29					
Top 5%	–5.21**	–2.21*	–1.86	–2.36*	–0.24					
Top 10%	–5.34**	–2.98**	–1.72	–2.17*	–0.37					
Top 20%	–6.21**	–3.58**	–3.13**	–1.93	–0.56					
Top 30%	–6.36**	–3.79**	–3.41**	–1.75	–0.66					
Top 40%	–6.57**	–4.08**	–3.13**	–1.75	–0.90					
Top 50%	–6.45**	–4.09**	–2.95**	–1.77	–1.14					
Top 60%	–6.23**	–3.88**	–3.06**	–1.73	–1.29					
Top 70%	–6.01**	–3.62**	–3.00**	–1.70	–1.47					
Top 80%	–5.79**	–3.42**	–2.93**	–1.64	–1.44					
Top 90%	–5.56**	–3.21**	–2.85**	–1.51	–1.35					
All	–5.09**	–3.05*	–2.81*	–1.37	–1.30					

Table IV. White bootstrap reality check. The statistics reported in this table are computed across eleven thick modelling-based forecasts and five estimation strategics (recursive, rolling and rolling with optimal chosen window estimation with the constant as the only focal variable; recursive and rolling estimation with optimal chosen window and four focal variables). The table reports *p*-values for the null that thin modelling-based forecasts outperform the available thick modelling-based forecasts

	Min	10%	25%	Median	75%	90%	Max
RC *p*-value	0.000	0.000	0.000	0.004	0.038	0.156	0.429

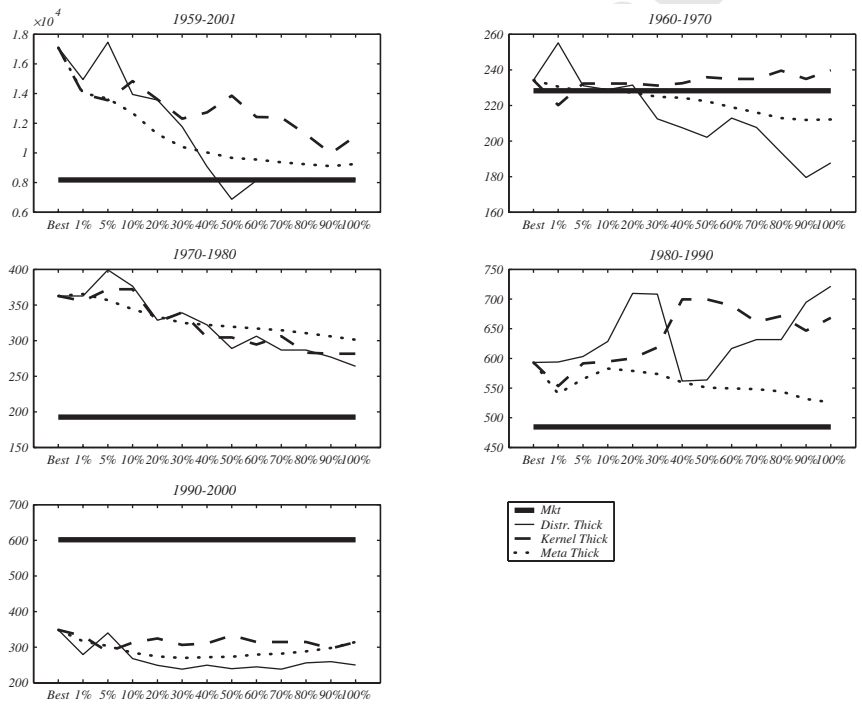

Figure 4. End-of-period wealth generated by asset allocation based on thin and thick modelling. Forecasts for excess returns are based on recursive estimation with one focal variable. On the horizontal axis we indicate the thickness of our approach in terms of the percentage of models (ranked by their within-sample performance) used in the construction of the different trading rules. Each panel reports the performance of a buy-and-hold strategy on S&P500 (Mkt), distributional thick modelling, meta thick modelling and kernel thick modelling strategies

models in terms of their adjusted R^2. Model-based portfolio allocations dominate the buy-and-hold strategy over the whole sample and in the decades 1970–1980 and 1980–1990. More complicated specification procedures tend to give a weaker outperformance relative to the buy-and-hold than the simple recursive specification. The evidence for the decade 1960–1970 is mixed in the sense that not all econometric-based strategies dominate on buy-and-hold strategy. In the last decade the buy-and-hold strategy is never outperformed, however the dominance of thick modelling over thin modelling becomes stronger.

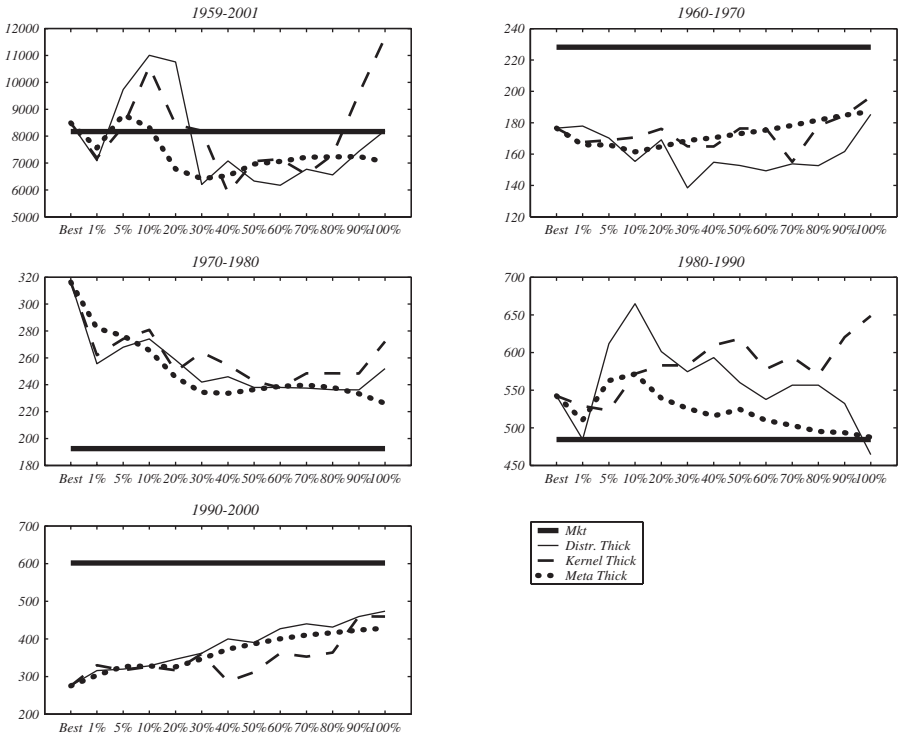

Figure 5. End-of-period wealth generated by asset allocation based on thin and thick modelling. Forecasts for excess returns are based on rolling estimation (optimal chosen window) and four focal variables. On the horizontal axis we indicate the thickness of our approach in terms of the percentage of models (ranked by their within-sample performance) used in the construction of the different trading rules. Each panel reports the performance of a buy-and-hold strategy on S&P500 (Mkt), distributional thick modelling, meta thick modelling and kernel thick modelling strategies

CONCLUSIONS

In this paper, we have reassessed the results on the statistical and economic significance of the predictability of stock returns provided by Pesaran and Timmermann (1995) for US data to propose a novel approach for portfolio allocation based on econometric modelling. We find that the results based on the thin modelling approach originally obtained for the sample 1960–1992 are considerably weakened in the decade 1990–2000.

We then show that the incorporation of model uncertainty substantially improves the performance of econometric-based portfolio allocation.

The portfolio allocation based on a strategy giving weights to a number of models rather than to just one model leads to systematic overperformance of portfolio allocations among two assets based on a single model. However, even thick modelling does not guarantee a constant overperformance with respect to a typical market benchmark for our asset allocation problem. To this end we have observed that combining thick modelling with a model specification strategy that imposes balanced regressions and chooses optimally the estimation window reduces the volatility of the asset alloca-

tion performance and delivers a more consistent outperformance with respect to the simple buy-and-hold strategy.

APPENDIX A: DATA

In the Pesaran–Timmermann (1995) data set (PT95) the data sources were as follows: stock prices were measured by the Standard & Poor's 500 index at close on the last trading day of each month. These stock indices, as well as a monthly average of annualized dividends and earnings, were taken from Standard & Poor's Statistical Service. The 1-month T-bill rate was measured on the last trading day of the month and computed as the average of the bid and ask yields. The source was the Fama–Bliss risk-free rates file on the CRSP tapes. The same for the 12-month discount bond rate. The inflation rate was computed using the producer price index for finished goods from Citibase, and the rate of change in industrial production was based on a seasonally adjusted index for industrial production (Citibase). The monetary series were based on the narrow monetary aggrgates published by the Fed of St. Louis and provided by Citibase.

The extended data set has been obtained merging the P&T original data set (1954.1–1992.12) with the new series retrieved from Datastream and FRED for the sample 1993.1–2001.9. All the financial variables are measured on the last trading day of each month.

	Code	Description
$P_t^{stock,US}$	TOTMKUS(RI)	US–DS MARKET—TOT RETURN IND
dy_t^{US}	TOTMKUS(DY)	US–DS market—Dividend yield
pe_t^{US}	TOTMKUS(PE)	US-DS MARKET—PER
$r1_t^{US}$	ECUSD1M	US EURO–$1 MONTH (LDN:FT)—MIDDLE RATE
ppi_t^{US}	USOCPRODF	US PPI—MANUFACTURED GOODS NADJ
$r12_t^{US}$	ECUSD1Y	US EURO–$1 YEAR (LDN:FT)—MIDDLE RATE
ip_t^{US}	USINPRODG	US INDUSTRIAL PRODUCTION
$M0_t^{US}$	USM0. . . . B	US MONETARY BASE CURA
$R10Y_t^{US}$	BMUS10Y(RY)	US YIELD-TO-MATURITY ON 10_YEAR GOV.BONDS

The data are available in Excel format from the following website: http://www.igier.unibocconi.it/favero (section working papers).

APPENDIX B: TESTING PERFORMANCE OF VARIOUS FORECASTING MODELS

In this paper we focus on out-of-sample tests of stock predictability. Out-of-sample tests are more stringent than in-sample tests and have important advantages over in-sample tests in assessing the predictability of stock returns. We analyse out-of-sample predictive ability using three recently developed statistics.

The first one is the market-timing test proposed by Pesaran and Timmermann (1992). The sign test is based on the proportion of times that the sign of a given variable y_t is correctly predicted

in the sample by the sign of the predictor x_t. Under the null hypothesis that x_t has no power in predicting y_t, the proportion of times that the sign is correctly predicted has a binomial distribution with known parameters, therefore a test of the null of predictive failure is constructed by comparing the observed proportion of signs correctly predicted with the proportion of signs correctly predicted under the null. The test statistic is computed as

$$\mathrm{Sn} = \frac{P - P^*}{\{V(P) - V(P^*)\}^{1/2}} \sim N(0,1)$$

where

$$P = \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

$$P^* = P_y P_x + (1 - P_y)(1 - P_x)$$

$$V(P^*) = \frac{1}{n} P^*(1 - P^*)$$

$$V(P) = n\left((2P_y - 1)^2 P_x(1 - P_x) + (2P_x - 1)^2 P_y(1 - P_y) + \frac{4}{n} P_y P_x (1 - P_y)(1 - P_x) \right)$$

Z_i is an indicator variable which takes a value of one when the sign of y_t is correctly predicted by x_t and zero otherwise, P_y is the proportion of times y_t takes a positive value, P_x is the proportion of times x_t takes a positive value.

The second one is the popular Diebold and Mariano (1995) statistic for equal predictive accuracy where we are testing the null hypothesis of a zero population mean loss differential between two forecasts. This test has a standard limiting distribution when comparing forecasts from non-nested models. However, we are comparing forecasts from nested models, so we follow the reccomendation of Clark and McCracken (2001b) and base our inference on a boostrap procedure similar to the one used in Kilian (1999). In order to derive the correct distribution for the statistic we apply the bootstrap in the following way. Let $d_{k,t}$, $t = 1, \ldots, n$ be the sequence of the realized difference in loss between model k and a benchmark model:

1. run the regression $E(d_t) = c + e_t$;
2. compute $\hat{e}_t$ and generate B bootstrap samples;[15]
3. generate B bootstrap responses $E(d_t)^{*1}, \ldots, E(d_t)^{*B}$ according to $E(d_t)^{*b} = \hat{c} + \hat{e}_t^{*b}$;
4. the new bootstrap data set is given by $(E(d_t)^{*b}, c)$;
5. compute the t-value of the constant and denote it by t^{*b};
6. derive the distribution of t^{*b};
7. compute the p-value as $\#(t^{actual} > t^{*b})/B$.

The third procedure we implement is the bootstrap reality check by White (2000), with consistent values given by Hansen (2001). In this case we are testing the null that a model (benchmark)

[15] There are different ways to generate the resamples: one approach is the stationary bootstrap by Politis and Romano (1994), another is the block bootstrap of Kunsch (1989).

performs better than other avaiable forecasting models in a given sample, taking care of data-snooping. The need to test for superior predictive ability arises from a situation in which, like our case, a family of forecasting models are compared in terms of their predictive ability defined in the form of a loss function. The question of interest is whether any alternative model is a better forecasting model than a benchmark model. When a large number of models are investigated prior to the selection of a model, then the search over models must be taken into account when making inference. After a search over several models, the relevant question is whether the excess performance of an alternative model is significant or not.

Let $X_k(t)$, $t = 1, \ldots, n$ be the sequence of realized performance of model k relative to a benchmark, $k = 0, \ldots, M$.

Let $b = 1, \ldots, B$ index the resamples of $\{1, \ldots, n\}$, given by $\theta_b(t)$, $t = 1, \ldots, n$ where B denotes the number of bootstrap resamples generated by the stationary boostrap of Politis and Romano (1994). The bth bootstrap resample is defined as: $X^*_{k,b}(t) = X_k(\theta_b(t)) - g(\overline{X}_{n,k})$, $b = 1, \ldots, B$, $t = 1, \ldots, n$ where $g(x) = \begin{cases} 0 & \text{if} \quad x \leq -A_{n,k} \\ x & \text{otherwise} \end{cases}$ where $A_{n,k}$ is a correction factor depending on an estimate of var $(n^{1/2}\overline{X}_{n,k})$. For $b = 1, \ldots, B$, we calculate $\overline{X}^*_{n,\max,b} = \max_k \overline{X}^*_{n,k,b}$ and the estimated p-value is given by

$$\hat{p} = \sum_{b=1}^{B} \frac{1\left(\overline{X}^*_{n,\max,b} > \overline{X}_{n,\max}\right)}{B}$$

In both cases it is very important to specify the loss function we have in mind. Evaluation of forecasting skills of a forecast producer may be best carried out using one of the purely statistical measures, while for a user forecast evaluation requires a decision-based approach.[16] From a user's perspective forecast accuracy is best judged by its expected economic value, the characterization of which requires a full specification of the user's decision environment. In our case, where the objective of forecasting is relatively uncontroversial, the importance of economic measures of forecast accuracy has been widely acknowledged and is straightforward. However, since we report economic measures of forecast accuracy in the next section, where we discuss the asset allocation performance, we decide to use the standard MSE loss function to test the different forecasts.

ACKNOWLEDGEMENTS

We are indebted to the editor, Allan Timmermann, two anonymous referees, Francesco Corielli, Clive Granger,Wessel Marquering, Alessandro Penati, Franco Peracchi, Hashem Pesaran, Eduardo Rossi, Guido Tabellini, as well as seminar participants at 'Ente Einaudi for Monetary and Financial Studies' in Rome, and University of California, San Diego for comments and suggestions.

[16] Whittle notes 'Prediction is not an end in itself, but only a means of optimizing current actions against the prospect of an uncertain future'. To evaluate forecasts we need to know how and by whom forecasts are used. See Pesaran and Skouras (2002) for further details.

REFERENCES

Aiolfi M, Favero CA, Primiceri G. 2001. Recursive 'thick' modelling of excess return and dynamic portfolio allocation. IGIER Working Paper No. 197.

Ait-Sahalia Y, Brandt M. 2001. Variable selection for portfolio choice. *Journal of Finance* **56**: 1297–1351.

Ang A, Bekaert G. 2002. Stock return predictability: is it there? Working Paper, Columbia Business School.

Avramov D. 2002. Stock return predictability and model uncertainty. *Journal of Financial Economics*, **64**: 423–458.

Barberis N. 2000. Investing for the long run when returns are predictable. *Journal of Finance* **55**(1): 225–264.

Bossaerts P, Hillion P. 1999. Implementing statistical criteria to select return forecasting models: what do we learn? *The Review of Financial Studies* **12**: 405–428.

Brandt M. 1999. Estimating portfolio and consumption choice: a conditional Euler equations approach. *Journal of Finance* **54**: 1609–1646.

Campbell JY, Shiller R. 1987. Cointegration and tests of present value models. *Journal of Political Economy* **95**: 1062–1088.

Campbell JY, Shiller R. 1988a. The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* **1**: 195–227.

Campbell JY, Shiller R. 1988b. Stock prices, earnings, and expected dividends. *Journal of Finance* **43**: 661–676.

Campbell JY, Viceira L. 1999. Consumption and portfolio decisions when expected returns are time-varying. *Quarterly Journal of Economics* **114**: 433–495.

Campbell JY, Lo A, McKinlay. 1997. *The Econometrics of Financial Markets.* Princeton University Press: Princeton, NJ.

Campos J, Hendry DF, Krolzig H-M. 2003. Consistent model selection by an automatic gets approach. *Oxford Bulletin of Economics & Statistics*. **65**: 803–820.

Chan, Stock J, Watson M. 1999. A dynamic factor model framework for forecast combination. *Spanish Economic Review* **1**: 91–121.

Chipman H, George EI, McCulloch RE. 2001. The practical implementation of Bayesian model selection. *IMS Lecture Notes, Monograph Series* Vol. 38.

Clemen RT. 1989. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* **5**: 559–581.

Clements MP, Hendry DF. 2001. *Forecasting Economic Time-Series*. Cambridge University Press: Cambridge.

Cochrane J. 1999. Portfolio advice for a multifactor world. NBER Working Paper No. 7170.

Diebold F, Mariano R. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–265.

Diebold F, Pauly P. 1987. Structural change and the combination of forecasts. *Journal of Forecasting* **6**: 21–40.

Doppelhofer G, Miller RI, Sala-i-Martin. 2000. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. NBER Working Paper No. 7750.

Elliott G, Timmermann A. 2002. Optimal forecast combinations under general loss functions and forecast error distributions. UC San Diego Department of Economics Discussion Paper 03–09.

Fama E, French KR. 1988. Dividend yields and expected stock returns. *Journal of Financial Economics* **22**: 3–25.

Giacomini R, White H. 2003. Test of conditional predictive ability. UC San Diego Department of Economics Discussion Paper 03–09.

Granger CWJ. 2003. Modeling conditional distribution. Mimeo, UC San Diego Department of Economics.

Granger CWJ, Jeon Y. 2003. Thick modeling. Mimeo, UC San Diego Department of Economics.

Granger CWJ, Newbold P. 1986. *Forecasting Economic Time Series*, 2nd edn. Academic Press: San Diego.

Granger CWJ, Timmermann A. 1999. Data mining with local model specification uncertainty: a discussion of Hoover and Perez. *The Econometrics Journal* **2**: 220–226.

Hansen PR. 2001. An unbiased and powerful test for superior predictive ability. Brown University Department of Economics Discussion Paper 01–06.

Hoeting J, Madigan D, Raftery A, Volinsky C. 1999. Bayesian model averaging: a tutorial. Technical Report 9814, Department of Statistics, Colorado State University.

Johansen S. 1995. *Likelihood-Based Inference Cointegrated Vector Autoregressive Models*. Oxford University Press: Oxford.

Kandel S, Stambaugh RS. 1996. On the predictability of stock returns: an asset-allocation perspective. *Journal of Finance* **51**(2): 385–424.
Keim D, Stambaugh RS. 1986. Predicting returns in the stock and bond markets. *Journal of Financial Economics* **17**: 357–390.
Lamont O. 1998. Earnings and expected returns. *Journal of Finance* **53**(5): 1563–1587.
Lander J, Orphanides A, Douvogiannis M. 1997. Earning forecasts and the predictability of stock returns: evidence from trading the S&P. Board of Governors of the Federal Reserve System, http:\\www.bog.frb.fed.org.
15 Lettau, Ludvigson S. 2001. Consumption, aggregate wealth and expected stock returns. *Journal of Finance* **56**(3): 815–849.
Marcellino M. 2002. Forecast pooling for short time series of macroeconomic variables. IGIER Working Paper No. 212.
Paye B, Timmermann A. 2002. How stable are financial prediction models? Evidence from US and international stock markets. UC San Diego Department of Economics Discussion Paper 02–13.
Pesaran MH, Timmermann A. 1992. A simple non-parametric test of predictive performance. *Journal of Business and Economics Statistics* **10**: 461– 465.
Pesaran MH, Timmermann A. 1995. Predictability of stock returns: robustness and economic significance. *Journal of Finance* **50**(4): 1201–1228.
Pesaran MH, Timmermann A. 2000. A recursive modelling approach to predicting UK stock returns. *The Economic Journal* **110**: 159–191.
Pesaran MH, Timmermann A. 2002. Market timing and return prediction under model instability. *Journal of Empirical Finance* **9**: 495–510.
16 Raftery A, Madigan D, Hoeting J. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**(437): 179–191.
16 Rapach DE, Wohar ME. 2002. Financial variables and the predictability of stock and bond returns: an out-of-sample analysis. Mimeo.
17 Samuelson PA. 1969. Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics* **51**: 239–246.
17 Siegel J. 1994. *Stocks for the Long Run*. Richard D. Irwin: Burr Ridge, I.
Sims C, Stock J, Watson M. 1990. Inference in linear time-series models with some unit roots. *Econometrica* **58**: 113–144.
Stock J, Watson M. 1999. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In *Cointegration, Causality, and Forecasting: a Festschrift in Honour of Clive W.J. Granger*, Engle RF, White H (eds), Chapter 1. Oxford University Press: Oxford.
17 Sullivan R, Timmermann A, White H. 1999. Data-snooping, technical trading rules performance and the bootstrap. *Journal of Finance* **54**: 1647–1692.
Zhou C. 1996. Stock market fluctuations and the term structure. Board of Governors of the Federal Reserve System, http://www.bog.frb.fed.org.
White H. 2000. A reality check for data snooping. *Econometrica* **68**: 1097–1126.

Authors' biographies:

Carlo A. Favero is Professor of Economics at Bocconi University, Director of IGIER and a Fellow of the International Macroeconomics programme at CEPR. He holds an MSc in economics from the LSE and a DPhil from Oxford University, where he was a member of the Oxford Econometrics Research Centre. He was Associate Professor of Econometrics at Bocconi University from 1994 to 2001 and Professor of Economics since 2002.

Marco Aiolfi is a third-year PhD student at Bocconi University. He graduated in economics at Bocconi and during his PhD he has been a visiting graduate student at the Department of Economics of UCSD.

Authors' address:

Carlo A. Favero and **Marco Aiolfi**, IGIER-Università Bocconi, Via Salasco 5, 20124 Milan, Italy.

AUTHOR QUERY FORM

Dear Author,
During the preparation of your manuscript for publication, the questions listed below have arisen.
Please attend to these matters and return this form with your proof.
Many thanks for your assistance.

Query References	Query	Remarks
1	AQ: Please supply keywords	
2	AQ: Granger (2002) not listed in Refs?	
3	AQ: Stock and Watson (2001, 2003) not listed in Refs?	
4	AQ: Stock and Watson (1996) not listed in Refs.	
5	AQ: Brown et al. (1975) not listed in Refs?	
6	AQ: Clark and McCracken (2001b) and Kilian (1999) not listed in Refs?	
7	AQ: Politis and Romano (1994) and Kunsch (1989) not listed in Refs?	
8	AQ: Pesaran and Skouras (2002) not listed in Refs?	
9	AQ: Not cited?	
10	AQ: Not cited?	
11	AQ: Not cited?	
12	AQ: McKinlay initials?	
13	AQ: Chan initials?	
14	AQ: Not cited?	
15	AQ: Lettau initials?	
16	AQ: Not cited?	
17	AQ: Not cited?	