Forecasting Methods in Finance

Allan Timmermann
UC San Diego, Rady School of Management

March 2, 2018

Abstract

Our review highlights some of the key challenges in financial forecasting problems along with opportunities arising from the unique features of financial data. We analyze the difficulty of establishing predictability in an environment with a low signal-to-noise ratio, persistent predictors, and instability in predictive relations arising from competitive pressures and investors' learning. We discuss approaches for forecasting the mean, variance, and probability distribution of asset returns. Finally, we cover how to evaluate financial forecasts while accounting for the possibility that numerous forecasting models may have been considered, leading to concerns of data mining.

1 Introduction

Finance is focused on intertemporal decision making under uncertainty and so forecasts of unknown future outcomes is integral to several areas of finance. Asset pricing requires forecasts of future cash flows, payoffs and discount rates. Risk management relies on forecasts of variances and covariances of returns on portfolios that frequently comprise large numbers of assets. Countless studies in corporate finance analyze firms' capital budgeting decisions which in turn depend on projected cash flows and firms' forecasts of the costs and benefits of issuing debt and equity. A large literature in banking analyzes the possibility of "runs" which reflects investors' forecasts of both a bank's solvency and liquidity as well as their expectation of other agents' (depositors') decisions on whether to run or stay put.

While economic and financial forecasting share many methods and perspectives, some important features help differentiate the two areas. First, competitive pressures and market efficiency mean that the "signal-to-noise" ratio in many financial forecasting problems–particularly predictability of asset returns–is very low compared to standard forecasting problems in macroeconomics in which the presence of a sizeable persistent component makes forecasting easier. The presence of weak predictors with low predictive power and the resulting importance of parameter estimation error is, therefore, the norm rather than the exception in financial forecasting.

Second, and related to the first point, fierce competition among asset managers in the financial markets means that predictable patterns in asset returns can be expected to self destruct as a result of investors' attempts to exploit predictability and the resulting adjustment in prices. The possibility of readily trading on price forecasts makes the scope for feedback effects from forecasts to actual outcomes stronger in finance than in other areas of economics. Model instability is therefore particularly important to financial forecasting.

Third, overfitting and issues related to data mining have increasingly become a concern in financial forecasting due to the ease with which numerous forecasting models can be fitted to a given data set and the difficulty of generating new and genuinely independent data sets on which to test the forecasting performance. In particular, how should the performance of a forecasting model be evaluated when this model is selected as the best performer among a larger set of competing specifications? This situation generates a multiple hypothesis testing problem that, if not accounted for, can lead to findings of spurious predictability patterns and serious distortions in inference.

Fourth, while volatility forecasting also features prominently in forecasting of macroeconomic variables–indeed the original application of ARCH models was to UK inflation (Engle, 1982)–it is more central to finance. This is particularly true in the area of risk management which can entail forecasting the correlations between very large sets of variables and so gives rise to high-dimensional forecasting problems. Moreover, access to high-frequency data, sampled every few seconds during trading sessions for the most liquid assets, means that measures of "realized" variances can be constructed and used to forecast future risks. This type of data does not, as yet, have obvious counterparts in economics where measurements tend to be conducted at a lower frequency.

Fifth, the presence of derivatives markets such as options or credit default swaps means that risk-neutral densities can be constructed under no-arbitrage conditions and used to forecast the probability distribution of asset prices. Once converted into physical probability distributions, such density estimates can be combined with forecasts obtained from other sources. Using options data in this manner introduces a host of complexities, however, related to having limited cross-sectional data on liquid traded options.

Sixth, financial forecasting problems often involve well-defined loss functions leading to optimization problems such as maximizing the expected utility from trading for an investor with mean-variance or power utility. In turn, this involves forecasting the probability distribution of portfolio payoffs or particular moments of this distribution. Given such utility functions, it is now routine to evaluate forecasting performance using economic measures such as certainty equivalent returns or average realized utilities from investments strategies based on a sequence of forecasts.

A variety of methods have been–or have the potential for being–used to deal with these challenges in financial forecasting. For example, methods for dealing with weak predictors and parameter estimation error such as forecast combination and, more broadly, ensemble forecasting methods developed in machine learning are beginning to find more widespread use. Forecasting methods

that take advantage of constraints from economic theory, e.g., by using filtering methods to back out persistent components in expected returns and expected dividend growth or by imposing bounds on the conditional Sharpe ratio, have also shown promise. Our review discusses these and other strategies for improving financial forecasting performance.

Our review proceeds as follows. Section 2 introduces the basic return predictability problem. Section 3 discusses challenges encountered in financial forecasting problems, including weak predictors (low signal-to-noise ratios), persistent predictors, model instability, and data mining. Section 4 discusses strategies for dealing with these challenges. Section 5 covers volatility and density forecasting methods, while Section 6 discusses methods for evaluating financial forecasts, emphasizing the use of economic performance measures, and Section 7 concludes.

2 Basics of return predictability

Let r_{t+1} denote the excess return on a risky asset held from period t to period $t+1$, net of a risk-free rate. Ignoring frictions due to transaction costs and restrictions on trading, under conditions of no arbitrage the following moment condition holds:

$$E_t[m_{t+1}r_{t+1}] = 0, \tag{1}$$

where m_{t+1} is the positively-valued stochastic discount factor (pricing kernel), see, e.g., Cochrane (2009) and $E_t[.] = E[.|\Omega_t]$ denotes conditional expectations given information at time t, Ω_t.

Equation (1) shows that the product of the pricing kernel and excess returns is a martingale difference sequence and so has mean zero conditional on the filtration generated by Ω_t. Solving for expected excess returns, we have

$$E_t[r_{t+1}] = \frac{-cov_t(r_{t+1}, m_{t+1})}{E_t[m_{t+1}]}, \tag{2}$$

where $cov_t(r_{t+1}, m_{t+1}) = E_t[(r_{t+1} - E_t[r_{t+1}])(m_{t+1} - E_t[m_{t+1}])]$ is the conditional covariance between r_{t+1} and m_{t+1}. This equation shows that predictability of excess returns is not ruled out by the absence of arbitrage. However, to be consistent with no-arbitrage conditions, any return predictability should reflect time variation either in the conditional covariance between excess returns and the stochastic discount factor, $cov_t(r_{t+1}, m_{t+1})$ or variation in the conditional expectation of the pricing kernel, $E_t[m_{t+1}]$.

A key challenge to interpretation of empirical evidence on return predictability is that the object which theory stipulates should be a martingale difference sequence, $m_{t+1}r_{t+1}$, is itself unobserved and model dependent.[1] Hence, interpretations of return predictability should always bear in mind the joint hypothesis

[1] For example, in a consumption based asset pricing model, the pricing kernel will reflect investors' intertemporal marginal rate of substitution between current and future consumption and, thus, depends on the assumed utility specification.

problem well-known from studies of market efficiency: predictability tests are really joint tests of market efficiency and a correct specification of investor preferences. For example, stock returns may be predictably higher during recessions than in expansions simply because investors' marginal utility of consumption (and, hence, risk premia) are higher during states with low growth.

By far the most commonly used prediction model in empirical studies is a simple linear specification for the equity premium:

$$r_{t+1} = \mu + \beta x_t + u_{t+1}, \tag{3}$$

where $x_t \in \Omega_t$ is a set of predictor variables known at time t. While the linear forecasting model in (3) may appear to be at odds with the more general first-order equation in (1), in fact it can be derived under quite general conditions.[2]

Further insights into the importance of forecasting for asset pricing can be gleaned from the log-linearized present value model of Campbell and Shiller (1988) which gives rise to the following approximate relation between the current log-price, p_t, and forecasts of future log-dividends, d_{t+1+j}, and continuously compounded returns, r_{t+1+j} :

$$p_t = \frac{k}{1-\rho} + E_t \left[\sum_{j=0}^{\infty} \rho^j [(1-\rho) d_{t+1+j} - r_{t+1+j}] \right], \tag{4}$$

where k and ρ are constants arising from the log-linearization.

Computing the price of a perpetual asset such as a stock therefore requires forecasting an infinite stream of cash flows (log-dividends, d_{t+1+j}) and discount rates (r_{t+1+j}). This complex task requires not only forecasting all future values of these variables themselves, but also forecasting the future values of any other variables used to predict cash flows and discount rates.[3]

Letting Δd_{t+i} denote the log-dividend growth rate, it follows that surprises to returns are driven either by changes in expected future dividends or changes in expected future returns:

$$\begin{aligned} r_{t+1} - E_t[r_{t+1}] &= E_{t+1} \sum_{j=0}^{\infty} \rho^j \Delta d_{t+1+j} - E_t \sum_{j=0}^{\infty} \rho^j \Delta d_{t+1+j} \\ &\quad - \left(E_{t+1} \sum_{j=1}^{\infty} \rho^j r_{t+1+j} - E_t \sum_{j=1}^{\infty} \rho^j r_{t+1+j} \right). \end{aligned} \tag{5}$$

Noting that $E_{t+1}[\bullet]$ and $E_t[\bullet]$ represent forecasts computed conditional on information at time $t+1$ and time t, respectively, deviations in realized returns from their previously expected values must be driven by changes in dividend or

[2] Assuming an affine pricing kernel and cash flows that are formed as a linear combination of a finite-dimensional, stationary vector autoregression, Farmer, Scmidt, and Timmermann (2017) show that (3) can be derived from a log-linearized asset pricing model.

[3] This task is typically accomplished using vector autoregressions (VARs).

return expectations. Importantly, all future values of these variables matter so changes in forecasts of payoffs or discount rates at *any* future horizon should lead to corresponding changes in returns.

3 Challenges to financial forecasting models

Researchers attempting to identify predictability in asset returns face several challenges which we describe in this section before discussing strategies for addressing such challenges in the next section.

3.1 Challenge I: Weak predictors

A very low signal-to-noise ratio in predictive return regressions is to be expected from market competition among asset managers and other investors who use vast resources to vie for higher returns.

One way to model the "weak predictor" feature of financial return regressions is to assume that the coefficient of the time-varying predictor in equation (3) is local-to-zero, i.e.,

$$\beta \propto \frac{b}{\sqrt{T}}, \tag{6}$$

for some constant, b, and a sample size, T. This approximation suggests that prediction models have limited power even in cases with a large sample size, T. Situations with weak predictors imply that parameter estimation error is roughly of the same order of magnitude as the signal embedded in the predictor, implying that conventional tests of predictive performance such as that proposed by Diebold and Mariano (1995) will have little power to detect return predictability.

Bayesian methods have been used to counter the important effect of parameter estimation error on return forecasts. By shrinking the coefficient estimates towards a prior (usually centered on zero for the slope coefficients on time-varying predictor variables), these methods dampen the effect of estimation error on the forecasts. Even though such forecasts may result in biased forecasts, they can exploit the bias-variance trade-off in such a way as to improve on the forecasting models' mean squared error performance.

From a variable selection perspective, weak predictors create a grey area where inclusion of individual predictors in return regressions is surrounded by considerable uncertainty that is unlikely to be conclusively resolved by conventional model selection methods.

The difficulty associated with establishing return predictability is countered by the fact that even small amounts of return predictability has the potential of translating into significant economic gains. Back of the envelope calculations by Campbell and Thomson (2008) suggest that an (out-of-sample) R^2 value as low as 0.005 (one-half of one percent) in a monthly return regression could generate a 40% increase in the average portfolio excess return of a mean-variance investor

with a modest degree of risk aversion.[4] See also Zhou (2010) for a discussion of this point.

3.2 Challenge II: Persistent predictors

Many of the predictors used to forecast stock returns–notably valuation ratios such as the dividend yield but also short term interest rates or interest rate spreads–are highly persistent. As pointed out by Stambaugh (1999), this can lead to biases in inference on the slope coefficient β in (3) provided that the innovation in the predictor is strongly correlated with unexpected shocks to returns.

Suppose the persistence in the predictor is captured by modeling this as a first-order autoregression

$$x_t = \mu_x + \rho x_{t-1} + v_t, \quad |\rho| < 1, \quad v_t \sim (0, \sigma_v^2). \tag{7}$$

Assuming Gaussian innovations, Marriott and Pope (1954) and Kendall (1954) show that there is a finite-sample bias in the estimated coefficient $\hat{\rho}$:

$$E[\hat{\rho} - \rho] \approx \frac{-(1+3\rho)}{T}.$$

Finally, suppose that $E(u_t|x_s, x_w) \neq 0, \quad s < t \leq w$ so that the u_t residuals from the linear return regression (3) are correlated with past or future values of the predictor, x, i.e., $\sigma_{uv} = E[u_t v_t] \neq 0$. This condition obviously holds for predictors such as the dividend-price ratio that have prices in the denominator.

Under these conditions, Stambaugh (1999, Proposition 4 and Corollary) shows that finite-sample biases in $\hat{\rho}$ translate into a finite-sample bias in $\hat{\beta}$:

$$\begin{aligned} E[\hat{\beta} - \beta] &= \frac{\sigma_{uv}}{\sigma_v^2} E[\hat{\rho} - \rho] \\ &= \frac{-\sigma_{uv}}{\sigma_v^2} \frac{(1+3\rho)}{T} + O(T^{-2}). \end{aligned} \tag{8}$$

Hence, if u_t and v_t are uncorrelated, there will not be a problem with a finite-sample bias in $\hat{\beta}$. Conversely, if $\sigma_{uv} \neq 0$, the bias can be substantial. For example, Stambaugh (1999) estimates a bias around 0.40 in a regression of returns on the dividend yield from 1977-1996 ($T = 240$).[5]

3.3 Challenge III: Model instability

The same set of return predictor variables typically does not work for extended periods of time; which variables get selected changes over time. No single prediction model is therefore clearly superior to other models. This complicates

[4] This calculation ignores trading costs and slippage in prices as a result of market impact from trading.

[5] Stambaugh demonstrates that Bayesian inference can yield sharper results, although such inference will depend on the priors, assumptions about the initial observation (fixed or stochastic), and stationarity assumptions for the predictor.

the selection of a forecasting model and produces an expected loss function that is a probability weighted average across all models.

There is a particular reason why we would expect instability to affect financial forecasts. Asset returns depend on asset prices which themselves reflect investors' expectations of future payoffs. A release of new public information that leads to a change in price forecasts can be expected to lead to simultaneous shifts in prices and, if shared by the broad market, will be incorporated into the market price with little delay.[6] If the information does not lead to any shift in risk premia, it cannot be expected to lead to improved forecasts of changes in prices or, more precisely, excess returns. Contrast this with new information about the state of the economy, e.g. the release of a jobs report which can be expected to lead to more accurate forecasts of future economic growth as it reveals more accurate information about the underlying state of the economy.

Consistent with the tendency for financial return models to "self-destruct", empirically the linear return prediction model in (3) has been found to be unstable in many empirical studies. Paye and Timmermann (2006) and Rapach and Wohar (2006) test for parameter stability and find that the null of stability is rejected for the most commonly used predictor variables using returns data from the US and a range of international stock markets.[7] Pastor and Stambaugh (2001) use a Bayesian approach to find breaks in the relation between expected returns and return volatility.[8]

Farmer, Schmidt, and Timmermann (2017) present evidence from nonparametric regressions which suggests that return predictability is concentrated in a small number of "pockets" with no significant predictability in the majority of the sample. As some of these pockets are short-lived, lasting a few weeks, this poses a particular challenge for real-time detection of predictability and attempts at exploiting such "pockets" to generate improved statistical and financial performance.

Conventional time-series regressions generally have a limited ability (weak power) to detect breaks in the regression parameters in "real time" without lengthy delays as only few post-break observations are available. This problem is exacerbated in cases with weak predictors whose effect gets veiled by parameter estimation error: Detecting a shift in the parameter of a variable that only possesses weak predictive power–and whose inclusion in the forecasting model

[6] McClean and Pontiff (2016) find evidence that a range of stock market anomalies, i.e., signs of mispricing, tend to vanish after they are published in academic journals. They document a 32% average reduction in anomalies due to publication-informed trading. Moreover, the post-publication reduction in the magnitude of such anomalies appears to be greater in cases with larger in-sample returns and for stocks that have less idiosyncratic risk and are more liquid. These anomalies are easier to detect and easier to implement trading strategies to exploit, and so we would expect them to disappear more rapidly once their existence becomes more broadly known.

[7] Some of this instability may be related to state-dependent forecasting performance' Rapach, Strauss, and Zhou (2010) and Dangl and Halling (2012) find evidence that return predictability is strong around economic recessions but much weaker during expansions.

[8] See Rossi (2013a) for an extensive review of evidence on model instability. Rossi (2013b) reviews the track record of exchange rate forecasting models and discusses how these are affected by model instability.

is questionable–is more difficult than if the variable had a large and highly significant effect. To put it bluntly, accurate detection of shifts to the parameter of a variable that generates an R^2 of less than one percent is always going to be difficult.

While it is a difficult task to detect breaks to financial forecasting models, converting such evidence into more accurate forecasts is even more challenging. This holds regardless of whether a break is detected through some econometric test or is based on subjective beliefs. For example, suppose it is believed that the election of Donald Trump as US president lead to a change in the forecasting models, e.g., for corporate cash flows as a result of changes in expectations of economic growth or prospects for tax reductions. In the immediate aftermath of Trump's election in November 2016, investors would have had very few data points from the new "regime" and so would not have been able to accurately estimate the parameters of the model. If not properly managed, this could result in erratic forecasts dominated by estimation error in the early stages after a break.

3.4 Challenge IV: Data mining and overfitting

Data mining is a concern that affects many forecasting problems. In its simplest form it is closely linked to overfitting. To illustrate the problem, consider two linear forecasting models, the first of which (M_1) uses a set of predictors, x_{1t}, while the second model (M_2) uses x_{1t} in addition to a set of predictors, x_{2t},

$$\begin{aligned} M_1 &: y_{t+1} = \beta_1' x_{1t} + \varepsilon_{1t+1} \\ M_2 &: y_{t+1} = \beta_{21}' x_{1t} + \beta_{22}' x_{2t} + \varepsilon_{2t+1}. \end{aligned} \tag{9}$$

Suppose that the coefficient estimates for these models are obtained by OLS so that

$$\begin{aligned} \hat{\beta}_1 &= \arg\min_{\beta_1} T^{-1} \sum_{t=0}^{T-1} (y_{t+1} - \beta_1' x_{1t})^2, \\ \hat{\beta}_2 &= \arg\min_{\beta_2} T^{-1} \sum_{t=0}^{T-1} (y_{t+1} - \beta_2' x_t)^2, \end{aligned} \tag{10}$$

where $\beta_2 = (\beta_{21}' \ \beta_{22}')'$ and $x_t = (x_{1t}' \ x_{2t}')'$. This results in two forecasts $y_{1t+1|t} = \hat{\beta}_1' x_{1t}$ and $y_{2t+1|t} = \hat{\beta}_2' x_t$.

Suppose that forecasting performance is measured using the same sample, $\{y_{t+1}\}_{t=0}^{T-1}$ that is used to estimate the parameters in (10). Because M_2 nests M_1, M_2 will always provide a (weakly) better in-sample fit:

$$T^{-1} \sum_{t=0}^{T-1} (y_{t+1} - \hat{\beta}_{21}' x_{1t} - \hat{\beta}_{22}' x_{2t})^2 \leq T^{-1} \sum_{t=0}^{T-1} (y_{t+1} - \hat{\beta}_1' x_{1t})^2, \tag{11}$$

where equality in (11) only holds if the second set of predictors, x_{2t}, are orthogonal to the forecast errors from the first model, $(y_{t+1} - \hat{\beta}_1' x_{1t})$. This means

that the biggest model (M_2) (almost) always comes out on top with the lowest in-sample MSE. This conclusion holds even if the bigger model can be expected to perform worse in population than the small model (M_1), perhaps due to its inclusion of redundant predictors. In other words, (11) can hold simultaneously with

$$E\left[(y_{t+1} - \hat{\beta}_{21}x_{1t} - \hat{\beta}_{22}'x_{2t})^2\right] > E\left[(y_{t+1} - \hat{\beta}_1'x_{1t})^2\right], \tag{12}$$

where $E[.]$ denotes population expectation. This overfitting effect makes data mining tempting, but also means that great care must be exercised when evaluating any improvements in forecasting performance resulting from the inclusion of additional predictor variables.

Ferson, Sarkissian, and Simin (2003) find that the effect of data mining for predictor variables can be exacerbated in a setup with persistent predictors which introduces a spurious regression bias. Spurious regression biases can arise if the return generating process contains a highly persistent expected return component which is correlated in finite samples with similarly persistent predictor variables. Data mining gets reinforced by this effect as the predictors that appear to be the best ones, and thus produce the highest R^2 values, may simply be maximizing the spurious regression bias.

Data mining concerns do not only arise in the context of time-series prediction models. Harvey, Liu and Zhu (2016) undertake a throughout study of the academic literature on factors that have been proposed to explain cross-sectional variation in expected returns. They propose an approach for dealing with the multiple hypothesis testing issues that arise when a large number of factors need to be considered and suggest that new factors need to generate t-statistics above 3.0 in order to be statistically significant after accounting for the multiple hypothesis testing problem.

4 Strategies for addressing the challenges

This section describes a variety of strategies that have been used to address the challenges outlined in the previous section. We start with forecast combinations, turn to filtering (unobserved components) methods and approaches for capturing model change, before covering Bayesian methods, machine learning techniques and theory-induced constraints on the forecasting models.

4.1 Forecast combination

Forecast combination methods have been used extensively in economic forecasting, but are less widespread in financial forecasting. This is somewhat surprising given that a large literature has established substantial benefits from combining forecasts in a wide set of areas.[9]

Why combine financial forecasts? One answer is that forecasters often employ many models with similar predictive performance making it difficult to

[9] See, e.g., Clements (1989) and Timmermann (2006).

identify a single, superior model. Another reason lies in state-dependent forecasting performance: certain models may work well under some market conditions but not in other and it can be difficult to tell, ex ante, which conditions will prevail in the future. Alternatively, the forecasting environment may simply be unstable, rendering individual forecasting models' past track records unreliable for their future performance. A third reason is simply that of "diversification": all models are misspecified and combining forecasts, e.g., by using an equally-weighted average of forecasts, has the effect of diversifying across model uncertainty.[10]

To see how forecast combination works, consider a vector of n individual forecasts of some variable, $f_{t+1|t} = (f_{1t+1|t}, f_{2t+1|t}, ..., f_{nt+1|t})'$, where $f_{jt+1|t}$ is the one-step-ahead forecast of y_{t+1} given information at time t, Ω_t, generated by the jth model. The simple equal-weighted forecast combination takes the form

$$f^c_{t+1|t} = \frac{1}{n}\sum_{j=1}^{n} f_{jt+1|t}. \tag{13}$$

This "$1/n$" strategy has proven highly successful in empirical applications, including to form portfolios (DeMiguel et al. (2007)). There are no weights to estimate from the data in this combination scheme and the weight on each forecast does not depend on the individual forecasts' past performance.

A more general approach to forecast combination that accounts for the individual models' past forecasting performance estimates the combination weights from a linear regression of the outcome, y_{t+1}, on the predictors

$$y_{t+1} = \beta_0 + \sum_{j=1}^{n} \beta_j f_{jt+1|t} + \varepsilon_{t+1}. \tag{14}$$

An intercept term (β_0) is often included so as to ensure that the combined forecast is (unconditionally) unbiased. If each of the individual forecasts is believed to be unbiased, alternatively one can impose the constraints $\beta_0 = 0$ and $\sum_{j=1}^{n} \beta_j = 1$ in (14) so as to preserve unbiasedness of the combined forecast.

Rapach, Strauss, and Zhou (2010) is a notable exception to the relative shortage of papers in financial forecasting that use forecast combination methods. They fit univariate forecasting models to returns on the US stock market, using a set of predictors from Welch and Goyal (2008), and form an equal-weighted average of these forecasts. Suppose the univariate return prediction models take the form

$$r_{t+1} = \gamma_{0j} + \gamma_{1j} x_{jt} + u_{jt+1}, \quad j = 1, ..., n \tag{15}$$

[10] As a case in point, the accuracy of individual forecasts of asset returns is often reduced by the effects of estimation error. One reason forecast combination succeeds in producing better return forecasts is by diversifying the effect of estimation errors when these are not too strongly correlated across forecasting models.

Estimates of each of these regressions can be used to generate a return forecast $\hat{r}_{jt+1|t} = \hat{\gamma}_{0j} + \hat{\gamma}_{ji} x_{it}$, which in turn can be used to form an equal-weighted forecast

$$\bar{r}^{EW}_{t+1|t} = \frac{1}{n} \sum_{j=1}^{n} \hat{r}_{jt+1|t}. \tag{16}$$

Empirically, Rapach et al. (2010) find that an equal-weighted combination of quarterly forecasts from 15 univariate models of the form in (15) produces significantly more accurate forecasts than those from a model that assumes a constant equity risk premium and so imposes $\gamma_{1i} = 0$ in (15).[11]

Elliott, Gargano, and Timmermann (2013) generalize the approach in Rapach et al. (2010) to complete subset regressions that use equal-weighted averages of all forecasting models that include a fixed number (k) of predictor variables. The univariate case with a single predictor proposed by Rapach et al. (2010) is a special case ($k = 1$).[12] Empirically, using a similar data set as that in Rapach et al. (2010), Elliott et al. (2013) find that equal-weighted combinations of forecasts from return prediction models that include a small set of predictors–typically around two or three–perform notably better than forecasts from individual forecasting models and also are better than univariate forecasts like those considered by Rapach et al. (2010).

An alternative to averaging forecasts from a set of parsimonious models is to reduce the dimensionality of the set of predictors by assuming a latent factor structure which allows for the extraction of a small set of common factors (diffusion indexes) that drive variation in the predictor variables. Ludvigson and Ng (2007), Kelly and Pruitt (2013), and Neely, Rapach, Tu, and Zhou (2014) are examples of studies that use common components to predict returns.

4.2 Filtering and unobserved components models

Investors' conditional expectations of future cash flows and discount rates are unobserved so cannot directly be used to compute prices in (4) or track return movements in equation (5). Assuming that prices embody investors' forward-looking expectations, we can treat these expectations as latent variables, estimates of which can be computed using filtering methods under the assumption that the log-linearized present value model holds.

To see how this might work, let $\mu_t = E_t[r_{t+1}]$ be the expected return while the expected dividend growth is denoted $g_t = E_t[\Delta d_{t+1}]$. Following van Binsbergen and Koijen (2010), suppose that these follow AR(1) processes:

$$\mu_{t+1} = \delta_0 + \delta_1(\mu_t - \delta_0) + \varepsilon^{\mu}_{t+1}, \tag{17}$$

$$g_{t+1} = \gamma_0 + \gamma_1(g_t - \gamma_0) + \varepsilon^{g}_{t+1}\ . \tag{18}$$

[11] It appears to be less important whether the mean, median, or a trimmed mean of the forecasts is used to compute the forecast combination.

[12] From a theoretical perspective, Elliott et al. (2015) show that the complete subset regression combination achieves variance reduction relative to a "kitchen sink" approach that includes the full list of predictor variables.

In turn, the realized dividend growth is expressed as its expected value, g_t, plus an unexpected shock, ε^d_{t+1} :

$$\Delta d_{t+1} = g_t + \varepsilon^d_{t+1}. \tag{19}$$

Using the log-linearized present value model, van Binsbergen and Koijen (2010) show that the log price-dividend ratio takes the form

$$pd_t = A - B_1(\mu_t - \delta_0) + B_2(g_t - \gamma_0), \tag{20}$$

where A, B_1, and B_2 are functions of the underlying parameters. van Binsbergen and Koijen show that this setup yields a state space system with the two state equations, (17) and (18), and two measurement equations, (19) and (20), the latter without an error term.

Using this representation, estimates of expected returns and dividend growth rates can be computed using a Kalman filter. Empirically, van Binsbergen and Koijen (2010) find that such filtered estimates have significant predictive power over future returns and dividend growth. Interestingly, both dividend growth and expected returns contain persistent components with stronger persistence estimated for the latter.

Pastor and Stambaugh (2009) develop an approach that also treats expected returns as a latent process. Their predictive system approach is composed of the following state space system for the observable returns, r_{t+1}, the observable predictor, x_{t+1}, and the unobserved expected return, μ_{t+1} :[13]

$$\begin{aligned} r_{t+1} &= \mu_t + u_{t+1} \\ x_{t+1} &= (1-\alpha)E_x + \alpha x_t + v_{t+1} \\ \mu_{t+1} &= (1-\beta)E_r + \beta\mu_t + w_{t+1}, \end{aligned} \tag{21}$$

where $(u_{t+1}, v_{t+1}, w_{t+1}) \sim N(0, \Sigma)$. Assuming this joint dynamics, the linear regression in equation (3) is misspecified if the predictor variable is imperfectly correlated with expected returns, i.e., unless $corr(v_{t+1}, w_{t+1}) = \pm 1$, and $\alpha = \beta$. Conversely, with imperfect predictors the full history of both the predictors and returns will generally matter to the conditional expectation of returns and can be computed from a weighted sum of past innovations to these variables. Pastor and Stambaugh (2009) develop tests for serial correlation in predictive return regressions which can be used to detect the presence of imperfect predictors. Moreover, they show that the sign of the correlation between residuals of the expected and unexpected return equations can be used as a diagnostic for an imperfect predictor.

4.3 Tracking model change

Different strategies for dealing with model instability in the area of financial forecasting have been considered. A simple, yet general representation of model- or

[13] For simplicity, we assume a univariate predictor, x_t, but Pastor and Stambaugh (2009) allow for multiple predictors.

parameter instability is to modify (3) to allow for a time-varying slope coefficient:

$$r_{t+1} = \mu + \beta_t x_t + u_{t+1}. \tag{22}$$

Several approaches fall under this umbrella. For example, we could specify the regression parameters as a (linear) function of observables, z_t, e.g., $\beta_t = \beta_0 + \beta_1 z_t$, in which case (22) becomes

$$r_{t+1} = \mu + \beta_0 x_t + \beta_1 z_t x_t + u_{t+1}, \tag{23}$$

see, e.g., Christopherson, Ferson and Glassman (1998) and Ferson and Schadt (1996) for applications to evaluation of fund performance.

Alternatively, one can assume that the parameters of the linear regression model (3) are latent variables that follow a process

$$\beta_t = \beta_{t-1} + \omega_t, \tag{24}$$

where $\omega_t \sim N(0, Q_t)$. This time-varying parameter model nests (3) as a special case if $Q_t = 0$ for all t.[14] Johannes, Korteweg, and Polson (2014) propose a more complex version of the time-varying parameter model that allows for mean reverting stochastic volatility and mean-reverting coefficients:

$$\begin{aligned} r_{t+1} &= \mu + \beta_0 x_t + \beta_{t+1} x_t + \sqrt{V^r_{t+1}} u_{t+1}, \\ \beta_{t+1} &= \rho \beta_t + \sigma_\beta \varepsilon_{\beta t+1}, \quad \varepsilon_{\beta t+1} \sim iidN(0,1) \\ \log(V^r_{t+1}) &= \alpha_r + \beta_r \log(V^r_t) + \sigma_r \eta_{rt+1}, \end{aligned} \tag{25}$$

Henkel, Martin, and Nardari (2011) consider Markov switching vector autoregressive (MSVAR) models to predict stock market returns. Let $y_{t+1} = (r_{t+1}, x'_{t+1})'$, so that r_{t+1} is the first element of the y_{t+1} vector. Then the MSVAR(1) model can be written as

$$y_{t+1} = \mu_{s_{t+1}} + A_{s_{t+1}} y_t + u_{t+1}, \tag{26}$$

where $u_{t+1} \sim N(0, \Sigma_{s_{t+1}})$, $\mu_{s_{t+1}}$ is now a column vector, $A_{s_{t+1}}$ is a state-dependent VAR matrix, $s_{t+1} \in \{1, 2\}$, and $P(s_{t+1} = j | s_t = i) = p_{ij}$. Hence, there is a constant probability of staying in one of the two states and both the mean and variance shift across the states. Note that the two states repeat in this model. This assumption is relaxed by Pettenuzzo and Timmermann (2011) who use the Chib (1998) change-point process to predict stock returns. In particular, they allow for K states so that $s_t \in \{1, ..., K\}$ in the sample $t = 1, ..., T$ and assume that the same state never repeats so that the state transition probability takes the form

$$P = \begin{pmatrix} p_{11} & p_{12} & 0 & \cdots & 0 \\ 0 & p_{22} & p_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & & p_{KK} \end{pmatrix}. \tag{27}$$

[14] See Dangl and Halling (2012) for an application of this type of model to stock market return predictability.

This model characterizes shifts between the K states during the historical sample. To predict future returns out-of-sample, Pettenuzzo and Timmermann (2011) assume that the new parameters after a break are drawn from a "meta distribution" which means that they can forecast returns out-of-sample in a way that accounts for the possible effect of future breaks. This requires making assumptions about the underlying stochastic process that determines both the frequency and magnitude of future breaks.

Empirically, many of these studies find that it is important to account for time variation in the parameters of return prediction models. Dangl and Halling (2012) find that the random walk coefficients model (24) quickly adapts to changes in the underlying return generating process. Johannes et al. (2014) also find that it is important to generalize the standard model to allow for time-varying coefficients and volatility clustering. Moreover, using the more general models to forecast returns appears to lead to improvements in portfolio performance. Henkel et al. (2011) find that return predictability in the stock market is closely linked to economic recession periods and that some predictors only have predictive power during recessions. Pettenuzzo and Timmermann find that accounting for the effects of past breaks and the possibility of future breaks in return prediction models can lead to significant economic gains when the resulting forecasts are used for asset allocation decisions.

A very different approach for handling model instability is to attempt to estimate the size and magnitude of recent breaks affecting the forecasting model and use this information to compute improved forecasts. To dampen the effect of estimation error on forecast computed on short data samples in the aftermath of a break, Pesaran and Timmermann (2007) propose using both pre- and post-break data, perhaps downweighting pre-break data to reduce biases in the resulting parameter estimates.

Alternatively, Bayesian methods that pull parameter estimates towards an economically reasonable prior can be used. This approach is akin to shrinking the post-break parameter estimates towards the prior and letting the degree of shrinkage taper off as more data points are cumulated in a new post-break regime.

There are clear limitations from time-series approaches which tend to have weak power to detect structural breaks. Moreover, some predictability may reflect low-frequency movements (e.g., business cycle variation) that affect the vast majority of firms, stocks, and countries and whose predictability can be difficult to ascertain given the paucity of observations at longer horizons.

Panel forecasting methods that exploit cross-sectional information to increase the power of the tests for instability can alternatively be used. Such methods can be expected to work better if the timing of breaks across cross-sectional units is not too heterogeneous, i.e., if breaks are common. Pooling cross-sectional and time-series information in this manner appears to be a promising venue for handling model instability in financial forecasting models.[15]

[15] See Smith and Timmermann (2017) for a recent application of this idea to a panel of stock

4.4 Bayesian Methods

Bayesian methods provide a coherent framework for handling parameter estimation error, model uncertainty and model instability. This can perhaps best be illustrated in the context of the simple linear regression model

$$r_{t+1} = \mu + \beta' x_t + u_{t+1}, \quad u_{t+1} \sim N(0, \sigma_u^2), \quad t = 1, ..., T-1, \tag{28}$$

where x_t is a vector of predictors. Denoting the conditioning data set by $Z_t = \{x_\tau, y_\tau\}$, $\tau = 1, ..., t$, we can integrate out uncertainty about the underlying parameters μ, β, and σ_u^{-2} to obtain the posterior predictive density:

$$p(r_{t+1}|Z_t) = \int p(r_{t+1}|\mu, \beta, \sigma_u^{-2}, Z_t) p(\mu, \beta, \sigma_u^{-2}|Z_t) d\mu d\beta d\sigma_u^{-2}. \tag{29}$$

Parameter uncertainty can be handled in a two-step procedure. In the first step, values of the parameters $\mu, \beta, \sigma_u^{-2}$ are drawn from $p(\mu, \beta, \sigma_u^{-2}|Z_t)$ using a Gibbs sampler. Under a set of standard normal-gamma priors, draws from the joint posterior distribution iterate forth and back between the distributions for the mean parameters (μ, β) and the precision parameter (σ_u^{-2}). Given these parameter values, in the second step, draws from the outcome distribution are simple to implement because, conditional on the first-step parameter values,

$$r_{t+1}|\mu, \beta, \sigma_u^{-2}, Z_t \sim N(\mu + \beta' x_t, \sigma_u^{-2}). \tag{30}$$

Model uncertainty can also readily be handled in the Bayesian framework. Suppose there are n models of the form in (28), i.e.,

$$r_{t+1} = \mu_j + \beta_j' x_{jt} + u_{jt+1}, \quad u_{jt+1} \sim N(0, \sigma_{ju}^2), \quad t = 1, ..., T-1, \tag{31}$$

where again x_{it} is a vector of predictors and $x_{it} \neq x_{jt}$ for $i \neq j$. Then we can generate a draw from the predictive density by weighting each of the n models by their posterior probabilities given the data, $p(M_j|Z_t)$:

$$p(r_{t+1}|Z_t) = \sum_{j=1}^{n} p(r_{t+1}|M_j, Z_t) \times p(M_j|Z_t). \tag{32}$$

Here $p(r_{t+1}|M_j, Z_t)$ is the posterior predictive density of model j which can be computed using similar steps as in equations (29) and (30), whereas $p(M_j|Z_t)$ can be computed using steps described in the section on Bayesian Model Averaging.

As a simple application, suppose we are considering models with different number of breaksor regimes. For example, $p(r_{t+1}|M_j, Z_t)$ could correspond to a model with j breaks. Then we can use equation (32) to integrate out uncertainty about the number of breaks, weighting the posterior predictive density of model M_j by its posterior probability.

---

market portfolios.

4.5 Machine learning methods

New methods for conducting predictive analytics, often in the context of data sets that have large cross-sectional or time-series dimensions, or both, have been developed in recent years–see Hastie, Tibshirani, and Friedman (2009) for an excellent introduction. These methods offer ways to flexibly estimate predictive models without imposing strong assumptions on their functional form and accounting for situations with large sets of predictors.

Key to the successful use of these methods appears to be control of their tendency to overfit due to their flexibility. The low signal-to-noise ratio encountered in many financial forecasting problems raises the risk that flexible fitting methods will simply fit noise in a given sample and be strongly affected by estimation error, more so than simpler linear models.

As an example of one of the recent "machine learning" methods, consider regression trees which provide piecewise (constant) approximations to an unknown functional form by splitting the state space into a set of J disjoint regions $S_1, S_2, ..., S_J$. Within each region, the function takes a constant value, μ_j, so that

$$\Upsilon(x_t, \Theta_J) = \sum_{j=1}^{J} \mu_j I(x_t \in S_j), \tag{33}$$

where x_t is a vector of predictor variables at time t and $\Theta_J = \{S_j, \mu_j\}$, $j = 1, ..., J$ is the set of parameters used to carve out the state space and the values of the constants, μ_j. $I(x_t \in S_j)$ is an indicator variable that equals one if $x_t \in S_j$, and otherwise equals zero. Finding the optimal way to discretize the sample space into the J regions is the tricky part, particularly if the dimension of the vector of predictor variables, x_t, is high. In contrast, once the boundaries for the state space have been determined, the values of μ_j are easy to estimate under conventional loss functions such as mean squared error (MSE) or mean absolute error (MAE).

Individual regression trees can be "boosted" by fitting successive trees to the residuals that remain after previous rounds of tree fitting. Specifically, boosted regression trees proceed iteratively by partitioning the data into sub regions given the preceding splits of the data. Trees are added with the objective of obtaining a better fit in regions that were poorly fitted by the initial trees. With B such boosting steps, a boosted regression tree can be obtained

$$f_B(x_t) = f_{B-1}(x_t) + \Upsilon(x_t, \Theta_{J,B}) = \sum_{b=1}^{B} \Upsilon(x_t, \Theta_{J,b}), \tag{34}$$

where $\Upsilon(x_t, \Theta_{J,b})$ is the regression tree fitted in the bth iteration. Parameter estimates of these trees can be obtained as the solution to an optimization problem of the form

$$\hat{\Theta}_{J,b} = \arg\min_{\Theta_{J,b}} \left(y_{t+1} - \left(f_{b-1}(x_t) + \Upsilon(x_t, \Theta_{J,b})\right)\right)^2, \tag{35}$$

assuming squared error loss.

Ensemble learning methods can be employed to address some of these concerns. To reduce the risk of overfitting, shrinkage can be used so that each tree only contributes a small amount to the overall fit. For example, using a value of $\lambda = 0.001$, (34) can be replaced by

$$f_B(x_t) = f_{B-1}(x_t) + \lambda \Upsilon(x_t, \Theta_{J,B}). \tag{36}$$

Replacing the conventional squared-error loss function in (35) with a mean absolute error loss function $T^{-1} \sum_{t=0}^{T-1} |r_{t+1} - f_B(x_t)|$ can also help to provide more robust results as the tendency to fit outliers is reduced by this loss function. Finally, subsampling methods have been shown to improve forecasting performance.

One advantage of methods such as boosted regression trees over conventional nonparametric methods is that they are less subject to the "curse of dimensionality" and, in fact, can be used to handle cases with large-dimensional predictors. Of course, this ability comes at a cost. First, there is no guarantee that the method will identify a global optimum in a high-dimensional parameter optimization problem. Second, this flexibility means that the risk of overfitting can be quite large.

While machine learning methods will undoubtedly be used extensively to "mine" financial data sets in the hope of finding empirical regularities that can be used in trading algorithms, it is important to be aware of their limitations. Several non-parametric and semi-parametric methods (e.g., sieve estimation) have been accessible to researchers for a long time and have not yet become part of applied financial researchers' toolkit. Often financial data sets have a limited time-series span that can be dominated by rare events such as the global financial crisis of 2007-2008 in which relationships between predictors and the dependent variable may become unstable. While model instability will have adverse effects on any forecasting approach, flexible methods that require large samples to obtain accurate estimates of the conditional mean function can be expected to be particularly strongly affected by shifts in the underlying data generating process.

4.6 Exploiting information from economic theory to restrict forecasts

One way to address the adverse effects of estimation error and model uncertainty is to use economic theory or priors to restrict the functional form and/or the values of the parameters of the forecasting model. For example, Pastor and Stambaugh (2009) use informative priors to constrain the sign of the correlation between shocks to expected and unexpected returns in their predictive systems analysis. Similarly, no-arbitrage constraints have been used to impose restrictions on dynamic Gaussian affine term structure models for bond returns by Ang and Piazzesi (2003) and Sarno et al. (2016).

Pettenuzzo, Timmermann and Valkanov (2013), use economic constraints to modify the posterior distribution of the parameters of the linear return regression (3) in a way that allows the return prediction model to learn from the

data. They consider two types of constraints, namely a constraint that imposes non-negative equity premia and a constraint that bounds the conditional Sharpe ratio and incorporates time-varying volatility in the predictive regression.

Note that the linear regression model in (3) does not rule out negative equity premium forecasts if unconstrained parameter estimates $\hat{\mu}, \hat{\beta}$ are used to generate forecasts, i.e., we could have

$$\hat{r}_{\tau+1|\tau} = \hat{\mu} + \hat{\beta} x_\tau < 0, \quad \tau = 1, ..., t-1. \tag{37}$$

Such negative forecasts of excess returns may not seem reasonable as we would expect risk-averse investors to command a positive risk premium for holding the stock market portfolio. To address this point, following Campbell and Thomson (2008) suppose that the equilibrium equity premium is expected to be non-negative, i.e.,

$$E[r_{\tau+1}|\Omega_t] = \mu + \beta x_\tau \geq 0 \quad \text{for } \tau = 1, ..., t. \tag{38}$$

Noting that the constraint in (38) must hold at all points in time, this yields t constraints in a sample with t observations.

Alternatively, we can impose bounds on the conditional Sharpe ratio, i.e., the ratio of the conditional mean over the conditional volatility. To account for time-varying volatility, the predictive regression in (3) can be modified as follows

$$r_{\tau+1} = \mu + \beta x_\tau + \exp(h_{\tau+1}) u_{\tau+1}, \quad u_{\tau+1} \sim N(0,1), \tag{39}$$

where $h_{\tau+1}$ is the log return volatility at $\tau + 1$. The dynamics of $h_{\tau+1}$ could, for example, be specified as a random walk model, $h_{\tau+1} = h_\tau + \xi_{t+1}$, $\xi_{t+1} \sim N(0, \sigma_\xi^2)$.

Bounds on the conditional Sharpe ratio, $SR_{\tau+1|\tau} = (\mu + \beta x_\tau)/\exp(h_\tau + 0.5\sigma_\zeta^2)$, take the form

$$SR^l \leq SR_{\tau+1|\tau} \leq SR^u \quad \text{for } \tau = 1, ..., t. \tag{40}$$

The bounds in equation (40) indirectly restrict the parameters $\theta = (\mu, \beta, \sigma_\xi^2)$ as well as the sequence of log return volatilities $h^t \equiv \{h_1, h_2, ..., h_t\}$.[16]

Pettenuzzo et al. (2014) assume standard Gaussian priors on the regression parameters and use a Gibbs sampler to estimate the forecasting model in (39). Empirically, they find that imposing economic constraints succeeds in reducing uncertainty about model parameters and reduces the risk of selecting a poor forecasting model. When evaluated in an out-of-sample forecasting analysis, the constrained models improve on both statistical and economic measures of forecasting performance.

An alternative approach advocated by Ferreira and Santa-Clara (2011) is to disaggregate returns into cash-flow and price-related components which can then be predicted separately. Ferreira and Santa-Clara decompose the continuously

[16] In practice, PTV impose bounds on the annualized Sharpe ratio of $SR^l = 0$, $SR^u = 1$.

compounded returns, r_{t+1}, into growth in the (log-) price earnings multiple, gm_{t+1}, growth in log earnings, ge_{t+1}, and the log dividend yield, dp_{t+1} :

$$r_{t+1} = gm_{t+1} + ge_{t+1} + dp_{t+1}. \tag{41}$$

Using this "sum of parts" approach, Ferreira and Santa-Clara forecast the three return components assuming no change in the price-earnings multiple so $E[gm_{t+1}] = 0$ (simplest model), a 20-year moving average for growth in earnings, and a random walk for the dividend yield.[17] Their empirical results suggest that forecasting the disaggregated return components in this manner works well out-of-sample. One reason this approach is found to work well is that it is akin to shrinkage towards current values (for the price-earnings multiple and the dividend yield) and a slowly moving estimate of earnings growth. These choices reduce the effect of estimation error on the forecasts, which is an important consideration.

4.7 Cross-validation (Out-of-sample) methods

Overfitting and data mining issues arise because of the correlation between the forecast error and the estimation error which causes the estimated MSE in (11) to be an underestimate of the (true) performance we would expect in a new sample. Cross-validation methods can be used to remove this correlation. Although these methods use the full data set, they avoid using the same data for model fitting (and selection) and for forecast evaluation.

One version of cross-validation that is particularly popular in finance (and in economics more broadly) is to split the sample $t = 1, ..., T$ into an in-sample portion $t = 1, ..., R$ $(R < T)$ used for model estimation and selection and an out-of-sample portion, $t = R + 1, ..., T$, used for forecast evaluation.

To see how out-of-sample forecasting methods can help reduce concerns of overfitting, consider the simple linear regression model

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N0, I_T), \tag{42}$$

where Y and ε are $T \times 1$ vectors of the dependent variable and regression residuals, respectively, and the predictors have been stacked into a $T \times k$ matrix (X) that has been rotated and standardized so that $X'X = I_k$. In this setting, Hansen and Timmermann (2015b) show that the in-sample (IS) residual sum of squares from the forecasting model is given by

$$RSS_{IS} = \varepsilon'\varepsilon - \varepsilon' XX'\varepsilon. \tag{43}$$

To capture the separation of the estimation and evaluation samples used by out-of-sample methods, suppose that β is estimated from an independent sample $\tilde{Y} = X\beta + \tilde{\varepsilon}$, with $\tilde{\varepsilon} \sim N(0, I_T)$ being independent of ε. The resulting out-of-sample (OoS) residual sum of squares is then given by

$$RSS_{OoS} = \varepsilon'\varepsilon + \tilde{\varepsilon}' XX'\tilde{\varepsilon} - 2\varepsilon' XX'\tilde{\varepsilon}. \tag{44}$$

[17] Ferreira and Santa-Clara (2011) also consider alternative forecasting models such as a linear regression that incorporates mean reversion in the price-earnings multiple.

When evaluating the in-sample and out-of-sample measures of forecasting performance in (43) and (44), note that the RSS computed using the true value of β is $\varepsilon'\varepsilon$. Labeling this value RSS_{true}, from (43) the overfit from using the in-sample measure of forecasting performance is given by

$$RSS_{true} - RSS_{IS} = \varepsilon' XX'\varepsilon \sim \chi^2_k, \tag{45}$$

see Hansen and Timmermann (2015b). This result makes intuitive sense: the more predictors are included in the forecasting model, the greater the potential for overfit, and the degree of overfit grows in direct proportion with the number of predictor variables, k. The key is to notice that the quadratic form $\varepsilon' XX'\varepsilon$ gets *subtracted* from the RSS so that estimation error wrongly contributes to overfitting.

In contrast, from equation (44) the overfit associated with the out-of-sample forecast evaluation scheme is (Hansen and Timmermann (2015b))

$$RSS_{true} - RSS_{OoS} = -\tilde{\varepsilon}' XX'\tilde{\varepsilon} + 2\varepsilon' XX'\tilde{\varepsilon}. \tag{46}$$

Recalling that ε and $\tilde{\varepsilon}$ are independent so the second term has zero mean, we see that the expectation of the overfit is negative and equals $-\chi^2_k$. Hence, parameter estimation error serves to *reduce* the out-of-sample forecasting performance, as we would expect.

Besides reducing the problem arising from individual models' tendency to overfit, out-of-sample forecast evaluation tests have the advantage that they can help uncover periods in which a forecasting model produces accurate forecasts as well as periods in which it fails to do so–see Giacomini and Rossi (2009, 2010) and Rossi (2013). This is particularly useful in financial forecasting where we would expect forecasting models only to work well during limited periods of time due to competitive market pressure. Closely related to this, if the out-of-sample forecasts are careful in using only information and forecasting methods that were available in real time, evidence from such tests can be used to test market efficiency.

While out-of-sample forecast evaluation methods can be an effective way to reduce overfitting, they are no panacea and also introduce new problems. First, nothing prevents individual researchers–or a group of researchers, each of whom only considers a single model–from experimenting with a multitude of forecasting models. Picking the model with the best out-of-sample forecasting performance distorts inference if not properly accounted for. Data mining is not impossible in out-of-sample forecast experiments–it is just a lot harder than when conducted in-sample. As an illustration, simple calculations in Hansen and Timmermann (2015b) show that for a regression model with $k = 4$ predictors, to achieve a rejection rate of 5% would take close to 5,000 different models in a sample with $T = 50$ observations. In stark contrast, the same rejection rate can be accomplished with a single model when based on in-sample forecasting performance.

Second, out-of-sample tests of forecasting performance can have substantially weaker power than in-sample tests. This is a straightforward implication

of using fewer observations $(T-R)$ to evaluate predictive accuracy than the corresponding in-sample tests (T). Inoue and Killian (2005) and Hansen and Timmermann (2015a) analyze the importance of this point. For nested regression models, Hansen and Timmermann (2015b) show that a standard test statistic for out-of-sample forecasting performance can be written as the difference between two Wald statistics of the same null hypothesis, namely a test that uses the full sample and a test that uses only a subsample. Subtracting a test statistic in this manner clearly reduces the power of the test and so results in less of an ability to identify predictability in out-of-sample experiments.

Third, how the sample split point that separates the in-sample and out-of-sample periods, i.e., the value of R, is chosen may affect the results. Hansen and Timmermann (2012) analyze the effect of data mining over the sample split point. Let $\rho \in (0,1)$ be the in-sample fraction of the data so that $R_\rho = \lfloor \rho T \rfloor$ observations are used for parameter estimation while $T - R_\rho$ observations are used for forecast evaluation. Hansen and Timmermann show that search over the sample split point, ρ, can lead to severe distortions in inference if not accounted for when evaluating predictive performance. In particular, they find that large values of ρ are more likely to be selected in experiments that choose ρ so as to maximize conventional test statistics of out-of-sample forecasting performance. This finding reflects the presence of the second term in (46) which has zero mean, but will have a larger variance when the out-of-sample period is short and so is likely to result in higher rejection rates for tests of out-of-sample forecasting performance. This point can matter a great deal in practice: Hansen and Timmermann (2012) find that the rejection rate for an out-of-sample test with a nominal size of 5% can be quadrupled with only 3-4 predictor variables as a result of inflation in the test statistic induced by mining over the sample split point.

4.8 Accounting for the multiple hypothesis testing problem

Out-of-sample forecast evaluation methods can significantly reduce the tendency of models to overfit which is associated with in-sample forecast evaluation. However, if many different models are being considered and this is not accounted for when evaluating the performance of the "best" forecasting model, severe distortions of inference on predictability may ensue.

This problem is analyzed in White (2000) who also provides an elegant solution to it. Suppose that there are k forecasting models whose out-of-sample mean squared error performances is measured relative to some benchmark (model 0) so that the mean squared error (MSE) differential of the benchmark relative to the kth forecasting model is given by $\bar{d}_k = (T-R)^{-1}\left[\sum_{t=R}^{T-1}(y_{t+1} - f_{0t+1|t})^2 - (y_{t+1} - f_{kt+1|t})^2\right.$ Stacking the MSE differentials for the k models gives a vector of sample averages $\bar{d} = (\bar{d}_1, ..., \bar{d}_k)'$. Since $\bar{d}_k > 0$ indicates that the kth forecasting model has outperformed the benchmark, the null hypothesis that none of the models

is capable of beating the benchmark takes the form

$$H_0 : \max_{j=1,\ldots,k} E[d_{jt+1|t}(\beta^*)] \leq 0, \tag{47}$$

see White (2000). Note that the null hypothesis is being evaluated at the probability limit of the parameter values β^*. If all forecasting models produce as accurate forecasts as those from the benchmark, the $\bar{d}$ vector, as well as the maximum of this vector, has a mean of zero.

To evaluate if the null in (47) is satisfied requires calculating the sampling distribution of the maximum of a k dimensional vector of possibly correlated terms. To this end, White (2000) establishes conditions under which $\sqrt{(T-R)}(\bar{d} - E[d(\beta^*)])$ converges to a multivariate normal distribution with an unknown covariance matrix, Λ. This insight reduces tests of the null in (47) to the task of evaluating whether the performance of the best forecasting model, chosen from a set of k models, is better than one would expect by chance from picking the maximum value drawn from a k-dimensional normal distribution with covariance matrix Λ.

To perform this "skill versus luck" calculation, White shows that a bootstrap that repeatedly draws resamples from the underlying forecast data can be used to evaluate whether the performance of the best model in the actual data is genuinely better than what one would expect as a result of having picked such a model from a possibly large-dimensional vector of forecasts. Importantly, this approach sidesteps the need for estimating the covariance matrix, Λ, which could be complicated in situations where k is large.

In an empirical application of the Reality Check bootstrap, Sullivan, Timmermann, and White (1999) consider the performance of 7,846 technical trading rules applied to daily returns on stock market indexes or futures prices. Their list of technical trading rules includes filter rules, moving averages, support and resistance type rules, and channel breakouts. While the best technical trading rule generates performance that is highly statistically significant when evaluated in isolation, i.e., as if only a single model had been considered, Sullivan et al. (1999) also find that the best trading rule fails to outperform a buy-and-hold benchmark once one accounts for the fact that this model was selected as the best performer from a larger set.

White (2000)'s approach has been expanded in a number of interesting directions. For example, Romano and Wolf (2005) develop a step-wise approach that performs White's bootstrap iteratively so as to identify *all* models with superior performance over a benchmark, as opposed to only testing if there exists at least one model that beats the benchmark. Romano and Wolf provide conditions under which their step-wise approach can asymptotically select all superior models (and eliminate inferior models) subject to controlling the family-wise error rate, i.e., the probability of incorrectly identifying at least one forecasting model as being superior to the benchmark.

Hansen, Lunde, and Nason (2011) develop a model confidence set approach that identifies the subset of forecasting models that, at a given level of confidence, includes the best-performing forecasting models, i.e., models whose per-

formance is not dominated by any other models in a pairwise test, or measured relative to the average performance of other models. Their approach involves an equivalence test for pairwise testing that model performance is identical, and an elimination rule for determining which model, if any, gets discarded from the model confidence set in a given step.

Both of these approaches hold considerable promise for applications in financial forecasting problems in which it is important to deal with the multiple hypothesis testing problem.

5 Volatility and density forecasting

A vast empirical literature shows evidence of persistence in the volatility of returns on a variety of asset classes (e.g., stocks, bonds, commodities, and currencies) and at different frequencies (e.g., daily, weekly, or monthly). Such persistence implies that time variation in return volatility is predictable.[18]

Early empirical evidence emanated from the ARCH and GARCH models proposed by Engle (1982) and Bollerslev (1986), respectively. GARCH models take the form

$$\begin{aligned} \varepsilon_{t+1} &= \sigma_{t+1|t}\eta_{t+1}, \quad \eta_{t+1} \sim iidN(0,1), \\ \sigma^2_{t+1|t} &= \omega + \sum_{i=1}^{p} \beta_i \sigma^2_{t+1-i|t-i} + \sum_{i=1}^{q} \alpha_i \varepsilon^2_{t+1-i}. \end{aligned} \tag{48}$$

The recursive form of this specification means that it is straightforward to generate volatility forecasts, iterating to obtain a forecast of the $h-$step-ahead volatility $\sigma^2_{t+h|t}$ from the $(h-1)$-step-ahead forecast, $\sigma^2_{t+k-1|t}$.

The GARCH(1,1) model

$$\sigma^2_{t+1|t} = \omega + (\alpha_1 + \beta_1)\sigma^2_{t|t-1} + \alpha_1 \sigma^2_{t|t-1}(\eta^2_t - 1) \tag{49}$$

has proven highly successful in many empirical applications (see Hansen and Lunde (2005)). For this model the persistence of the conditional variance can be measured by $\alpha_1 + \beta_1$ as $E_{t-1}[\eta^2_t - 1] = 0$. It is not uncommon to find estimates for which $\hat{\alpha}_1 + \hat{\beta}_1$ is slightly below one, indicating a high degree of persistency, particularly at high frequencies. This is consistent with the presence of a strong predictive component in the second moment of the return distribution. Assuming that $\alpha_1 + \beta_1 < 1$, the steady-state (average) variance from the GARCH process in (49) is

$$E[\sigma^2_{t+1|t}] \equiv \sigma^2 = \frac{\omega}{1 - \alpha_1 - \beta_1},$$

and a forecast of the conditional one-period variance h periods from now, given current information, can be computed as

$$\sigma^2_{t+h|t} \equiv E_t[\sigma^2_{t+h}] = \sigma^2 + (\alpha_1 + \beta_1)^{h-1}(\sigma^2_{t+1|t} - \sigma^2),\ h \geq 1.$$

[18] See Andersen et al. (2006) for an excellent review of volatility forecasting methods.

Regime switching models offer an alternative way of capturing persistence in volatility. Consider the simple regime-switching model with a state-dependent mean and volatility:

$$r_{t+1} = \mu_{s_{t+1}} + \sigma_{s_{t+1}}\eta_{t+1}, \quad \eta_{t+1} \sim iidN(0,1), \tag{50}$$

where $\mu_{s_{t+1}}$ and $\sigma_{s_{t+1}}$ are the mean and volatility in state s_{t+1}. As in the literature on return predictability, typically it is assumed that there is a small set of possible states, i.e., $s_{t+1} \in \{1, 2, ..., K\}$ for a small value of K. For $K = 2$ there are only two possible states–often identified as high- and low-volatility states when these models are fitted to asset returns. Persistence in the level of volatility can be modeled by assuming that s_{t+1} follows a first-order homogeneous Markov chain

$$\Pr(s_{t+1} = j | s_t = i) = p_{ij}. \tag{51}$$

The Markov chain is persistent provided that $p_{11} + p_{22} > 1$ in the case with two states. Time-variation in the persistence of return volatility can be captured by using dynamic state transition probabilities of the form

$$\Pr(s_{t+1} = j | s_t = i, z_t) = p_{ij}(z_t), \tag{52}$$

where $p_{ij}(z_t)$ could be specified as a logit or probit function so as to ensure that $p_{ij}(z_t) \in [0, 1]$.

GARCH and Markov switching models can be used to capture features of financial data such as fat tails (kurtosis) and skew. For example, time-varying skews can be incorporated into the GARCH model by allowing positive and negative shocks to have a different effect on future volatility as captured by the threshold GARCH model of Glosten, Jagannathan, and Runkle (1993):

$$\sigma^2_{t+1|t} = \omega + \alpha_1\varepsilon_t^2 + \lambda\varepsilon_t^2 I(\varepsilon_t < 0) + \beta_1\sigma^2_{t|t-1}, \tag{53}$$

where $I(\varepsilon_t < 0)$ is an indicator variable that takes a value of one if $\varepsilon_t < 0$, and otherwise equals zero.

Similarly, the return distribution generated by the MS model in (50) with two states will be skewed provided that $\mu_1 \neq \mu_2$.

GARCH and Markov switching specifications such as (48) and (50) are "adaptive" in that they do not predict the initial large shock, i.e., an outlier in ε_{t+1} that is not preceded by a large value of ε_t. However, once such a shock has occurred, these models will elevate their forecast of next periods' variance using a decay rate that reflects mean reversion provided that $\alpha_1 + \beta_1 < 1$ in the GARCH model (49) or $p_{ii} < 1$ in the MS model in (51). This property, coupled with the persistence of the underlying volatility process, accounts for the empirical success of these classes of models when used to model volatility of asset returns.

5.1 Realized volatility

Realized volatility models make use of data sampled at a higher frequency than that typically used to estimate GARCH models. For example, a realized volatility model could be based on data sampled every five minutes during the market opening hours of the trading day, whereas a GARCH model typically uses daily returns data.[19] To see how realized volatility models work, suppose that changes in log-prices, dp_t, evolve according to an arithmetic random walk process with constant drift μ and time-varying volatility σ_t :

$$dp_t = \mu dt + \sigma_t dW_t, \tag{54}$$

where dW_t are increments to an underlying Wiener process. The volatility of asset returns over some interval $[t-1, t]$ is unobserved but can be approximated by sampling returns on a discrete grid of points $t-1 < \tau_0 < \tau_1 < .... < \tau_N = t$. This gives rise to the realized variance[20]

$$RV_t = \sum_{j=1}^{N} (dp_{\tau_j} - dp_{\tau_{j-1}})^2. \tag{55}$$

We would expect RV_t to be a somewhat noisy estimate of the underlying volatility for small values of N. However, because the realized variance can be viewed as a noisy proxy of the unobserved (underlying) volatility process, one can imagine using it as the basis for a forecasting equation

$$RV_{t+1} = \alpha + \sum_{j=1}^{J} \beta_j RV_{t+1-j} + \varepsilon_{t+1}. \tag{56}$$

A notable example of this model that has been extensively used in empirical work is the HAR-RV model proposed by Corsi (2009). Corsi proposes projecting the daily integrated volatility ($\tilde{\sigma}^2_{t+1}$) on lagged values of daily (RV^d_t), weekly (RV^w_t), and monthly (RV^m_t) realized volatility estimates:

$$\tilde{\sigma}^2_{t+1} = \alpha + \beta_d RV^d_t + \beta_w RV^w_t + \beta_m RV^m_t + \varkappa_{t+1}, \tag{57}$$

where, for example, $RV^w_t = \frac{1}{5}(RV^d_t + RV^d_{t-1} + RV^d_{t-2} + RV^d_{t-3} + RV^d_{t-4})$ is the weekly average realized volatility. Using past realized volatilities at different frequencies often improves the predictive power over a model that only includes a few lags of the daily realized volatility measure.

Alternatively, RV_t can be added as an additional covariate to GARCH models or modeled jointly with the log-variance from a GARCH process to get a model such as

$$\begin{aligned} \log \sigma^2_{t+1|t} &= \omega + \beta \log \sigma^2_{t|t-1} + \gamma RV_t, \\ \log RV_t &= \delta_0 + \delta_1 \log h_{t|t-1} + \delta_2 \log RV_{t-1} + v_t \end{aligned} \tag{58}$$

[19] See, e.g., Ait-Sahalia et al. (2005) for an analysis of how often data should be sampled when it is affected by market microstructure noise.

[20] Properties of the estimated realized variance such as consistency are discussed by Hansen and Lunde (2011).

as proposed by Hansen et al. (2012).[21]

High frequency data on price movements gives researchers the ability to better decompose total volatility into separate continuous volatility and jump volatility components. This can be important for forecasting purposes because the two components possess very different time-series properties. Notably, jump volatility tends to be much less persistent than continuous volatility. See Andersen, Bollerslev and Diebold (2007) and Patton and Shephard (2015) for examples of this approach.

Some studies suggest an important role in volatility forecasting for implied volatility measures extracted from options data. For example studies such as Blair, Poon and Taylor (2001) find that once information in implied volatilities is used to predict future volatility, there is little or no additional information in high-frequency return movements.

5.2 Stochastic volatility

GARCH models such as (48) assume that volatility is driven by past and current shocks to returns which are observable and, thus, can be estimated by maximum likelihood methods subject to assumptions on the value of the initial volatility state. Stochastic volatility models instead assume that the volatility process is affected by a sequence of volatility-specific shocks, ζ_t.

The basic stochastic volatility model takes the form

$$r_{t+1} = \mu + \exp(h_{t+1})\eta_{t+1}, \quad \eta_{t+1} \sim iidN(0,1), \tag{59}$$

where the log-volatility at time $t+1$, h_{t+1}, can be modeled as a stationary, mean-reverting process:

$$h_{t+1} = \lambda_0 + \lambda_1 h_t + \zeta_{t+1}, \quad \zeta_{t+1} \sim iidN(0, \sigma_\zeta^2) \tag{60}$$

for $|\lambda_1| < 1$. The shocks η_τ and ζ_s are mutually independent for all values of τ and s. It is also not uncommon to impose that $\lambda_0 = 0$ and $\lambda_1 = 1$, which turns (60) into a driftless random walk model.

Kim et al. (1998) develop methods for estimation of stochastic volatility models. The presence of a volatility-specific shock sequence ζ_s in (60) means that conventional maximum likelihood methods cannot be used for estimation. Instead, the Gibbs sampler can be used to obtain draws from the joint posterior distribution $p(\mu, h^t, \lambda_0, \lambda_1, \sigma_\zeta^{-2})$, where $h^t = (h_1, ..., h_t)$. With these in place, one can obtain draws from the predictive density

$$\begin{aligned} p(r_{t+1}|\Omega_t) \quad = \quad & \int p(r_{t+1}|h_{t+1}, \mu, h^t, \lambda_0, \lambda_1, \sigma_\zeta^{-2}, \Omega_t) \\ & \times p(h_{t+1}|\mu, h^t, \lambda_0, \lambda_1, \sigma_\zeta^{-2}, \Omega_t) \\ & \times p(\mu, h^t, \lambda_0, \lambda_1, \sigma_\zeta^{-2}|\Omega_t) d\mu dh^{t+1} d\lambda_0 d\lambda_1 d\sigma_\zeta^{-2}, \end{aligned} \tag{61}$$

[21] Paye (2012) finds that univariate specifications of monthly stock market volatility such as (56) are difficult to beat in out-of-sample forecast comparisons when compared to models that add a variety of macroeconomic predictor variables to the forecasting equation.

where $p(r_{t+1}|h_{t+1}, \mu, h^t, \lambda_0, \lambda_1, \sigma_\zeta^{-2}, \Omega_t)$ is the predictive density conditional on the model parameters, including the value of the log-volatility, $p(h_{t+1}|\mu, h^t, \lambda_0, \lambda_1, \sigma_\zeta^{-2}, \Omega_t)$ captures shifts in the future log-volatility, h_{t+1}, away from the current log-volatility, h_t, and $p(\mu, h^t, \lambda_0, \lambda_1, \sigma_\zeta^{-2}|\Omega_t)$ measures the effect of parameter uncertainty. Equation (61) can be used to generate draws from the predictive density of returns which can, in turn, be used to compute optimal portfolio weights or to evaluate functions of the conditional return distribution.

As in the case of GARCH models or similar volatility specifications, observable predictors can also be added to the stochastic volatility model. See, e.g., Pettenuzzo et al. (2016) for a mixed data sampling (MIDAS) approach that adds predictor variables to the SV specification in (60) and Ghysels, Sinko, and Valkanov (2007) and Andreou, Ghysels and Kourtellos (2011) for broader summaries of MIDAS models.

5.3 Multivariate volatility models and copulas

Portfolio allocation and risk management problems involve modeling and forecasting the joint distribution of a possibly very large set of asset returns. A number of approaches have been developed to capture predictable time variation both in the probability distribution of individual assets' returns as well as in their joint distribution. Consider how to model the returns on an $n \times 1$ vector of asset returns, r_{t+1}, using a conditional scale-location model

$$r_{t+1} \sim N(\mu_t, H_{t+1|t}). \tag{62}$$

The multivariate dynamic conditional correlation model of Engle and Sheppard (2001) and Engle (2002) takes the form

$$H_{t+1|t} = D_{t+1|t} R_{t+1|t} D_{t+1|t}, \tag{63}$$

where $D_{t+1|t}$ is a diagonal $n \times n$ matrix with time-varying volatilities whose ith diagonal element, $h_{iit+1|t}^{1/2}$, can be generated from a univariate GARCH process of the form

$$h_{iit+1|t} = \omega_i + \sum_{p=1}^{p_i} \alpha_{ip} \varepsilon_{it+1-p}^2 + \sum_{q=1}^{q_i} \beta_{iq} h_{iit+1-q|t-q}. \tag{64}$$

The matrix $R_{t+1|t}$ captures time-varying correlations and is modeled by Engle and Sheppard (2001) as

$$R_{t+1|t} = (Q_{t+1|t}^*)^{-1} Q_{t+1|t} (Q_{t+1|t}^*)^{-1}, \tag{65}$$

where $Q_{t+1|t}^*$ is a diagonal matrix with individual elements $q_{iit+1|t}^{1/2}$ and $Q_{t+1|t}$ follows a GARCH-type dynamic equation

$$Q_{t+1|t} = (1 - \sum_{j=1}^{J} \alpha_j - \sum_{k=1}^{K} \beta_k)\bar{Q} + \sum_{j=1}^{J} \alpha_j \varepsilon_{t+1-j} \varepsilon_{t+1-j}' + \sum_{k=1}^{K} \beta_k Q_{t+1-k|t-k}.$$

As in the univariate case, α_j measures the news impact of past shocks, while β_k captures persistence in correlations.

Scale-location models capture dynamics in the return distribution through movements in the first two moments (mean and variance). A more general approach is provided by modeling changes in the dependencies between returns through their copulas. In particular, under weak conditions it follows from Sklar's Theorem (Sklar, 1959) that the joint return distribution of $r_{t+1} = (r_{1t+1}, r_{2t+1}, ..., r_{nt+1})$ can be decomposed into its univariate marginal distributions $P_1(r_{1t+1}), ..., P_n(r_{nt+1})$ and a copula, C, that incorporates the joint distribution of the n marginal distributions:[22]

$$P(r_{t+1}) = C(P_1(r_{1t+1}), ..., P_n(r_{nt+1})). \tag{66}$$

In a predictive context, it is appropriate to use a conditional copula representation of the form (Patton, 2013):

$$P_{t+1|t}(r_{t+1}) = C_{t+1|t}(P_{1t+1|t}(r_{1t+1}), ..., P_{nt+1|t}(r_{nt+1})). \tag{67}$$

To estimate copula models, one can adopt a two-stage approach which, first, estimates the conditional marginal distributions $P_{jt+1|t}(r_{1t+1})$, $j = 1, ..., n$, using, e.g., a GARCH model and, second, uses the resulting probability integral transforms $P_{jt+1|t}(r_{1t+1})$ as inputs into the conditional copula model $C_{t+1|t}$.

5.4 Density forecasting

Forecasts of the entire probability distribution of the outcome–density forecasts– are becoming increasingly common to use. The simplest class of models is the conditional scale-location models in (62) which only require modeling the conditional mean and volatility functions, at least under the assumption of Gaussian shocks. Within this class of models, GARCH and Markov switching specifications such as (48) and (50) have been fairly dominant, although stochastic volatility models such as (60) are also popular.

The assumption of (conditionally) Gaussian innovations can of course easily be relaxed. For example, one can use a skewed t-distribution (Hansen (1994)) to capture skews and fat tails. Semi-nonparametric density methods have also been proposed, see, e.g., Gallant and Tauchen (1989). These can use hermite polynomials to add flexibility to a conditional scale-location model and take the form

$$p_{t+1|t}(r_{t+1}) = \frac{1}{c\sigma_{t+1|t}} \left(\sum_{i=0}^{m} \omega_i \eta_{t+1|t}^i \right)^2 \phi(\eta_{t+1|t}),$$

where $\phi(.)$ is the standard Gaussian density function, $\eta_{t+1|t} = (r_{t+1} - \mu_{t+1|t})/\sigma_{t+1|t}$ is the standardized residual, and $c = \int \left(\sum_{i=0}^{m} \omega_i \eta_{t+1|t}^i \right)^2 \phi(\eta_{t+1|t}) d\eta_{t+1|t}$ is a normalization factor which ensures that the density integrates to one. When

[22] Copulas have some distinct advantages such as being invariant under increasing and continuous transformations of the marginal distributions.

$m = 0$, the standard Gaussian density is recovered as a special case, while higher values of m adds flexibility to the density specification.

5.4.1 Bayesian Approaches

Bayesian approaches used typically require fully specifying the underlying forecasting model and so will generate a predictive density as part of the analysis. For example, this is the case for the stochastic volatility model in equation (61).

Density forecasting is another area in which forecast combination has found good use. Geweke and Amisano (2011) propose to construct optimal pools of density forecasts. Let $p_{it|t-1}$ be the conditional density forecast of y_t from the ith model, given information at time $t-1$. Given a sample of data, $\{y_t\}$, $t = 1, ..., T$, Geweke and Amisano propose to choose weights $\omega = (\omega_1, ..., \omega_n)'$ so as to maximize the sample log predictive score function

$$\omega = \arg\max_{\omega} T^{-1} \sum_{t=1}^{T} \ln \left(\sum_{i=1}^{n} \omega_i p_{it|t-1}(y_t) \right). \tag{68}$$

Geweke and Amisano consider combinations of a variety of GARCH and hidden Markov normal mixture models (corresponding to different specifications, $p_{it|t-1}$) and find that the optimal pool often yields interior solutions to (68) putting considerable weight on at least two different density models, as opposed to assigning all or almost all weight to a single model.

Bayesian model averaging (BMA) is another useful technique that can be used to combine density forecasts. The BMA density forecast can be computed as a weighted average of n individual density forecasts

$$p^{BMA}(y_{t+1}|Z_t) = \sum_{i=1}^{n} p(y_{t+1}|M_i, Z_t) p(M_i|Z_t), \tag{69}$$

where $p(M_i|Z_t)$ is the posterior probability for model i given the data, Z_t. Using Bayes rule, this, in turn, can be computed from

$$p(M_i|Z_t) = \frac{p(Z_t|M_i)p(M_i)}{\sum_{j=1}^{n} P(Z_t|M_j)p(M_j)}, \tag{70}$$

where $p(M_i)$ is the prior probability of (density) model M_i, while $p(Z_t|M_i)$ is known as the marginal likelihood of model i. This is given by

$$p(Z_t|M_i) = \int p(Z_t|\theta_i, M_i) p(\theta_i|M_i) d\theta_i, \tag{71}$$

where $p(\theta_i|M_i)$ is the prior density of model i's parameters and $p(Z_t|\theta_i, M_i)$ is the likelihood of the data, given the parameters under the ith model.

As can be seen from this description, the list of requirements for implementing BMA forecasts can be quite involved. First, one must have a set of models, $M_1, ..., M_n$. For each of these models, one must be able to compute $p(M_i|Z_t)$ in

(71) which can be time consuming for many types of models. Prior model probabilities, $p(M_1), ..., p(M_n)$ are also needed–these are often set to $1/n$–as are priors for the model parameters for each of the models, $p(\theta_1|M_1)$, ..., $p(\theta_n|M_n)$.[23]

5.4.2 Extracting densities from option prices

Market prices on derivatives such as options provide a special opportunity to obtain density estimates. Specifically, using a cross-section of options with identical expiration date, T, but different strikes, X, one can compute option-implied density estimates. Let r_f denote the riskfree rate, while S_T is the price of the underlying asset at time T. Then the price of a call option at time $t < T$ is given by

$$C_t(T, X) = e^{-r_f(T-t)} \int_0^\infty (S_T - X, 0) f_t(S_T) dS_T, \tag{72}$$

where $f_t(S_T)$ is the option-implied risk-neutral density for S_T and $F_t(S_T)$ is the corresponding CDF. An estimate of $f_t(S_T)$ can be obtained by differentiating the expression for $C_t(T, X)$:

$$f_t(S_T) = e^{r_f(T-t)} \frac{\partial^2 C_t(T, X)}{\partial X^2}\bigg|_{X=S_T}. \tag{73}$$

A variety of approximation methods exist for estimating $f_t(S_T)$, e.g., using a butterfly spread of options with neighboring strike prices:

$$f_t(X_n) \approx e^{r_f(T-t)} \left(\frac{C_t(T, X_{n+1}) - 2C_t(T, X_n) + C_t(T, X_{n-1})}{(\Delta X)^2} \right). \tag{74}$$

Complications arise when using data from options markets in this manner. First, the above formulas allow us to estimate the risk-neutral densities from option prices. Such densities are different from the density under the physical measure. Under assumptions about the process driving the underlying asset price and investor risk aversion (in incomplete markets), formulas can be derived that link the physical and risk-neutral probability distributions; see Christoffersen et al. (2013) for a thorough review and discussion of this topic. Second, options on many securities are often not very liquid, making it difficult to get a broad cross-section from which option-implied densities can be estimated from equations such as (73). Measurement errors due to market microstructure effects and illiquid markets can also introduce biases in implied volatility estimates–see Christensen and Prabhala (1998) and Poon and Granger (2003).

[23] Avramov (2002) uses BMA to a combination of 2^{14} possible model specifications obtained as the exhaustive set of combinations of 14 different predictors. Avramov computes the economic (utility) loss from ignoring model uncertainty and finds that this can be sizeable. He also finds that BMA forecasts are more robust than forecasts from individual models.

6 Evaluating financial forecasts

Evaluation of forecasts plays an important role in finance due to the presence of well-defined utility or loss functions that can be used to assess predictive accuracy. Another factor that plays a role is the relative ease with which financial forecasts can be exploited in investment strategies. Whereas it can be difficult to measure the impact on economic welfare of small improvements in the accuracy of macroeconomic forecasts, it is usually easier to compute the effect of increased predictive accuracy on trading profits. The conventional tool for doing so is what is usually referred to as "back tests", i.e., simulated real-time trading results based some a financial forecasting approach.[24]

6.1 Economic versus statistical loss

Ait-Sahalia and Brandt (2001) discuss how determining an investor's optimal portfolio weights can be viewed as a prediction problem given the investor's utility function, u, and a budget constraint relating future wealth, $W_{t+1} = \omega_t' R_{t+1}$, to portfolio weights, ω_t, and a vector of gross asset returns, R_{t+1}.[25]

Suppose the investor chooses portfolio weights, ω_t to maximize

$$\omega_t^* = \arg\max_{\omega_t} E_t \left[u \left(\omega_t' R_{t+1} \right) \right], \tag{75}$$

where $\omega_t' \iota = 1$ for $\iota = (1, 1, ..., 1)'$. Assuming mean-variance preferences with coefficient of absolute risk aversion $A \geq 0$

$$E_t[u(W_{t+1})] = E_t[W_{t+1}] - \frac{A}{2} Var_t(W_{t+1}) \equiv \mu_t - \frac{A}{2}\Sigma_t,$$

the optimal portfolio weights are a non-linear function of the first and second moments of the wealth distribution:[26]

$$\omega_t = \Sigma_t^{-1} \iota \frac{AW_t - \iota' \Sigma_t^{-1} \mu_t}{AW_t \iota' \Sigma_t^{-1} \iota} + \frac{\Sigma_t^{-1} \mu_t}{AW_t}, \tag{76}$$

see Ait-Sahalia and Brandt (2001). To evaluate (76), one option is to use plug-in estimates of μ_t and Σ_t. This is not a very desirable strategy, however, as it ignores estimation error. A more desirable strategy is to parameterize the investor's portfolio choice $\omega(x_t, \beta)$ as a function of a set of predictors known at time t, x_t and parameters, β.

To keep the solution feasible and allow for multiple predictor variables, Ait-Sahalia and Brandt propose using a single-index specification $\omega(x_t, \beta) = \omega(x_t'\beta)$. Deriving the first-order condition from the optimization problem in (75), we have

$$E \left[u' \left(W_t(\omega(x_t'\beta)' R_{t+1}) R_{t+1} \right) | x_t \right] = 0. \tag{77}$$

[24] There is also a literature on evaluation of volatility forecasts which we do not focus on here; see, e.g., Patton (2011).

[25] For simplicity, we assume initial wealth of $W_t = 1$.

[26] This expression assumes there is no risk-free asset available.

This equation can be used to estimate the parameters of the portfolio policy function by means of the GMM estimation method, using functions of the data, $g(x_t)$ to convert (77) into sample moment restrictions.

In situations where ω_t is of sufficiently low dimension to allow for a grid search, simple numerical methods can be used to determine the investor's optimal portfolio weights given a predictive distribution of returns, $p\left(R_{t+1}|\,\Omega_t\right)$. To see how this works, consider the optimal asset allocation of an investor with power utility function who can choose between a risk-free asset and a risky stock market portfolio with gross returns of R_{ft+1} and R_{t+1}, respectively:

$$U\left(\omega_t, R_{t+1}\right) = \frac{\left[\left(1-\omega_t\right)R_{ft+1} + \omega_t R_{t+1}\right]^{1-A}}{1-A}, \tag{78}$$

where A is now the investor's coefficient of relative risk aversion. The investor chooses ω_t to solve the optimal asset allocation problem

$$\omega_t^* = \arg\max_{\omega_t} \int U\left(\omega_t, R_{t+1}\right) p\left(R_{t+1}|\,\Omega_t\right) dr_{t+1}. \tag{79}$$

The integral in (79) can be numerically approximated using J draws from the predictive distribution

$$\widehat{\omega}_{t,i} = \arg\max_{\omega_t} \frac{1}{J}\sum_{j=1}^{J}\left\{\frac{\left[\left(1-\omega_t\right)R_{ft+1} + \omega_t R_{t+1}\right]^{1-A}}{1-A}\right\}. \tag{80}$$

For example, in a frequentist setting a GARCH model could be used as the predictive distribution from which to make draws. This procedure would typically condition on the parameter estimates. In a Bayesian setting, one could draw from the posterior predictive distribution such as (61) to account for parameter uncertainty.

6.2 Economic versus statistical loss

Using economic rather than purely statistics loss functions in the forecast evaluation step can make a substantial difference. Kandel and Stambaugh (1996) provide a range of illustrative examples showing situations where seemingly weak return predictability can have important economic consequences, e.g., in portfolio allocation problems.

Empirically, Cenesizoglu and Timmermann (2012) consider the performance of a range of forecasting models with time-varying mean and volatility. The parameters of all models are estimated recursively and forecasting performance is measured out-of-sample. For the economic loss function, the forecasts are used to select optimal portfolio weights either under mean-variance preferences or under power utility and performance is evaluated using these utility functions.

Cenesizoglu and Timmermann (2012) find that it is common for return prediction models to produce higher out-of-sample mean squared forecast errors than a model with a constant equity premium, yet simultaneously add economic

value when their forecasts are used to guide portfolio decisions. For sure, there is generally a positive correlation between a return prediction model's out-of-sample MSE performance and its ability to add economic value. However, this relation tends to be weak and only explains a small part of the cross-sectional variation in different models' ability to generate economic value in the out-of-sample analysis.

These results suggest that underperformance along conventional measures of forecasting performance such as root mean squared forecast errors contain only limited information on whether return prediction models that allow for a time-varying mean or variance will help or hurt investors when these return forecasts are used to guide portfolio decisions.

7 Conclusion

As the cost of fitting an ever-larger number of increasingly sophisticated and flexible forecasting models to financial data sets declines, the ability of researchers to separate spurious patterns from genuine predictability becomes harder. Recent advances in dealing with the multiple hypothesis testing problem underlying the data mining problem offer some promise, although they often require revisiting the entire list of predictors that have been considered in the literature–as opposed to considering individual predictors in isolation–which can be a costly exercise for individual researchers.[27]

We are likely to see in future research a close contest between increasingly powerful forecasting methods with the ability to find predictive patterns even in settings with low signal-to-noise ratios and more sophisticated methods for dealing with overfitting and data mining. Ideally, economic theory will play an important role in this contest, helping researchers to identify robust predictability patterns, but also bearing in mind that sometimes predictive modeling is ahead of theory and it may not always be clear how to pinpoint from theory which variables should possess predictive power over the outcome of interest.

Data limitations reduce the accuracy of financial forecasts in many situations. For example, our ability to predict variation in asset returns at the business cycle frequency using, e.g., slow-moving predictors such as interest rate spreads or valuation ratios, is ultimately limited by the small number of business cycles observed in historical data. Similarly, the tendency of predictable patterns in returns to self-destruct as a result of competitive pressures means that there are limits on how informative historical data will be for predicting future returns.

These points make many financial forecasting problems both challenging and fascinating. However, because the potential payoffs from uncovering even small degrees of predictability are so high, undoubtedly finance will remain an innovative and fast-paced arena for developing and testing new forecasting methods.

[27] See Sullivan et al. (1999) for a study of technical trading rules and Harvey et al. (2015) for a study of factors for explaining cross-sectional variation in expected returns.

References

[1] Aït-Sahalia, Y., and M. W. Brandt, 2001, Variable selection for portfolio choice. Journal of Finance 56:1297–351.

[2] Aït-Sahalia, Y., P.A. Mykland, and L. Zhang, 2005, How often to sample a continuous-time process in the presence of market microstructure noise. Review of financial studies 18.2: 351-416

[3] Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold. 2006. Volatility and correlation forecasting. Handbook of economic forecasting 1:777–878.

[4] Andersen, T.G., T. Bollerslev, and F.X. Diebold, 2007, Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. The review of economics and statistics 89, 4, 701-720.

[5] Andreou, E., E. Ghysels, and A. Kourtellos, 2011, Forecasting with mixed-frequency data. In M. Clements and D. Hendry (eds.), Oxford Handbook of Economic Forecasting, 225–45. Oxford University Press.

[6] Ang, A. and M. Piazzesi, 2003, A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. Journal of Monetary Economics 50(4), 745 – 787.

[7] Avramov, D., 2002, Stock return predictability and model uncertainty. Journal of Financial Economics 64:423–58.

[8] Blair, B.J., S-H Poon, and S.J. Taylor, 2001, Forecasting S&P 100 volatility: The incremental information content of implied volatilities and high-frequency index returns. Journal of Econometrics. 105, 1, 5-26.

[9] van Binsbergen, J.H., and R. Koijen, 2010, Predictive regressions: a present-value approach. Journal of Finance 65, 4, 1439-1471.

[10] Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–27.

[11] Campbell, J.Y., and R.J. Shiller, 1988, The dividend-price ratio and expectations of future dividends and discount factors. Review of financial studies 1, 3, 195–228.

[12] Campbell, J.Y., and S.B. Thompson, 2008, Predicting excess stock returns out of sample: can anything beat the historical average? Review of Financial Studies 21, 4, 1509–1531.

[13] Cenesizoglu, T. and A. Timmermann, 2012, Do return prediction models add economic value? Journal of Banking & Finance 36, 11, 2974- 2987.

[14] Chib, S., 1998. Estimation and comparison of multiple change point models. Journal of Econometrics 86, 221–241.

[15] Christensen, B.J. and N.R. Prabhala, 1998, The Relation between Implied and Realized Volatility. Journal of Financial Economics 50, 2, 125–50.

[16] Christoffersen, P., K. Jacobs, and B. Chang, 2013, Forecasting with Option Implied Information. Handbook of Economic Forecasting, edited by G. Elliott and A. Timmermann, Volume 2, Chapter 10, 581-656. Elsevier.

[17] Christopherson, J., Ferson, W., Glassman, D., 1998. Conditioning manager alphas on economic information: Another look at the persistence of performance. Review of Financial Studies 11, 111-142.

[18] Clemen, R.T., 1989, Combining forecasts: A review and annotated bibliography. International journal of forecasting 5, 4, 559–583.

[19] Cochrane, J.H., 2009, Asset Pricing (Revised Edition). Princeton university press.

[20] Corsi, F. 2009, A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7, 2, 174-196.

[21] Dangl, T., and M. Halling. 2012. Predictive regressions with time-varying coefficients. Journal of Financial Economics 106, 157–81.

[22] DeMiguel, V., L. Garlappi, and R. Uppal, 2007, Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?. Review of Financial studies 22,5: 1915-1953.

[23] Diebold, F. X., and R. S. Mariano. 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13:253–63.

[24] Elliott, G., A. Gargano, and A. Timmermann. 2013. Complete subset regressions. Journal of Econometrics 177:357–73.

[25] Elliott, G., A. Gargano, and A. Timmermann, 2015, Complete subset regressions with large-dimensional sets of predictors. Journal of Economic Dynamics and Control 54:86–110.

[26] Engle, R.F., 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica, 987–1007.

[27] Engle, R.F., 2002, Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics 20, 339–50.

[28] Engle, R. F., and K. Sheppard, 2001, Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. working papers 8554, National Bureau of Economic Research.

[29] Farmer, L., L. Schmidt, and A. Timmermann, 2017, Pockets of predictability. Unpublished manuscript, University of Virginia, Chicago, and UCSD.

[30] Ferreira, M.A., and P. Santa-Clara, 2011, Forecasting stock market returns: The sum of the parts is more than the whole. Journal of Financial Economics 100, 3, 514-537.

[31] Ferson, W.E., and Schadt, R.W., 1996, Measuring fund strategy and performance in changing economic conditions. Journal of Finance 51, 425-461.

[32] Ferson, W.E., S. Sarkissian, and T.T., Simin, 2003, Spurious regressions in financial economics?. Journal of Finance 58.4, 1393-1413.

[33] Gallant, A.R., and G. Tauchen, Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications. Econometrica 57, 1091-1120.

[34] Geweke, J., and G. Amisano. 2011. Optimal prediction pools. Journal of Econometrics 164:130–41.

[35] Ghysels, E., A. Sinko, and R. Valkanov, 2007, MIDAS regressions: Further results and new directions. Econometric Reviews 26:53–90.

[36] Giacomini, R., and B. Rossi. 2009. Detecting and predicting forecast breakdowns. Review of Economic Studies 76:669–705.

[37] Giacomini, R., and B. Rossi. 2010. Forecast comparisons in unstable environments. Journal of Applied Econometrics 25:595–620.

[38] Glosten, L.R., R. Jagannathan, and D.E. Runkle. 1993, On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of finance, 48, 5, 1779-1801.

[39] Hamilton, J.D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 357–84.

[40] Hansen, B.E., 1994, Autoregressive Conditional Density Estimation. International Economic Review 35, 705-730.

[41] Hansen, P. R., Z. Huang, and H. Shek, 2012, Realized GARCH: A joint model for returns and realized measures of volatility. Journal of Applied Econometrics 27, 877-906.

[42] Hansen, P. R., and A. Lunde. 2005. A forecast comparison of volatility models: does anything beat a GARCH (1, 1)? Journal of applied econometrics 20:873–89.

[43] Hansen, P.R, and A. Lunde, 2011, Forecasting volatility using high frequency data. The Oxford Handbook of Economic Forecasting, Oxford: Blackwell, 525-556.

[44] Hansen, P. R., A. Lunde, and J. M. Nason. 2011. The model confidence set. Econometrica 79:453–97.

[45] Hansen, P. R., and Timmermann, A., 2012, Choice of sample split in out of-sample forecast evaluation. Unpublished manuscript, UNC Chapel Hill and UCSD.

[46] Hansen, P.R., and A. Timmermann, 2015a, Equivalence Between Out-of-Sample Forecast Comparisons and Wald Statistics. Econometrica 83 (6), 2485-2505.

[47] Hansen, P.R. and A. Timmermann, 2015b, Comment. Journal of Business and Economic Statistics 33:1, 17-21.

[48] Harvey, Campbell R., Y. Liu, and H. Zhu, 2016, ... and the cross-section of expected returns. Review of Financial Studies 29, 1, 5-68.

[49] Hastie, T., R. Tibshirani, J. Friedman, 2009, The elements of statistical learning, vol. 2. Springer.

[50] Henkel, S.J., J.S. Martin, and F. Nardari, 2011, Time-varying short-horizon predictability. Journal of Financial Economics 99, 3, 560-580.

[51] Inoue, A., and L. Kilian, 2005, In-sample or out-of-sample tests of predictability: Which one should we use? Econometric Reviews 23:371–402.

[52] Johannes, M., A. Korteweg, and N. Polson, 2014, Sequential learning, predictability, and optimal portfolio returns. Journal of Finance 69, 2, 611-644.

[53] Kandel, S., and R. F. Stambaugh. 1996. On the predictability of stock returns: An asset-allocation perspective. Journal of Finance 51:385–424.

[54] Kelly, B., and S. Pruitt, 2013, Market expectations in the cross-section of present values. The Journal of Finance 68, 5, 1721-1756.

[55] Kendall, M.G., 1954. Note on bias in the estimation of autocorrelation. Biometrika 41, 403-404.

[56] Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: Likelihood inference and comparison with arch models. The Review of Economic Studies 65(3), 361–393.

[57] Ludvigson, S.C., and S. Ng., 2007, The empirical risk–return relation: A factor analysis approach. Journal of Financial Economics 83, 1, 171-222.

[58] McLean, R.D., and J. Pontiff, 2016, Does Academic Research Destroy Stock Return Predictability? Journal of Finance 71, 1, 5-32.

[59] Marriott, F.H.C., Pope, J.A., 1954. Bias in the estimation of autocorrelations. Biometrika 41, 390-402.

[60] Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G., 2014, Forecasting the equity risk premium: the role of technical indicators. Management Science 60, 7, 1772-1791.

[61] Pastor, L., and R.F. Stambaugh, 2001, The equity premium and structural breaks. Journal of Finance 56, 1207-1245.

[62] Pastor, L., and R.F. Stambaugh, 2009, Predictive systems: living with imperfect predictors. Journal of Finance 64, 4,1583–1628.

[63] Patton, A.J., 2011, Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics 160, 1, 246-256.

[64] Patton, A.J. 2013, Copula methods for forecasting multivariate time series. Handbook of economic forecasting vol. 2, 899-960.

[65] Patton, A.J., and K. Sheppard, 2015, Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics 97, 3, 683-697.

[66] Paye, B.S., 2012, Déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. Journal of Financial Economics 106, 3, 527-546.

[67] Paye, B.S., and A. Timmermann. 2006. Instability of return prediction models. Journal of Empirical Finance 13:274–315.

[68] Pesaran, M.H., and A. Timmermann, 2007, Selection of estimation window in the presence of breaks. Journal of Econometrics 137:134–61.

[69] Pettenuzzo, D., and A. Timmermann, 2011, Predictability of stock returns and asset allocation under structural breaks. Journal of Econometrics 164:60–78.

[70] Pettenuzzo, D., A. Timmermann, and R. Valkanov, 2014, Forecasting Stock Returns under Economic Constraints. Journal of Financial Economics 114, 517-553.

[71] Pettenuzzo, D., A. Timmermann, and R. Valkanov, 2016, A MIDAS Approach to modeling first and second moment dynamics. Journal of Econometrics 193, 2, 315-334.

[72] Poon, S-H., and C.W.J. Granger, 2003, Forecasting volatility in financial markets: A review. Journal of Economic Literature XLI, 478-539.

[73] Rapach, D.E., and M.E. Wohar, 2006, Structural breaks and predictive regression models of aggregate us stock returns. Journal of Financial Econometrics, 4, 2, 238–274.

[74] Rapach, D.E., J.K. Strauss, and G. Zhou, 2010, Out-of-sample equity premium prediction: combination forecasts and links to the real economy. Review of Financial Studies 23, 2, 821–862.

[75] Romano, J., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73,4, 1237-1282.

[76] Rossi, B. 2013a, Advances in Forecasting under Instability. Chapter 21, pages 1203-1324 in G. Elliott and A. Timmermann (eds.) Handbook of Economic Forecasting vol 2B. North-Holland.

[77] Rossi, B. 2013b, Exchange rate predictability. Journal of Economic Literature 51:1063–119.

[78] Sarno, L., P. Schneider, and C. Wagner (2016). The economic value of predicting bond risk premia. Journal of Empirical Finance 37, 247–267.

[79] Sklar, M., 1959, Fonctions de repartition 'a n dimensions et leurs marges. Universite Paris 8.

[80] Smith, S., and A. Timmermann, 2017, Detecting Breaks in Real Time: A Panel Forecasting Approach. Unpublished Manuscript

[81] Stambaugh, R. F. 1999. Predictive regressions. Journal of Financial Economics 54:375–421.

[82] Sullivan, R., A. Timmermann, and H. White, 1999, Data-snooping, technical trading rule performance, and the bootstrap. Journal of Finance 54, 5, 1647-1691.

[83] Timmermann. A., 2006, Forecast combinations. Handbook of economic forecasting vol. 1, 135–196.

[84] Welch, I., and A. Goyal, 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21,4, 1455-1508.

[85] White, H. 2000. A reality check for data snooping. Econometrica 68:1097–126.

[86] Zhou, G. , 2010, How much stock return predictability can we expect from an asset pricing model? Economics Letters 108, 184-186.