14

Bagging Constrained Equity Premium Predictors

Eric Hillebrand, Tae-Hwy Lee, and Marcelo C. Medeiros[1]

14.1 Introduction

We study bootstrap aggregation (bagging) methods to improve the out-of-sample prediction of a simple univariate linear model (following Campbell and Thompson (CT), 2008) for the equity premium by imposing restrictions in the regression. In imposing restrictions, such as positivity of the regression coefficient or positivity of the prediction, usually the coefficient or prediction is simply set to zero if the estimated value is negative. This amounts to the application of an indicator function. Bühlmann and Yu (2002) showed that bagging can reduce the variance of the estimator in this situation by "smoothing" the indicator function. In this chapter we show in theory, through simulations, and in an empirical application using the same data set as CT (2008), that bagging coefficient and forecast restrictions can improve the predictive power of a linear model.

Excess returns prediction has attracted academics and practitioners for many decades since the early 1920s, when Dow (1920) studied the role of dividend ratios as a possible predictor for returns. In the 1980s, a number of authors presented empirical evidence of ex-post (in-sample) return predictability. Fama and Schwert (1977), Fama and Schwert (1981), Rozeff (1984), Keim and Stambaugh (1986), Campbell (1987), Campbell and Shiller (1988a,b), and Fama and French (1988, 1989) showed that excess returns could be successfully predicted based on lagged values of variables such as dividend-price ratio

[1] We would like to thank the editors Niels Haldrup, Mika Meitz, and Pentti Saikkonen, two anonymous referees, and Mike McCracken for their very thorough comments. All remaining errors are ours. An older version of this chapter was circulated under the title "Let's Do It Again: Bagging Equity Premium Predictors".

and dividend yield, earnings-price ratio and dividend-earnings ratio, interest rates and spreads, inflation rates, book-to-market ratio, volatility, investment-capital ratio, consumption, wealth, and income ratio, and aggregate or net equity issuing activity.

Subsequent work, however, demonstrated that these results do not hold during the bull market period of the 1990s; see Lettau and Ludvigson (2001) or Schwert (2002). For example, during this period when stock prices soared, the dividend yield systematically drifted downwards, thus generating negative sample correlation between returns and dividend yield, contrary to the positive historical correlation. Furthermore, since early results concerned only ex-post predictability, they were of little practical interest. Studies of ex-ante (out-of-sample) return predictability have found either that previous successful results were restricted to particular sub-samples (Pesaran and Timmermann 1995) or that return predictability was a statistical illusion; see Bossaerts and Hillion (1999). In addition, several authors pointed out that the apparent predictability of stock returns might be spurious as many of the predictor variables were highly persistent, leading to possibly biased coefficients and incorrect t-tests in predictive regressions; see, for example, Nelson and Kim (1993), Cavanagh, Elliot, and Stock (1995), and Stambaugh (1999). These problems are exacerbated when large numbers of variables are considered and only results that are apparently statistically significant are reported; see Foster, Smith, and Whaley (1997) and Ferson, Sarkissian, and Simin (2003).

The inconclusive evidence has inspired the use of time-varying regression models. As pointed out by Pesaran and Timmermann (2002) and Timmermann (2008) "forecasters of stock returns face a moving target that is constantly changing over time. Just when a forecaster may think that he has figured out how to predict returns, the dynamics of market prices will, in all likelihood, have moved on, possibly as a consequence of the forecaster's own efforts." On the other hand, alternative econometric methods were advocated for correcting the abovementioned bias and conducting valid inference: for example Cavanagh, Elliot, and Stock (1995), Mark (1995), Kilian (1999), Ang and Bekaert (2007), Jansson and Moreira (2006), Lewellen (2004), Torous, Valkanov, and Yan (2004), Campbell and Yogo (2006), and Polk, Thompson, and Vuolteenaho (2006).

More recently, Goyal and Welch (2008) argued that none of the conventional predictor variables proposed in the literature seemed capable of systematically predicting stock returns out-of-sample. Their empirical evidence suggested that most models were unstable or spurious, and most models were no longer significant even in-sample. The authors show that the earlier apparent statistical significance was especially confined to the years of the Oil Shock of 1973–1975; see also Butler, Grullon, and Weston (2006).

Our approach is motivated by CT (2008), who show that many predictive regressions outperform the historical average return forecast once a restriction is imposed on the sign of the coefficient in the regression. Imposing a constraint on the coefficient amounts to applying shrinkage estimation. Shrinkage methods are designed to reduce estimator variance at the possible cost of incurring bias. CT (2008) find out-of-sample predictive power of the common stock return predictors over the historical average. The advantage is small (not statistically significant) but nonetheless economically meaningful for mean-variance investors. We impose these a priori parameter restrictions (which we call CT restrictions) in a regression function of equity premium conditional on various predictors, then we smooth the restrictions by bagging (Bühlmann and Yu 2002; Inoue and Kilian 2008). The resulting bagging forecast has lower variance than the forecast using the CT restricted estimator. Whether mean squared forecast error (MSFE) is also reduced in the process depends on how much bagging increases bias. We explore this question in simulations and in an application to the same data set used in CT (2008).

After bagging an indicator-type restriction as it is imposed in CT (2008), the resulting asymptotic shape of the estimator follows the cumulative distribution function of a normal random variable (Bühlmann and Yu 2002, proposition 2.1). That is, instead of taking one value (say, zero) on one side of the restriction, and a positive value on the other side of the restriction undergoing an abrupt transition, the estimator now transitions smoothly from one side of the restriction to the other. See, in particular, figure 1 in Bühlmann and Yu (2002). Thereby, bagging provides another perspective on nonlinear, smoothly transitioning estimators as pioneered in, for example, Teräsvirta (1994, 2006).

The chapter is organized as follows. In Section 14.2 we review bagging and present the bagging approach to restricted parameter estimation. In Section 14.3, we present a Monte Carlo simulation. Section 14.4 describes the data set and presents empirical results. Section 14.5 concludes.

14.2 Bagging restrictions on regression functions

A linear model assumes that the regression function $\mathbb{E}(y|\mathbf{X})$ is linear in the predictors $\mathbf{X} = (x_1, \ldots, x_k)'$. When k is small, such as $k = 1$, we may have good reason to believe that the coefficient of X must be positive or must exceed some known value. In that case, we may use the a priori belief to *shrink* the parameter space. Such a priori beliefs are less intuitive when k is large, so we restrict the consideration to the case $k = 1$. A simple method is to use a hard-thresholding indicator function to define a constrained least squared estimator $\bar{\beta} = \max(\tilde{\beta}, 0) = \tilde{\beta} \cdot \mathbf{1}(\tilde{\beta} > 0)$ with $\tilde{\beta}$ being an unconstrained least squares estimator of β. By shrinking the parameter space for the slope coefficient of x_1 in the prediction model towards zero, one reduces the variance, and possibly the overall mean squared forecast error, at the cost of bias. While imposing such a constraint can improve the predictive power of the linear model if the constraint is correct, the restricted estimator $\bar{\beta}$ involves a discontinuous hard-threshold indicator (jump) function at the boundary of the constrained parameter space. We show that we can further improve the predictive ability of the constrained linear model by smoothing the indicator function using bagging. We consider two types of restrictions as considered in CT (2008), namely positivity of the forecast, motivated by the requirement that the mean of the equity premium should be positive, and positivity of the regression coefficient, motivated by simple insights from financial theory—for example that the dividend yield should have a positive influence on the equity premium.

Goyal and Welch (GW 2008) show that predictive regressions cannot beat the historical average. Campbell and Thompson (CT 2008), on the other hand, show that many predictive regressions beat the historical average once constraints are imposed. Consider

$$y_{t+1} = \alpha + \beta x_t + u_{t+1}, \quad t = 1, \ldots, T, \tag{14.1}$$

where y is the excess return on the S&P500 over the 3-month T-bill interest rate. The regressor x_t stands for a predictor variable such as dividend yield, earnings yield, book-to-market ratio, return on equity, long-term government bond yield, term spread, default spread, inflation rate, equity share of new issues, and so on. GW find that forecasts from the unrestricted model $\tilde{y}_{n+1} = \tilde{\alpha}_n + \tilde{\beta}_n x_n$ (with $\tilde{\alpha}_n, \tilde{\beta}_n$ unrestricted OLS) are worse than forecasts with the exclusion restriction $\beta = 0$, which amounts to the historical average (HA) $\frac{1}{n}\sum_{t=1}^{n} y_t$. CT show that the positivity constraint $\beta > 0$ produces a better forecast than the exclusion restriction. In this chapter we aim to show that bagging can further improve the CT constrained forecast.

14.2.1 *Bagging*

Bootstrap aggregating, or bagging, means to estimate a parameter on each of a set of J sub-samples drawn from the original data set $\mathcal{D}$ and then average over the J estimates. As $J \to \infty$, the bagged estimator will differ from the estimator obtained from the entire data set $\mathcal{D}$ only if it is a nonlinear or adaptive estimator (Hastie, Tibshirani, and Friedman 2001, p. 246). An estimator is said to be "unstable" if a small change in the training set will lead to a significant change in the estimator (Breiman 1996). In our application to an indicator function, the bagged predictor smoothes the instability caused by estimation and model uncertainty and the hard threshold function.

The mechanism of bagging has been explained in various ways, for example Breiman (1996) under squared-error loss and Lee and Yang (2006) under convex loss (e.g., a tick function for quantiles). Bühlmann and Yu (2002) show that for a nonsmooth unstable predictor, bagging reduces variance of the first-order term. In particular, they show that bagging can reduce the mean-squared forecast error by averaging over the randomness of variable selection. Buja and Stuetzle (2006) and Friedman and Hall (2007) expand a smooth unstable function into linear and higher order terms, and show that bagging reduces the variance of the higher order terms. Grandvalet (2004) argues that bagging stabilizes prediction by equalizing the influence of training samples. Stock and Watson (2012) show that bagging is asymptotically equivalent to Bayesian shrinkage. Applications of bagging include inflation (Inoue and Kilian 2008), financial volatility (Hillebrand and Medeiros 2010), equity premium (Huang and Lee 2010), short-term interest rates (Audrino and Medeiros 2011), and employment data (Rapach and Strauss 2010).

To fix notation, let

$$\mathcal{D}_t = \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t} \qquad (t = R, \ldots, T)$$

be a training set at time t and let $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ be a forecast of Y_{t+1} or of the binary variable $\mathbf{1}(Y_{t+1} \geq 0)$ using this training set $\mathcal{D}_t$ and the explanatory variable vector $\mathbf{X}_t$. The optimal forecast $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ for Y_{t+1} will be the conditional mean of Y_{t+1} given $\mathbf{X}_t$ if we have the squared error loss function, or the conditional quantile of Y_{t+1} on $\mathbf{X}_t$ if the loss is a tick function as in Koenker and Basset (1978).

Suppose each training set $\mathcal{D}_t$ consists of R observations generated from the underlying probability distribution $\mathbf{P}$. The forecast $\{\varphi(\mathbf{X}_t, \mathcal{D}_t)\}_{t=R+1}^{T}$ can be improved if more training sets can be generated from $\mathbf{P}$ and if the forecast can be formed from averaging the multiple forecasts obtained from the multiple training sets. Ideally, if $\mathbf{P}$ were known and multiple training sets $\mathcal{D}_t^{(j)}$ $(j = 1, \ldots, J)$ could be drawn from $\mathbf{P}$, an ensemble aggregating predictor $\varphi_A(\mathbf{X}_t)$ could be constructed by averaging of $\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)})$ with respect to $\mathbf{P}$, that is,

$$\varphi_A(\mathbf{X}_t) = \mathbb{E}_{\mathbf{P}}\varphi(\mathbf{X}_t, \mathcal{D}_t),$$

where $\mathbb{E}_{\mathbf{P}}(\cdot)$ denotes expectation with respect to $\mathbf{P}$, and the subscript A in φ_A denotes "aggregation."

In practice, $\mathbf{P}$ is not known. We may estimate $\mathbf{P}$ by its empirical distribution, $\hat{\mathbf{P}}(\mathcal{D}_t)$, for a given data set $\mathcal{D}_t$. Then, from the empirical distribution $\hat{\mathbf{P}}(\mathcal{D}_t)$, multiple sub-samples $\mathcal{D}_t^*$ can be drawn by an appropriate bootstrap method. The question of which bootstrap algorithms can provide consistent densities for moment estimators and quantile estimators in time series settings is addressed, for example, in Hall, Horowitz, and Jing (1995) and

Fitzenberger (1997). Bagging predictors, $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$, can then be computed by averaging over the sub-samples. More specifically, the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be obtained following the steps: (1) Given a training set of data at time t, $\mathcal{D}_t = \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^t$, construct the jth bootstrap sample $\mathcal{D}_t^{*(j)} = \{(Y_s^{*(j)}, \mathbf{X}_{s-1}^{*(j)})\}_{s=t-R+1}^t$, $j = 1, \ldots, J$, according to the empirical distribution of $\hat{\mathbf{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$. (2) Compute the bootstrap predictor $\varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)})$ from the jth bootstrapped sample $\mathcal{D}_t^{*(j)}$. (3) Compute the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ by averaging over J bootstrap predictors

$$\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*) = \frac{1}{J}\sum_{j=1}^{J} \varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}).$$

14.2.2 *Bagging Restrictions*

In the context of a simple univariate regression, we let

$$\varphi(x_t, \mathcal{D}_t) = \mathbb{E}(y_{t+1}|x_t) = \alpha + \beta x_t,$$

where $\mathcal{D}_t = \{(y_{s+1}, x_s)\}_{s=t-R+1}^t$ for $t = R, \ldots, T$.

The two types of restrictions considered in CT (2008) are positivity of the coefficient β (PC) and positivity of the forecast $\varphi(x_t, \mathcal{D}_t)$ (PF). We compare the following forecasts:

1. **HA** (Historical Average forecast with exclusion restriction $\beta = 0$):

$$y_{T+1}^{\mathrm{HA}} = \frac{1}{T}\sum_{t=T-R+1}^{T} y_t.$$

2. **UF** (Unrestricted Forecast):

$$y_{T+1}^{\mathrm{UF}} = \tilde{\alpha}_T + \tilde{\beta}_T x_T,$$

   where $\tilde{\alpha}_T, \tilde{\beta}_T$ are unrestricted OLS estimators. UF is used in Goyal and Welch (2008).

3. **PC** (forecast with Positive Coefficient restriction $\beta > 0$):

$$y_{T+1}^{\mathrm{PC}} = \bar{\alpha}_T + \bar{\beta}_T x_T, \qquad (14.2)$$

   where $\bar{\beta}_T = \max\{\tilde{\beta}_T, 0\} = \mathbf{1}\left(\tilde{\beta}_T > 0\right)\tilde{\beta}_T$, and $\bar{\alpha}_T = \mathbf{1}\left(\tilde{\beta}_T > 0\right)\tilde{\alpha}_T + \mathbf{1}\left(\tilde{\beta}_T \leq 0\right) y_{T+1}^{\mathrm{HA}}$. PC is used in CT (2008).

4. **PC-GH** (forecast with Positive Coefficient restriction $\beta > 0$ using bagging as in Gordon and Hall (GH), 2009):

$$y^{\text{PC-GH}}_{T+1,J} = \frac{1}{J}\sum_{j=1}^{J}\left(y^{\text{PC},*(j)}_{T+1}\right) = \hat{\alpha}_{T,J} + \hat{\beta}_{T,J}x_T, \tag{14.3}$$

where $\hat{\beta}_{T,J} = \frac{1}{J}\sum_{j=1}^{J}\bar{\beta}^{*(j)}_T$ and $\hat{\alpha}_{T,J} = \frac{1}{J}\sum_{j=1}^{J}\bar{\alpha}^{*(j)}_T$. The $\bar{\alpha}^{*(j)}_T$ and $\bar{\beta}^{*(j)}_T$ are the positivity-constrained estimators obtained from the jth sub-sample. As the number of sub-samples $J \to \infty$, $\hat{\beta}_{T,J}$ converges to $\mathbb{E}^*\bar{\beta}_T$, $\hat{\alpha}_{T,J}$ converges to $\mathbb{E}^*\bar{\alpha}_T$, and $y^{\text{PC-GH}}_{T+1,J}$ converges to $\mathbb{E}^*(y^{\text{PC}}_{T+1})$, where $\mathbb{E}^*$ denotes expectation with respect to the empirical distribution $\hat{\mathbf{P}}(\mathcal{D}_T)$.

5. **PF** (forecast with Positive Forecast restriction $\varphi(x_t,\mathcal{D}_t) > 0$):

$$y^{\text{PF}}_{T+1} = \mathbf{1}\left(y^{\text{UF}}_{T+1} > 0\right)y^{\text{UF}}_{T+1}.$$

6. **PF-GH** (forecast with Positive Forecast restriction $\varphi(x_t,\mathcal{D}_t) > 0$ using bagging as in GH 2009):

$$y^{\text{PF-GH}}_{T+1,J} = \frac{1}{J}\sum_{j=1}^{J}\left(y^{\text{PF},*(j)}_{T+1}\right),$$

where again there is convergence $\lim_{J\to\infty} y^{\text{PF-GH}}_{T+1,J} = \mathbb{E}^*(y^{\text{PF}}_{T+1})$, $\mathbb{E}^*(\cdot)$ being expectation with respect to the empirical distribution $\hat{\mathbf{P}}(\mathcal{D}_T)$.

With this notation, we can rephrase the aim of this chapter. Comparative statements can be understood in the mean-square error sense, but we also consider two alternative loss functions in the empirical exercise: GW (2008) find that unrestricted forecasts $y^{\text{UF}}_{n+1} = \tilde{\alpha}_n + \tilde{\beta}_n x_n$ (with $\tilde{\alpha}_n, \tilde{\beta}_n$ OLS estimators) are worse than HA forecasts (with exclusion restriction $\beta = 0$). CT (2008) show that PC and PF produce better forecasts than HA. We will show that PC-GH and PF-GH further improve PC and PF.

The reason for this improvement is the following equivalence that is motivated by the second equality in Eq. (14.3).

Proposition 14.1. *Bagging the positive coefficient forecast y^{PC}_{T+1} is equivalent to computing the forecast $y^{PC\text{-}GH}_{T+1,J}$ from the Gordon and Hall (2009) bagging estimators $\hat{\alpha}_{T,J}$, $\hat{\beta}_{T,J}$. That is,*

$$\frac{1}{J}\sum_{j=1}^{J} y^{PC,*(j)}_{T+1} = y^{PC\text{-}GH}_{T+1,J}.$$

Proof.

$$\frac{1}{J}\sum_{j=1}^{J} y_{T+1}^{\mathrm{PC},*(j)} = \frac{1}{J}\sum_{j=1}^{J}(\bar{\alpha}_T^{*(j)} + \bar{\beta}_T^{*(j)} x_T) = \left(\frac{1}{J}\sum_{j=1}^{J}\bar{\alpha}_T^{*(j)}\right) + \left(\frac{1}{J}\sum_{j=1}^{J}\bar{\beta}_T^{*(j)}\right) x_T$$
$$= \hat{\alpha}_{T,J} + \hat{\beta}_{T,J} x_T = y_{T+1,J}^{\mathrm{PC\text{-}GH}}. \qquad \square$$

Even though Proposition 14.1 is a very simple insight, it provides a powerful reason, in view of Breiman (1996), why bagging the coefficient estimates and obtaining $y_{T+1}^{\mathrm{PC\text{-}GH}}$ can improve the PC constrained forecast y_{T+1}^{PC} used in CT (2008).

14.2.3 *AMSE comparison*

With our objective in view, we compare the asymptotic mean-square error (AMSE) of the unrestricted, the restricted, and the bagging estimator and forecast. Keeping the decomposition of MSE into variance and bias in mind, we show that under certain circumstances, bagging estimators and forecasts have a shrinkage advantage over simple constrained estimators and forecasts as used in CT (2008). That is, their reduction in variance outweighs their increase in bias. For this purpose, we collect some results from the literature, in particular from Bühlmann and Yu (2002) and from Gordon and Hall (2009), and express them in a unified framework.

Let θ denote either a prediction $\varphi(x_t, \mathcal{D}_t)$ from a regression or the parameter vector (α, β). The unrestricted estimator $\tilde{\theta}_T$ is thus,

$$\tilde{\theta}_T = \tilde{\alpha}_T + \tilde{\beta}_T x_T, \text{ or } \tilde{\theta}_T = (\tilde{\alpha}_T, \tilde{\beta}_T).$$

The restricted estimator for θ subject to a lower bound θ_1 is

$$\bar{\theta}_T = \max\{\tilde{\theta}_T, \theta_1\}.$$

The Gordon and Hall (GH 2009) bagging estimator is

$$\hat{\theta}_T = \lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}\max\{\tilde{\theta}_T^{*(j)}, \theta_1\} = \lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}\bar{\theta}_T^{*(j)} = \mathbb{E}(\max\{\tilde{\theta}_T^{*}, \theta_1\}|\mathcal{D}_T)$$

for the situation where a lower bound θ_1 is known. As before, $\mathcal{D}_T$ is the available data set at time T, $\mathcal{D}_T^*$ is a bootstrap sample, and $\tilde{\theta}_T^*$ is a bootstrap replication of $\tilde{\theta}_T$ from $\mathcal{D}_T^*$. There are J such bootstrap replications; expectation statements hold for $J \to \infty$; in practice J is of course finite. We use the moving block bootstrap method for sub-sampling.

Consider the case where θ is the regression coefficient β. Let the data-generating process be

$$y_{t+1} = \alpha_0 + \beta_0 x_t + u_{t+1},$$

where $\mathbb{E}(u_t) = 0$, $\mathbb{V}(u_t) = \sigma_u^2 < \infty$. Let the data-generating slope parameter be of a local-to-threshold form

$$\beta_0 = \beta_1 + b\sigma_\beta T^{-1/2}. \tag{14.4}$$

Here, β_1 is the threshold applied in the constraint. For example, in the equity premium application where the hypothesis of predictability is studied against a null of no predictability, $\beta_1 = 0$. Note that the term "data-generating" then refers to the choice of hypotheses and does not make a statement about financial economics. Equation (14.4) does not make a statement whether or not returns are asymptotically predictable. If a researcher wanted to study the base case of some predictability rather than no predictability, they would specify a nonzero β_1, possibly of a parametric form derived from financial economic theory. Note that $(\beta_0 - \beta_1)/(\sigma_\beta\sqrt{T}) = b$, so σ_β^2 and b are parameters that control the distance of β_0 from β_1 in finite samples and resemble "variance" and critical value in a "studentized" fraction. (This is not strictly a studentization because no object is actually random here.)

The estimated model is a simple linear regression of y_{t+1} onto x_t, and the estimators under consideration are the OLS estimator $\tilde{\beta}_T$ of UF, the simple PC constrained estimator

$$\bar{\beta}_T = \max\left\{\tilde{\beta}_T \mathbf{1}\left(\tilde{\beta}_T > \beta_1 + c\sigma_\beta T^{-1/2}\right),\ \beta_1\right\} = \max\left\{\tilde{\beta}_T \mathbf{1}\left(\sqrt{T}(\tilde{\beta}_T - \beta_1)/\sigma_\beta > c\right),\ \beta_1\right\},$$

and the bagging estimator PC-GH

$$\hat{\beta}_T = \frac{1}{J}\sum_{j=1}^{J} \bar{\beta}_T^{*(j)} = \frac{1}{J}\sum_{j=1}^{J} \max\left\{\tilde{\beta}_T^{*(j)} \mathbf{1}\left(\sqrt{T}(\tilde{\beta}_T^{*(j)} - \beta_1)/\sigma_\beta > c\right),\ \beta_1\right\}.$$

In this setup, c is a critical value that the scaled and studentized difference between the estimator and the threshold value β_1 has to exceed before the estimator is adopted over the threshold value. In this sense, c is the researcher's guess for b. The two parameters b and c are generally not the same, and the distance $b - c$, the quality of the guess, figures in the asymptotic distributions in Proposition 14.2 below.

In Bühlmann and Yu (2002) and in Gordon and Hall (2009) one finds the following results on the asymptotic distributions of the estimators and their dependence on the data-generating perturbation b and on the critical decision value c.

Proposition 14.2. *(Bühlmann and Yu (2002), Gordon and Hall (2009))*
Let $Z \sim \mathrm{N}(0,1)$, $\Phi(\cdot)$ the cumulative distribution function, and $\phi(\cdot)$ the probability density function of the standard normal distribution. Then, as $T \to \infty$,

$$\sqrt{T}\sigma_\beta^{-1}(\tilde{\beta}_T - \beta_1) \xrightarrow{d} Z + b. \tag{14.5}$$

$$\sqrt{T}\sigma_\beta^{-1}(\bar{\beta}_T - \beta_1) \xrightarrow{d} (Z+b)\mathbf{1}\,(Z+b-c>0)\,. \tag{14.6}$$

$$\sqrt{T}\sigma_\beta^{-1}(\hat{\beta}_T - \beta_1) \xrightarrow{d} (Z+b)\Phi(Z+b-c) + \phi(Z+b-c). \tag{14.7}$$

The positive coefficient (PC) constraint in Eq. (14.2) means that $c = 0$ and $\beta_1 = 0$. For this case, we compare the asymptotic mean-squared error (AMSE) of the limiting random variables in Eqs. (14.5), (14.6), and (14.7) of Proposition 14.2. Equation (14.7) also shows that the bagging estimator is a smoothly transitioning sigmoid function.

Asymptotic bias (Abias): For all $z \in \mathbb{R}$, note that

$$z + b \le (z+b)\,\mathbf{1}(z+b>0) < (z+b)\Phi(z+b) + \phi(z+b),$$

which results in the following order of the asymptotic biases:

$$\text{Abias of } \tilde{\beta}_T \le \text{Abias of } \bar{\beta}_T < \text{Abias of } \hat{\beta}_T,$$

where

$$\begin{aligned}
\text{Abias of } \tilde{\beta}_T &= \mathbb{E}(Z+b) - b = 0,\\
\text{Abias of } \bar{\beta}_T &= \mathbb{E}\left[(Z+b)\,\mathbf{1}(Z+b>0)\right] - b,\\
\text{Abias of } \hat{\beta}_T &= \mathbb{E}\left[(Z+b)\Phi(Z+b) + \phi(Z+b)\right] - b.
\end{aligned}$$

Therefore, Abias gets worse as we impose the restriction and as we add bagging. For example, when $b = 0$,

$$\begin{aligned}
\text{Abias of } \tilde{\beta}_T &= \mathbb{E}(Z) = 0,\\
\text{Abias of } \bar{\beta}_T &= \mathbb{E}\left[Z\mathbf{1}(Z>0)\right] = 0.5\sqrt{2/\pi} \approx 0.3989,\\
\text{Abias of } \hat{\beta}_T &= \mathbb{E}\left[Z\Phi(Z) + \phi(Z)\right] = \pi^{-1/2} \approx 0.5642.
\end{aligned}$$

Asymptotic variance (Avar): While Abias increases, asymptotic variance is reduced by imposing a restriction and more so by bagging the restriction. For $b = 0$, the asymptotic distribution of the simple constrained estimator $\bar{\beta}_T$ is a standard normal truncated to the positive half-line and thus has asymptotic variance $\mathbb{V}(Z\mathbf{1}(Z>0)) = (1 - 1/\pi)/2) = 0.3408$. The Avar of the bagging estimator is $\mathbb{V}(Z\Phi(Z) + \phi(Z)) = 1/3 + \sqrt{3}/(2\pi) - 1/\pi = 0.2907$. AMSE$(\hat{\beta}_T) = 0.6088$ and AMSE$(\bar{\beta}_T) = 0.4998$ are substantially smaller than AMSE$(\tilde{\beta}_T) = 1$. Note that the

Table 14.1. AMSE comparisons

b	$\tilde{\beta}_T$ Abias	$\tilde{\beta}_T$ Avar	$\tilde{\beta}_T$ AMSE	$\bar{\beta}_T$ Abias	$\bar{\beta}_T$ Avar	$\bar{\beta}_T$ AMSE	$\hat{\beta}_T$ Abias	$\hat{\beta}_T$ Avar	$\hat{\beta}_T$ AMSE
–2.0	0.0000	1.0000	1.0000	2.0085	0.0057	4.0397	2.0502	0.0135	4.2169
–1.0	0.0000	1.0000	1.0000	1.0833	0.0684	1.2420	1.1996	0.0839	1.5231
–0.5	0.0000	1.0000	1.0000	0.6978	0.1705	0.6574	0.8491	0.1674	0.8884
0.0	0.0000	1.0000	1.0000	0.3989	0.3409	0.5001	0.5642	0.2907	0.6090
0.5	0.0000	1.0000	1.0000	0.1978	0.5535	0.5926	0.3491	0.4438	0.5657
1.0	0.0000	1.0000	1.0000	0.0833	0.7511	0.7581	0.1996	0.6045	0.6443
1.5	0.0000	1.0000	1.0000	0.0294	0.8884	0.8893	0.1049	0.7475	0.7585
2.0	0.0000	1.0000	1.0000	0.0083	0.9602	0.9602	0.0501	0.8562	0.8587
2.5	0.0000	1.0000	1.0000	0.0019	0.9886	0.9886	0.0218	0.9271	0.9276
3.0	0.0000	1.0000	1.0000	0.0003	0.9974	0.9974	0.0085	0.9672	0.9673
3.5	0.0000	1.0000	1.0000	0.0002	0.9997	0.9997	0.0032	0.9871	0.9871
4.0	0.0000	1.0000	1.0000	0.0002	0.9998	0.9998	0.0011	0.9953	0.9953

ratio AMSE($\hat{\beta}_T$)/AMSE($\bar{\beta}_T$) = 1.2181. The AMSE of PC-GH is worse due to much larger Abias($\hat{\beta}_T$). However, the case $\beta = 0$ (i.e., $b = 0$) is not an interesting one. We impose the restriction $\beta > 0$ (i.e., $b > 0$). For other values of b, analytical evaluations are difficult and we use numerical evaluation. The results for ~~some~~ a grid~~s~~ of b on [-2 4] are presented in Table 14.1.

AMSE Comparison: Summarizing the results of Table 14.1, we note the following observations.

1. When $\beta > 0$ (i.e., $b > 0$), if we impose the *correct* restriction $\beta > 0$, we can improve AMSE. For example, when $b = 1$ in Table 14.1, AMSE($\hat{\beta}_T$)/AMSE($\bar{\beta}_T$) = 0.85. Hence PC-GH reduces AMSE by 15% from PC. When $b = 2$, AMSE($\hat{\beta}_T$)/AMSE($\bar{\beta}_T$) = 0.8943, and thus PC-GH is about 11% better than PC.
2. When $\beta \gg 0$ (i.e., $b \gg 0$), if we impose the *obvious* restriction $\beta > 0$, we do not gain much. For example, when $b = 3$ in Table 14.1, the restriction is hardly binding and the gain is small. Bagging makes a minor contribution for AMSE($\hat{\beta}_T$)/AMSE($\bar{\beta}_T$) = 0.9698. When the restriction becomes even more obviously correct with $b = 4$, the gain becomes even smaller. PC is the same as UF in AMSE. PC-GH is only slightly better than PC with AMSE($\hat{\beta}_T$)/AMSE($\bar{\beta}_T$) = 0.9955.
3. When $\beta \ll 0$ (i.e., $b \ll 0$), if we impose the *wrong* restriction $\beta > 0$, then the increase in Abias dominates the reduction in Avar. AMSE substantially deteriorates by imposing the wrong constraint and more so by bagging the constraint, as can be seen from Table 14.1 for $b = -1$ or -2.

14.3 Simulation

Simulation design: In order to evaluate the performance of the restricted and bagging predictors, we construct a simulation experiment that is motivated by the stock-return prediction problem. First, we generate $\{w_t\}_{t=1}^{T=200}$ from

$$w_t = \rho w_{t-1} + e_t, \qquad e_t \sim NID\left(0, \sigma_e^2\right), \ \sigma_e = 0.2,$$

where $\rho \in \{0, 0.5, 0.9, 0.99\}$ setting different levels of persistence of the regressor. Next, we let $x_t = \exp(w_t)/\text{std}(\exp(w_t))$ so that $\{x_t\}$ be positive to mimic the predictor variables that we will consider in the equity premium prediction in Section 14.4. The predictor x is normalized with the standard deviation std$(\cdot)$. Then, we generate $\{y_t\}_{t=1}^{T=200}$ from

$$y_t = 0.02 + \beta x_{t-1} + u_t, \qquad u_t \sim NID\left(0, \sigma_u^2\right), \ \mathbb{E}(u_t e_s) = 0, \forall t, s,$$

where $\sigma_u \in \{0.1, 1\}$. The data-generating value of β deviates from the bound β_1 by the local drift parameter b with rate $T^{-1/2}$, that is, $\beta = \beta_1 + b\,\sigma_\beta\, T^{-1/2}$. We set the values of β_1, σ_β, and b as follows: $\beta_1 = 0$, $\sigma_\beta^2 = \sigma_u^2\left(\sum_{t=1}^{100}(x_t - \bar{x})^2\right)^{-1}$, and $b \in \{1, 3, 5, 10, 15, 20, 30, 50, 100\}$.

We use the first half of the ~~total~~ sample size 200 to estimate β using the unrestricted, restricted, and the bagging (PC-GH) estimators. The PC-GH estimator is computed over $J = 200$ bootstrap samples. Using each of the above estimators at time $t = 100, \ldots, 199$, we compute the unrestricted, restricted, and bagging forecasts $\theta_{t+1}(x_t)$ of y_{t+1}. The forecasts over the second half of observations are compared with the actual value y_{t+1}. We compute the following out-of-sample predictive ability measure of each model

$$100 \cdot R_{\text{OS}}^2 = 100\left(1 - \frac{\frac{1}{100}\sum_{t=101}^{200}(y_t - \theta_t(x_{t-1}))^2}{\frac{1}{100}\sum_{t=101}^{200}(y_t - \theta_t^{\text{HA}})^2}\right),$$

where the historical average $\theta_t^{\text{HA}} = \frac{1}{100}\sum_{s=t-100}^{t-1} y_s$ (for $t = 101, \ldots, 200$) is taken as a benchmark forecast as in CT (2008). The same statistic $\left(100 \cdot R_{\text{OS}}^2\right)$ was used in CT (2008) to compare various forecast models. We repeat the steps above over 1000 Monte Carlo replications and compute the average of the out-of-sample $\left(100 \cdot R_{\text{OS}}^2\right)$.

Simulation results: In this simulation, unlike in the empirical application to equity premium prediction, the unrestricted forecast UF always dominates HA by construction of the simulation design with $b > 0$. Hence, it is not interesting to compare the forecasts with HA for this simulation section as we do it in the empirical section. Here we present the gain in $100 \cdot R^2_{\mathrm{OS}}$ over the unrestricted forecast UF from imposing a constraint and from bagging. Figures 14.1 and 14.2 report this gain, defined as

$$\text{Gain-in-}R^2 = \left(100 \cdot R^2_{\mathrm{OS}}\right)_{\mathrm{model}} - \left(100 \cdot R^2_{\mathrm{OS}}\right)_{\mathrm{UF}}, \tag{14.8}$$

where *model* = PC, PC-GH, PF, or PF-GH. Each of Figures 14.1 and 14.2 reports the gain in R^2 for two different values of σ_u. The four panels of each figure show the situation for one of the four different values of ρ. In each panel, the abscissa shows the different values of b, and the ordinate shows the gain as defined in Eq. (14.8) for that specific 3-tuple (σ_u, ρ, b). Summarizing, we make the following observations.

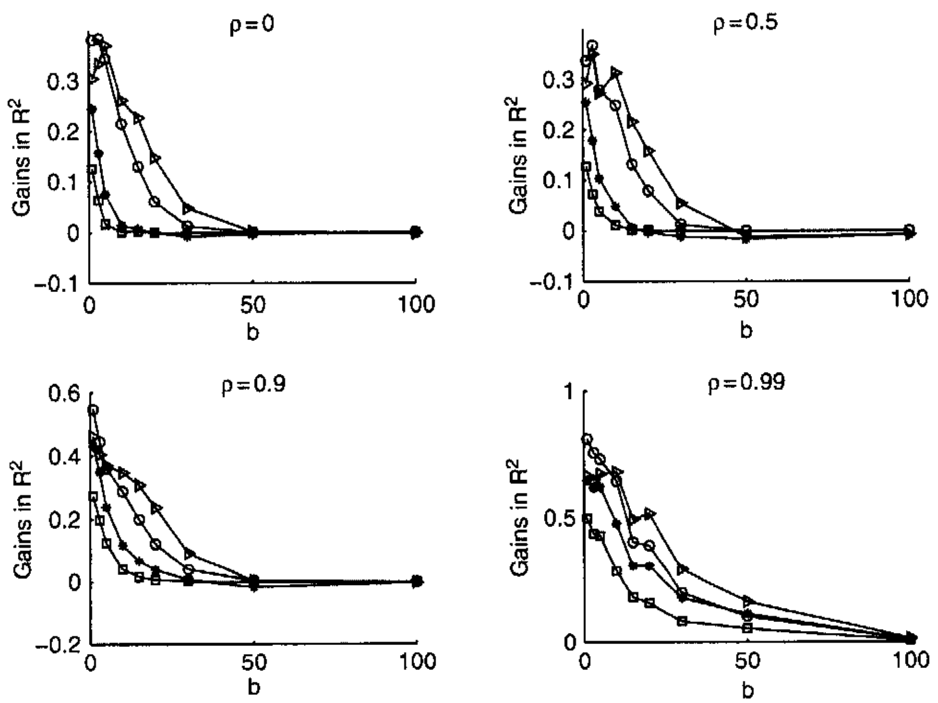

Figure 14.1. Gains in $100 \cdot R^2_{\mathrm{OS}}$ from imposing constraint and bagging over UF when $\sigma_u = 0.1$.

Notes: The gain of a model in $100 \cdot R^2_{\mathrm{OS}}$ over UF is $\left(100 \cdot R^2_{\mathrm{OS}}\right)_{\mathrm{model}} - \left(100 \cdot R^2_{\mathrm{OS}}\right)_{\mathrm{UF}}$ for each of model = PC (line with circles o), PC-GH (line with triangles △), PF (line with squares □), or PF-GH (line with asterisks *).

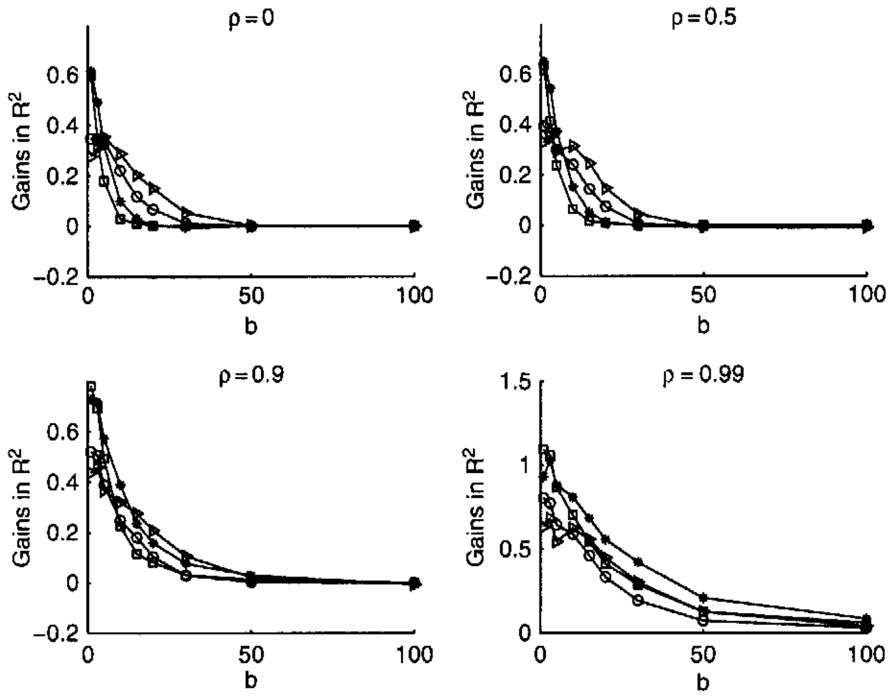

Figure 14.2. Gains in $100 \cdot R^2_{OS}$ from imposing constraint and bagging over UF when $\sigma_u = 1$.
Notes: See Figure 14.1.

1. For a wide range of b, the gains over UF from imposing the PC constraint and PF constraint are positive.
2. For a wide range of b, bagging further improves the constrained forecasts. PC is further improved by PC-GH, and PF is further improved by PF-GH.
3. The gain~~s~~ from the PF constraint is smaller when σ_u is smaller (Figure 14.1) than when σ_u is larger (Figure 14.2), indicating that the PF constraint may be more useful when the market becomes more volatile.
4. Note that high persistence (high ρ) leads to a large value of $\sum_{t=1}^{100}(x_t - \bar{x})^2$ and thus reduces σ_β, which reduces the local drift $(b\,\sigma_\beta\,T^{-1/2})$ from the bound. Therefore, for higher ρ, the effects of the constraint or of bagging are bigger for a given value of b, and the effects are present over a larger range of b.[2] (b.², not b-squared)

[2] The effect of the persistence (measured by ρ) in the predictor x of the predictive regression on the finite sample estimation bias in β has been widely studied since the seminal paper by Stambaugh (1999). Examples are Valkanov (2003), Lewellen (2004), Campbell and Yogo (2006), and Zhu (2013). In that context it is interesting to observe the effect of the persistence ρ on gains from using the constraint and bagging, which is examined here from Figures 14.1–14.2.

14.4 Let's do it again: equity premium prediction

14.4.1 *Data and constraints*

Data: We use the data set of Campbell and Thompson (2008), which was kindly provided by Sam Thompson. The data frequency is monthly; the sample period is 1871–2005. Excess returns on the S&P 500 are calculated from the returns time series (1871M2 through 2005M12, CRSP since 1927) and the 3-month Treasury-Bill interest rate (denoted as r_t^f, 1920M1 through 2005M12, 1870M2 through 1919M12 calculated following Goyal and Welch (2008)). The predictor variables are the dividend yield ***d/p*** (1872M2 through 2005M12), earnings yield ***e/p*** (1872M2 through 2005M12), smoothed earnings yield ***se/p*** following Campbell and Shiller (1988b), Campbell and Shiller (1998) (1881M1 through 2005M12), book-to-market ratio ***b/m*** (1926M6 through 2005M12), smoothed return on equity ***roe*** as described in Campbell and Thompson (2008) (1936M6 through 2005M12), the 3-month Treasury-Bill ***tbl*** (1920M1 through 2005M12), long-term government bond yield ***lty*** (1870M1 through 2005M12), the term spread ***ts***, that is, the difference between long-term and short-term treasury yields (1920M1 through 2005M12), the default spread ***ds***, that is, the difference between corporate and Treasury bond yields (1919M1 through 2005M12), the lagged inflation rate ***inf*** (1871M5 through 2005M12), and the equity share of new issues ***nei*** proposed by Baker and Wurgler (2000). See Table 14.2 for a summary.

Table 14.2. Data and restrictions

x	sign(β)		estimation starts from	forecasting starts from
d/p	+	dividend yield	1872M2	1927M1
e/p	+	earnings yield	1872M2	1927M1
se/p	+	smoothed earnings yield	1881M2	1927M1
b/m	+	book-to-market ratio	1926M6	1946M6
roe	+	smoothed return on equity	1936M6	1956M6
tbl	−	3-month Treasury Bill	1920M1	1940M1
lty	−	long-term government bond yield	1871M2	1927M1
ts	+	term spread	1920M1	1940M1
ds	+	default spread	1919M1	1939M1
inf	−	inflation rate	1871M5	1927M1
nei	−	equity share of new issues	1927M12	1947M12

Notes: We use the same data set as Campbell and Thompson (2008), which was kindly provided by Sam Thompson. The data frequency is monthly. See Section 14.4.1 for details. The PC constraints of CT (2008) for each predictor are shown in column 2. When the sign constraint on the coefficient β is negative, the positive constraint PC should be understood as a negative constraint. We do not distinguish the cases in the text for better readability. The term spread (ts) is the long-term minus the short-term Treasury yields. The default spread (ds) is the corporate bond yield minus the Treasury bond yield.

Comparing out-of-sample predictive ability: For the prediction exercise presented in this section, we use the estimation sample periods and forecast periods as shown in Table 14.2, which are the same as in Campbell and Thompson (2008). We take the "recursive scheme" that uses expanding windows for estimating models. We move the estimation window forward by one month to estimate the models. We keep rolling the estimation sample forward until the last one-month ahead forecast is made for the month of 2005M12.

Constraints: We apply sign restrictions on the coefficients β depending on the predictor and a positivity restriction on the forecast y_{t+1} of the risk premium, as well as a combination of these two. The coefficient restrictions for the different predictors are listed in Table 14.2; they are the same as in CT (2008). We impose the hard constraint and we apply bagging to smooth it. This results in the following set of forecasts: UF, PC, PF, PCF (applying positivity on coefficient and forecast jointly), PC-GH, PF-GH, and PCF-GH (bagging the joint restriction).

14.4.2 *Empirical results*

We compare the prediction performance in terms of three criteria. Table 14.3 compares MSFE in $100 \cdot R^2_{\mathrm{OS}}$ proposed by CT (2008). Table 14.4 modifies the $100 \cdot R^2_{\mathrm{OS}}$ with the adjustment in MSFE of Clark and West (2006). Table 14.5 reports the utility function of an investor with simple mean-variance preferences U = expected portfolio return - $\gamma/2$ portfolio variance as proposed by CT (2008).

14.4.2.1 COMPARING MSFE IN $100 \cdot R^2_{\mathrm{OS}}$ OF CT (2008)

Table 14.3 presents the results that are directly related to Goyal and Welch (2008) and Campbell and Thompson (2008). The reported numbers are out-of-sample R^2 statistics R^2_{OS} multiplied by 100.

$$100 \cdot R^2_{\mathrm{OS}} = 100\left(1 - \frac{\frac{1}{P}\sum_{t=T-P+1}^{T}(y_t - \theta_t(x_{t-1}))^2}{\frac{1}{P}\sum_{t=T-P+1}^{T}(y_t - \theta_t^{\mathrm{HA}})^2}\right), \tag{14.9}$$

where P denotes the number of out-of-sample forecasts; $\theta_t(x_{t-1})$ is the prediction from the UF, PC, PF, PCF, PC-GH, PF-GH, and PCF-GH model, respectively. These models are organized in the columns of Tables 14.3 through 14.5. The rows of the tables show the different univariate predictors that are used for x, from dividend yield ***d/p*** through new issues ***nei***. There are 11 such predictors. The last row reports a simple, equally weighted, combined forecast from all 11 individual forecasts. The first reported numbers in Table 14.3 are out-of-sample statistics $100 \cdot R^2_{\mathrm{OS}}$. The second reported numbers in parentheses are the

Table 14.3. Relative gains in MSFE over HA

	UF	PC	PF	PCF	PC-GH	PF-GH	PCF-GH
d/p	−0.65 (−0.57)	0.05 (0.12)	0.07 (0.18)	0.08 (0.21)	0.31 (0.56)	0.53 (1.13)	0.38 (0.71)
e/p	0.12 (0.51)	0.18 (0.79)	0.14 (0.57)	0.18 (0.79)	0.25 (1.00)	0.20 (0.90)	0.25 (1.00)
se/p	0.33 (0.85)	0.42 (1.11)	0.38 (1.01)	0.43 (1.16)	0.48 (1.19)	0.50 (1.35)	0.50 (1.25)
b/m	−0.39 (−0.41)	−0.39 (−0.41)	0.03 (0.04)	0.03 (0.04)	−0.43 (−0.43)	0.34 (0.66)	0.02 (0.02)
roe	−0.92 (−2.15)	−0.06 (−0.38)	−0.92 (−2.15)	−0.06 (−0.38)	−0.68 (−1.43)	−0.85 (−2.05)	−0.68 (−1.43)
tbl	0.49 (0.48)	0.48 (0.46)	0.54 (0.91)	0.52 (0.89)	0.53 (0.51)	0.52 (0.92)	0.56 (0.90)
lty	−0.19 (−0.33)	−0.19 (−0.33)	0.20 (0.70)	0.20 (0.70)	−0.16 (−0.28)	0.23 (0.82)	0.22 (0.75)
ts	0.43 (0.58)	0.45 (0.60)	0.43 (0.59)	0.44 (0.61)	0.59 (0.78)	0.30 (0.41)	0.57 (0.76)
ds	−0.24 (−0.69)	−0.24 (−0.69)	−0.24 (−0.69)	−0.24 (−0.69)	−0.50 (−1.24)	0.12 (0.35)	−0.42 (−1.07)
inf	−0.22 (−0.92)	−0.21 (−0.87)	−0.18 (−0.78)	−0.17 (−0.73)	−0.16 (−0.60)	−0.08 (−0.34)	−0.11 (−0.43)
nei	0.32 (0.34)	0.32 (0.34)	0.48 (0.55)	0.48 (0.55)	0.30 (0.32)	0.39 (0.45)	0.47 (0.54)
CF	0.65 (2.10)	0.67 (2.00)	0.50 (2.40)	0.52 (2.39)	0.72 (2.19)	0.46 (2.83)	0.56 (2.61)

Notes: The reported numbers are out-of-sample statistics R^2_{OS} multiplied by 100, following Campbell and Thompson (2008), as defined in Eq. (14.9). This is to compare seven forecasts in seven columns using each of the 11 predictors (in 11 rows). The last row (row 12) presents the equally-weighted combined-forecast (CF) of the 11 forecasts in each column. R^2_{OS} measures the relative gain of a predictive regression over HA. The numbers in parentheses are the values of statistics to test the null hypothesis that MSFE gain is zero.

values of statistics to test the null hypothesis that MSFE gain is zero[3]. We note the following observations from Table 14.3.

1. Positive values of $100 \cdot R^2_{OS}$ indicate that a model is better than HA. Many values of $100 \cdot R^2_{OS}$ in Table 14.3 are negative, indicating that it is not easy to beat the historical average. However, many of these values become larger or turn positive when the PC, PF, or PCF constraints are imposed. The values of statistics tend to get larger as well when the constraints are imposed, even if none of them are significantly positive.

[3] Assuming that the estimation sample is finite while we let $P \to \infty$, we can apply Giacomini and White (2006). Then the Diebold–Mariano statistic is asymptotically standard normal under the null that $\lim_{P\to\infty} P^{-1/2} \sum_{t=T-P+1}^{T} E\left(\hat{u}^2_{0,t+1} - \hat{u}^2_{1,t+1}\right) = 0$ where the subscript 0 denotes the HA model and the subscript 1 denotes any of the other models, which may nest the HA model. In this case we are not testing for the null hypothesis that $E\left(u^2_{0,t+1} - u^2_{1,t+1}\right) = 0$. Instead we are testing for the null hypothesis that $\lim_{P\to\infty} P^{-1/2} \sum_{t=T-P+1}^{T} E\left(\hat{u}^2_{0,t+1} - \hat{u}^2_{1,t+1}\right) = 0$ using the estimated forecast errors. We thank a referee for pointing this out.

2. The constraints work. Many of the 11 PC forecasts are better than the unconstrained forecast. Ten of 11 PCs are at least as good as UF. All 11 PFs are at least as good as UFs. So are all 11 PCFs.
3. Bagging works for 7 of 11 cases when PC is compared to PC-GH, for 7 of 11 cases when PF is compared to PF-GH, and for 7 out of 11 when PCF is compared to PCF-GH.
4. In general CF produces the best forecast in each column, dominating all individual forecasts in most of the seven columns. This is observed in Table 14.3 (but not in Tables 14.4, 14.5). Bagging works for PC and PCF constraints, as seen from pairwise comparing the numbers in the last row. It is interesting to note that none of the MSFE gains are significantly positive in rows 1–11, but all of the CFs in rows significantly positive with all the t statistics over 2.00. The statistics are even larger with bagging.

In summary, it is hard to beat HA with UF, but imposing the constraints and bagging can improve UF. Many constrained and bagged predictions outperform HA. Combined forecasts (CF) outperform HA for all constrained models and bagging further improves the forecast power.

14.4.2.2 ADJUSTED $100 \cdot R^2_{\mathrm{OS}}$ OF CLARK AND WEST (2006)

Campbell and Thompson write in CT (2008, p. 1515, footnote 5) that "Clark and West (2006) point out that if the return series is truly unpredictable, then in a finite sample the predictive regression will on average have a higher mean squared prediction error because it must estimate an additional coefficient. Thus, the expected out-of-sample R^2 under the null of unpredictability is negative, and a zero out-of-sample R^2 can be interpreted as weak evidence for predictability. We do not pursue this point here because, like Goyal and Welch (2008), we ask whether predictive regressions or historical average return forecasts have delivered better out-of-sample forecasts, not whether stock returns are truly predictable." Following these lines, we have studied (in Table 14.3) whether the constrained or bagged predictive regressions can beat HA, but we have not studied whether stock returns are truly predictable by using various predictors x.

As suggested by Campbell and Thompson (2008) we use the out-of-sample R^2_{OS} in Eq. (14.9), which compares the MSFE $\frac{1}{P}\sum_{t=T-P+1}^{T}(y_t - \theta_t(x_{t-1}))^2$ of a predictive regression with the MSFE $\frac{1}{P}\sum_{t=T-P+1}^{T}(y_t - \theta_t^{\mathrm{HA}})^2$ of HA. To compare the MSFEs, the test statistics of Diebold and Mariano (1995) and West (1996) use the MSFE differential

$$\frac{1}{P}\sum_{t=T-P+1}^{T}\left[(y_t - \theta_t^{\mathrm{HA}})^2 - (y_t - \theta_t(x_{t-1}))^2\right] \qquad (14.10)$$

to test the null hypothesis that $\mathbb{E}\left[(y_t - \theta_t^{\mathrm{HA}})^2 - (y_t - \theta_t(x_{t-1}))^2\right] = 0$. Note that R^2_{OS} in Eq. (14.9), reported in Table 14.3, is obtained from dividing (14.10) by $\frac{1}{P}\sum_{t=T-P+1}^{T}(y_t - \theta_t(x_{t-1}))^2$.

Clark and West (CW 2006) show that when a predictive regression model (using x) is compared with the HA, under the null hypothesis of no predictive ability of x, its MSFE is expected to be greater than the HA's MSFE. CW propose an adjustment to the MSFE differential in Eq. (14.10) in order to account for the disadvantage of the sample MSFE of the predictive model. The CW-adjusted MSFE differential is

$$\frac{1}{P}\sum_{t=T-P+1}^{T}\left[(y_t - \theta_t^{\mathrm{HA}})^2 - \left\{(y_t - \theta_t(x_{t-1}))^2 - (\theta_t^{\mathrm{HA}} - \theta_t(x_{t-1}))^2\right\}\right]. \tag{14.11}$$

We use the following "CW-adjusted-R^2_{OS}" defined by dividing Eq. (14.11) by $\frac{1}{P}\sum_{t=T-P+1}^{T}(y_t - \theta_t^{\mathrm{HA}})^2$:

$$\text{CW-adjusted-}100{\cdot}R^2_{\mathrm{OS}} = 100\left(1 - \frac{\frac{1}{P}\sum_{t=T-P+1}^{T}\left\{(y_t - \theta_t(x_{t-1}))^2 - (\theta_t^{\mathrm{HA}} - \theta_t(x_{t-1}))^2\right\}}{\frac{1}{P}\sum_{t=T-P+1}^{T}(y_t - \theta_t^{\mathrm{HA}})^2}\right).$$

Table 14.4 presents the CW-adjusted-100 $\cdot R^2_{\mathrm{OS}}$ and the test statistics in parentheses to test the null hypothesis that the MSFE gain (14.11) with the Clark and West adjustment is zero[4]. We note the following observations.

1. Positive values of the CW-adjusted-100 $\cdot R^2_{\mathrm{OS}}$ indicate that a predictive regression model using x is better than HA. Most values in Table 14.4 are positive. The numbers in Table 14.4 are not only larger than those in Table 14.3, which is by construction, but also more highly significant according to the test statistics in parentheses.
2. The constraints work, with a few exceptions for PC and PF.
3. Bagging works well for many cases for PC-GH and PCF-GH.
4. While the combined forecast (CF) is nowhere the best, it is consistently better than HA across all constraints. Bagging works for CF as in the previous tables. The combined forecast with the PC constraint is further improved by bagging (PC-GH is better), the CF with PF is further improved by bagging (PF-GH is better), and the CF with PCF is also further improved by bagging (PCF-GH is better).

14.4.2.3 HOW MUCH R^2 IS ECONOMICALLY MEANINGFUL?

Although the gains from imposing constraints and bagging presented in Table 14.2 are small, they can be economically meaningful for mean-variance

[4] The CW statistic of Clark and West (2006) is closely related to the ENC-T of Clark and McCracken (2001).

Table 14.4. Relative gains in MSFE over HA with the adjustment of Clark and West (2006)

	UF	PC	PF	PCF	PC-GH	PF-GH	PCF-GH
d/p	−0.11 (−0.10)	0.49 (1.23)	0.51 (1.24)	0.51 (1.30)	0.83 (1.49)	0.77 (1.64)	0.86 (1.58)
e/p	0.32 (1.34)	0.38 (1.64)	0.34 (1.41)	0.38 (1.64)	0.46 (1.85)	0.39 (1.72)	0.46 (1.85)
se/p	0.71 (1.84)	0.80 (2.12)	0.76 (2.00)	0.81 (2.16)	0.87 (2.16)	0.80 (2.17)	0.88 (2.20)
b/m	1.15 (1.20)	1.15 (1.20)	0.95 (1.30)	0.95 (1.30)	1.23 (1.24)	0.76 (1.48)	0.97 (1.30)
roe	−0.62 (−1.45)	−0.03 (−0.20)	−0.62 (−1.45)	−0.03 (−0.20)	−0.28 (−0.60)	−0.57 (−1.36)	−0.28 (−0.60)
tbl	2.04 (1.97)	2.02 (1.95)	1.14 (1.95)	1.13 (1.92)	2.16 (2.04)	1.07 (1.89)	1.23 (1.98)
lty	0.81 (1.40)	0.81 (1.40)	0.53 (1.80)	0.53 (1.80)	0.84 (1.45)	0.52 (1.86)	0.56 (1.86)
ts	1.49 (2.00)	1.48 (2.00)	1.46 (1.99)	1.45 (1.99)	1.67 (2.19)	1.32 (1.81)	1.62 (2.16)
ds	−0.04 (−0.11)	−0.04 (−0.11)	−0.04 (−0.11)	−0.04 (−0.11)	−0.23 (−0.56)	0.31 (0.91)	−0.16 (−0.42)
inf	−0.13 (−0.55)	−0.12 (−0.50)	−0.11 (−0.45)	−0.09 (−0.40)	−0.06 (−0.21)	−0.02 (−0.08)	−0.02 (−0.09)
nei	1.93 (2.05)	1.93 (2.05)	1.90 (2.19)	1.90 (2.19)	1.91 (2.03)	1.77 (2.09)	1.89 (2.18)
CF	0.73 (2.35)	0.76 (2.28)	0.54 (2.58)	0.57 (2.58)	0.81 (2.45)	0.49 (2.96)	0.61 (2.79)

Notes: The reported numbers are out-of-sample statistics R^2_{OS} multiplied by 100, of Campbell and Thompson (2008), with the adjustment of Clark and West (2006), as discussed in Section 14.4.2.2. As in Table 14.3, we compare seven forecasts in seven columns using each of the 11 predictors (in 11 rows). The last row (row 12) presents the equally-weighted combined-forecast (CF) of the 11 forecasts in each column. The numbers in parentheses are the values of statistics to test the null hypothesis that MSFE gain with the Clark and West adjustment is zero.

investors. As Barberis (2000) points out, "...the evidence of predictability in asset returns affects optimal portfolio choice for investors with long horizons [...] even after incorporating parameter uncertainty, there is enough predictability in returns to make investors allocate substantially more to stocks." To see how gains in R^2 may be translated into economic gain, following CT (2008), we consider an investor with single-period horizon and mean-variance preferences

$$\begin{aligned} U &= \text{expected portfolio return} - \frac{\gamma}{2}\,\text{portfolio variance}, \qquad (14.12)\\ &= \mathbb{E}\left[wy_{t+1} + \left(1-w\right)r^f_{t+1}\right] - \frac{\gamma}{2}\mathbb{V}\left[wy_{t+1} + \left(1-w\right)r^f_{t+1}\right], \end{aligned}$$

where γ captures relative risk aversion. The excess return on a risky asset over the riskless interest rate is given by $y_{t+1} = \alpha + \beta x_t + u_{t+1}$ as in Eq. (14.1). Following CT (2008), we assume x_t has unconditional mean zero and unconditional variance σ_x^2, and the risk-free interest rate is constantly equal to zero. The random shock u_{t+1} has unconditional mean zero and unconditional variance $\sigma_u^2 > 0$. As a result, y_{t+1} has unconditional mean $\mathbb{E}\,(y_{t+1}) = \alpha$ and unconditional variance $\mathbb{V}\,(y_{t+1}) = \beta^2\sigma_x^2 + \sigma_u^2$, assuming independence of x and u.

Without observing x_t, the portfolio weight in the risky asset is

$$w_0 = \frac{1}{\gamma}\frac{\mathbb{E}(y_{t+1})}{\mathbb{V}(y_{t+1})} = \frac{1}{\gamma}\frac{\alpha}{\beta^2\sigma_x^2 + \sigma_u^2},$$

and the equity premium (EP) is

$$EP_0 = \mathbb{E}\left[w_0 y_{t+1} + \left(1 - w_0\right) r_{t+1}^f\right)\right] = \frac{1}{\gamma}\frac{\alpha^2}{\beta^2\sigma_x^2 + \sigma_u^2} = \frac{1}{\gamma}\frac{\mathbb{E}(y_{t+1})^2}{\mathbb{V}(y_{t+1})} = \frac{1}{\gamma}S^2,$$

where S is the Sharpe ratio. Conditional on x_t, the portfolio weight in the risky asset becomes

$$w_t = \frac{1}{\gamma}\frac{\mathbb{E}(y_{t+1}|x_t)}{\mathbb{V}(y_{t+1}|x_t)} = \frac{1}{\gamma}\frac{\alpha + \beta x_t}{\sigma_u^2},$$

with EP

$$EP_1 = \mathbb{E}\left[w_t y_{t+1} + \left(1 - w_t\right) r_{t+1}^f\right)\right] = \frac{1}{\gamma}\frac{\alpha^2 + \beta\sigma_x^2}{\sigma_u^2} = \frac{1}{\gamma}\frac{S^2 + R^2}{1 - R^2},$$

where $R^2 = \frac{\beta\sigma_x^2}{\beta\sigma_x^2 + \sigma_u^2}$. Note that w_0 is constant while w_t is time-varying. The increase in EP from observing x_t is

$$EP_1 - EP_0 = \left(\frac{1 + S^2}{1 - R^2}\right)\frac{1}{\gamma}R^2 > \frac{1}{\gamma}R^2.$$

The relative gain in EP is

$$\frac{EP_1 - EP_0}{EP_0} = \left(\frac{1 + S^2}{1 - R^2}\right)\frac{R^2}{S^2} > \frac{R^2}{S^2}.$$

If R^2 is large w.r.t. S^2, then an investor can use the information in the predictive regression to obtain a large proportional increase in return. CT (2008) report $S^2 = 0.0120$ for the CT data set (monthly 1871–2005).

For example, from the bagging results, the out-of-sample R^2_{OS} for dividend yield (***d/p***) of PCF-GH is 0.0016. The relative EP gain is about 13% for dividend yield as a predictor compared to the HA forecast:

$$\frac{R^2}{S^2} = \frac{0.0016}{0.0120} = 0.13 \text{ or } 13\%.$$

Similarly, the out-of-sample R^2_{OS} for earnings yield (***e/p***) of PCF-GH is 0.0024 and thus the relative EP gain is about 20% when earnings yield is used as predictor and compared to the HA forecast:

$$\frac{R^2}{S^2} = \frac{0.0024}{0.0120} = 0.20 \text{ or } 20\%.$$

14.4.2.4 UTILITY FUNCTION OF CT (2008)

As discussed in the previous subsection, CT (2008) show the economic significance of numerically small R^2's by interpreting them relative to the squared Sharpe ratio. Rapach, Strauss, and Zhou (2010) use the investor utility value in equation (14.12) of CT (2008) to compare the economic values of different forecasts. We adopt the same utility function (14.12) to compare the seven conditional predictive regression models (UF, PC, PF, PCF, PC-GH, PF-GH, and PCF-GH) relative to HA. Table 14.5 reports the utility level (14.12) of an investor using the predictive regression over the utility level of the HA forecast.

The historical average (HA) forecast does not use x_t in forecasting y_{t+1}. The realized average utility level for the HA forecast over the out-of-sample period is

$$\hat{U}_0 = \widehat{\mathbb{E}}\left[w_0 y_{t+1} + \left(1 - w_0\right)r^f_{t+1}\right] - \frac{\gamma}{2}\widehat{\mathbb{V}}\left[w_0 y_{t+1} + \left(1 - w_0\right)r^f_{t+1}\right],$$

where $\widehat{\mathbb{E}}(\cdot)$ and $\widehat{\mathbb{V}}(\cdot)$ are the sample mean and sample variance over the out-of-sample period for the portfolio return $\left[w_0 y_{t+1} + \left(1 - w_0\right)r^f_{t+1}\right]$ that was formed using the HA forecasts of y_{t+1}.

The seven conditional forecasts use a predictor x_t in forecasting y_{t+1}. The realized out-of-sample average utility level for each of these forecasts is

$$\hat{U}_1 = \widehat{\mathbb{E}}\left[w_t y_{t+1} + \left(1 - w_t\right)r^f_{t+1}\right] - \frac{\gamma}{2}\widehat{\mathbb{V}}\left[w_t y_{t+1} + \left(1 - w_t\right)r^f_{t+1}\right],$$

where $\widehat{\mathbb{E}}(\cdot)$ and $\widehat{\mathbb{V}}(\cdot)$ are the out-of-sample sample-mean and sample-variance of the return $\left[w_t y_{t+1} + \left(1 - w_t\right)r^f_{t+1}\right]$ on the portfolio that was formed for each

Table 14.5. Utility gains over HA

	UF	PC	PF	PCF	PC-GH	PF-GH	PCF-GH
d/p	0.12	0.12	0.12	0.12	0.13	0.54	0.13
e/p	−0.17	−0.17	−0.17	−0.17	−0.11	−0.12	−0.11
se/p	−0.48	−0.48	−0.48	−0.48	−0.41	−0.16	−0.41
b/m	−0.85	−0.85	−0.85	−0.85	−0.80	−0.01	−0.80
roe	−0.47	−0.05	−0.47	0.05	−0.22	−0.44	−0.22
tbl	0.98	0.98	0.98	0.98	0.96	1.03	0.96
lty	0.29	0.29	0.29	0.29	0.30	0.44	0.30
ts	2.76	2.76	2.76	2.76	3.06	2.07	3.06
ds	−0.06	−0.06	−0.06	−0.06	0.17	−0.02	0.17
inf	0.58	0.58	0.58	0.58	0.66	0.74	0.66
nei	3.29	3.29	3.29	3.29	3.31	2.78	3.31
CF	2.16	2.23	1.51	1.60	2.34	1.27	1.68

Notes: The reported numbers are out-of-sample statistics for the utility gain $\left(\hat{U}_1 - \hat{U}_0\right)$ of an investor with mean-variance preferences. It was discussed in CT (2008) and used by Rapach, Strauss, and Zhou (2010). See Section 14.4.2.4 for details. As in Table 14.3, we compare seven forecasts in seven columns using each of the 11 predictors (in 11 rows). The last row (row 12) presents the equally-weighted combined-forecast (CF) of the 11 forecasts in each column.

of the seven conditional forecasts for a given predictor. Table 14.5 then repeats this exercise for each of the 11 predictors considered and for the equally-weighted, combined forecast (CF) and reports the gains in utility $\left(\hat{U}_1 - \hat{U}_0\right)$ from the conditional forecasts over the utility $\hat{U}_0$ of the HA forecast. We set $\gamma = 3$; this value is commonly employed in the literature. The results with $\gamma = 2, 4$ are qualitatively similar. We note the following observations.

1. Positive values of the utility gain $\left(\hat{U}_1 - \hat{U}_0\right)$ indicate that a model is better than HA. While there are negative values for some predictors, the incidence of negative values is lower, however, than in Table 14.3. Even some unrestricted forecasts take positive values[5].
2. The PC and PF constraints do not work well without bagging. Without bagging, only 1 PC out of 11 is better than UF. Without bagging, the PF constraint does not work for all 11 predictors. Without bagging, only 1 PCF out of 11 is better than UF.
3. Bagging works for ~~the~~ both PC and PF constraints: 10 PC-GH out of 11 are better than UF, 9 PF-GHs are better than UF, and 10 PCF-GH out of 11 are better than UF.
4. For CF in row 12, all CFs are positive, beating HA. The PC constraint seems to be more effective than the PF constraint. Bagging improves on CF for PC and PCF forecasts but not for PF. This is similar to Tables 14.3 and 14.4 in MSFE.

In summary, in terms of the utility level, bagging the constraints can improve the forecast. Combined forecasts (CF) outperform HA for all constrained models and bagging further improves their forecast power, as in Tables 14.3 and 14.4.

14.5 Conclusions

The vast literature on equity return prediction has considered a wide array of models and methods. CT (2008) propose restrictions on the regression coefficient or on the return prediction. Their shrinkage approach reduces MSFE by increasing the forecast bias and reducing the forecast error variance.

In this chapter, we apply bagging to reduce the forecast error variance compared to simple constrained estimators at the potential cost of an increase

[5] We note that recent work by McCracken and Valente (2012) can possibly be applied here to assess the statistical significance of the utility gains. They show that the difference in mean-variance utilities can be asymptotically normal despite the fact that the models are nested. Since the asymptotic variance that accounts for the effects of the parameter estimation is often difficult to estimate, they use a bootstrap proposed by Calhoun (2011).

in bias. We review the theory behind bagging, in particular Breiman (1996), Bühlmann and Yu (2002), and Gordon and Hall (2009), and explore the bias-variance trade-off and shrinkage properties of bagging in simulations. We show that for a large variety of signal-to-noise and regressor persistence scenarios, bagging can further improve predictive power as long as the imposed constraint is not completely obvious and far from binding, but, loosely speaking, true enough.

In the stock return prediction problem, we find that the constraint on the sign of the regression coefficient and/or the positivity constraint on the forecast itself improves prediction. Smoothing the hard constraint at zero for the return forecast by bagging over a large set of bootstrap replications, we improve this edge in predictive power, which we measure by the out-of-sample R^2 as in CT (2008), by the utility function of CT (2008) as reported in Rapach and Strauss (2010), and by the adjusted out-of-sample R^2 of Clark and West (2006). In particular, after accounting for the natural MSFE-disadvantage of a regressor model compared to the historical mean under the null, the advantage of bagging constraints becomes very clear. Simple combination forecasts do consistently well, but not always best. In our application, they could also be improved by bagging.

References

Ang, A. and G. Bekaert (2007): "Stock Return Predictability: Is It There?," *Review of Financial Studies* **20**, 651–707.

Audrino, F. and M. C. Medeiros (2011): "Modeling and Forecasting Short-term Interest Rates: The Benefits of Smooth Regimes, Macroeconomic Variables, and Bagging," *Journal of Applied Econometrics* **26**, 999–1022.

Baker, M. and J. Wurgler (2000): "The Equity Share in New Issues and Aggregate Stock Returns," *Journal of Finance* **55**, 2219–2257.

Barberis, N.C. (2000): "Investing for the Long Run When Returns Are Predictable," *Journal of Finance* **55**, 225–264.

Bühlmann, P. and B. Yu (2002): "Analyzing Bagging," *Annals of Statistics* **30**, 927–961.

Bossaerts, P. and P. Hillion (1999): "Implementing Statistical Criteria to Select Return Forecasting Models: What do we Learn?," *Review of Financial Studies* **12**, 405–428.

Breiman, L. (1996): "Bagging Predictors," *Machine Learning* **36**, 105–139.

Buja, A. and W. Stuetzle (2006), "Observations on Bagging", University of Pennsylvania and University of Washington, Seattle.

Butler, A., G. Grullon, and J. Weston (2006): "Can Managers Forecast Aggregate Market Returns?," *Journal of Finance* **60**, 963–986.

Calhoun, G. (2011): "An Asymptotically Normal Out-of-sample Test of Equal Predictive Accuracy for Nested Models", Iowa State University.

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

Campbell, J. and R. Shiller (1988a): "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies* **1**, 195–228.

Campbell, J. and R. Shiller (1988b): "Stock Prices, Earnings, and Expected Dividends," *Journal of Finance* **43**, 661–676.

Campbell, J. and R. Shiller (1998): "Valuation Ratios and the Long-Run Stock Market Outlook," *Journal of Portfolio Management* **24**, 11–26.

Campbell, J. and S. Thompson (2008): "Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?," *Review of Financial Studies* **21**(4), 1511–1531.

Campbell, J. and M. Yogo (2006): "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics* **81**, 27–60.

Cavanagh, C., G. Elliot, and J. Stock (1995): "Inference in Models with Nearly Integrated Regressors," *Econometric Theory* **11**, 1131–1147.

Clark, T. E. and M. W. McCracken (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* **105**: 85–110.

Clark, T. E. and K. D. West (2006): "Using Out-of-sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics* **135**, 155–186.

Diebold, F. X. and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* **13**: 253–263.

Dow, C. H. (1920): "Scientific Stock Speculation," *The Magazine of Wall Street, New York*.

Estrella, A. and G. A. Hardouvelis (1991): "The Term Structure as a Predictor of Real Economic Activity," *Journal of Finance* **46**(2), 555–576.

Fama, E. and K. French (1988): "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics* **22**, 3–25.

Fama, E. and K. French (1989): "Business Conditions and Expected Returns on Stocks and Bonds," *Journal of Financial Economics* **25**, 23–49.

Fama, E. and G. Schwert (1977): "Asset Returns and Inflation," *Journal of Financial Economics* **5**, 115–146.

Fama, E. and G. Schwert (1981): "Stock Returns, Real Activity, Inflation and Money," *American Economic Review* **71**, 545–565.

Faust, J. and J. H. Wright (2009): "Comparing Greenbook and Reduced Form Forecasts Using a Large Real Time Dataset," *Journal of Business and Economic Statistics* **27**(4): 468–479.

Ferson, W., S. Sarkissian, and T. Simin (2003): "Spurious Regressions in Financial Economics?," *Journal of Finance* **58**, 1393–1413.

Fitzenberger, B. (1997): "The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions", *Journal of Econometrics* **82**, 235–287.

Foster, F., T. Smith, and R. E. Whaley (1997): "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R2," *Journal of Finance* **52**, 591–607.

Friedman, J. H. and P. Hall (2007), "On Bagging and Nonlinear Estimation", *Journal of Statistical Planning and Inference* **137**, 669–683.

Giacoinini, R. and H. White (2006): "Test of Conditional Predictive Ability," *Econometrica*, **74**(6), 1545–1578.

Gordon, I. and P. Hall (2009): "Estimating a Parameter When It Is Known that the Parameter Exceeds a Given Value," *Australian and New Zealand Journal of Statistics* **51**(4), 449–460.

Goyal, A. and I. Welch (2008): "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies* **21**(4), 1455–1508.

Hall, P., J. Horowitz, and B.-Y. Jing (1995): "On Blocking Rules for the Bootstrap with Dependent Data," *Biometrika* **82**, 561–574.

Hastie, T., R. Tibshirani, and J. Friedman (2001): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Hillebrand, E. and M. Medeiros (2010): "Forecasting Realized Volatility Models: The Benefits of Bagging and Nonlinear Specifications," *Econometric Reviews* **29**(5), 571–593.

Huang, H. and T.-H. Lee (2010): "To Combine Forecasts or to Combine Information," *Econometric Reviews* 29(5), 534–557. bold

Inoue, A. and L. Kilian (2008): "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation," *Journal of the American Statistical Association* **103**(482), 511–522.

Jansson, M. and M. J. Moreira (2006): "Optimal Inference in Regression Models with Nearly Integrated Regressors," *Econometrica* **74**, 681–715.

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

Kilian, L. (1999): "Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?," *Journal of Applied Econometrics* **14**, 491–510.

Koenker, R. and G. Basset (1978): "Asymptotic Theory of Least Absolute Error Regression," *The Journal of the American Statistical Association* **73**, 618–622.

Lee, T.-H. and Y. Yang (2006): "Bagging Binary and Quantile Predictors for Time Series," *Journal of Econometrics* **135**, 465–497.

Lettau, M. and S. Ludvigson (2001): "Consumption, Aggregate Wealth, and Expected Stock Returns," *Journal of Finance* **56**, 815–849.

Lewellen, J. (2004): "Predicting Returns with Financial Ratios," *Journal of Financial Economics* **74**, 209–235.

Mark, N. (1995): "Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability," *American Economic Review* **85**, 201–218.

McCracken, M. W. and G. Valente (2012): "Testing the Economic Value of Asset Return Predictability", Federal Reserve Bank of St. Louis, WP 2012-049A.

Nelson, C. and M. Kim (1993): "Predictable Stock Returns: The Role of Small Sample Bias," *Journal of Finance* **48**, 641–661.

Pesaran, M. and A. Timmermann (2002): "Market Timing and Return Prediction under Model Instability," *Journal of Empirical Finance* **9**, 495–510.

Pesaran, M. H. and A. Timmermann (1995): "Predictability of stock returns: robustness and economic significance," *Journal of Finance* **50**, 1201–1228.

Polk, C., S. Thompson, and T. Vuolteenaho (2006): "Cross-Sectional Forecasts of the Equity Premium," *Journal of Financial Economics* **81**, 101–141.

Rapach, D. and J. K. Strauss (2010): "Bagging or Combining (or Both)? An Analysis Based on Forecasting U.S. Employment Growth," *Econometric Reviews* **29**(5), 511–533.

Rapach, D., J. 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.

Rozeff, M. (1984): "Dividend Yields are Equity Risk Premiums," *Journal of Portfolio Management* **11**, 68-75.

Schwert, G. (2002): "Anomalies and market efficiency," in *Handbook of the Economics of Finance*, edited by M. Harris, and R. Stulz. North Holland: Amsterdam.

Stambaugh, R. (1999): "Predictive Regressions," *Journal of Financial Economics* **54**, 375-421.

Stock, J. H. and Watson, M. W. (2007): "Why has U.S. Inflation Become Harder to Forecast?" *Journal of Money, Credit, and Banking* **39**(1), 3-33.

Stock, J. H. and Watson, M. W. (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," ~~Harvard University and Princeton University~~. *

Teräsvirta, T. (1994): "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association* **89**, 208-218.

Teräsvirta, T. (2006): "Forecasting Economic Variables with Nonlinear Models", *Handbook of Economic Forecasting*, Volume 1, Chapter 8, edited by G. Elliott, C. W. J. Granger and A. Timmermann, Elsevier, B.V.

Timmermann, A. (2008): "Elusive Return Predictability," *International Journal of Forecasting*, 1-18.

Torous, W., R. Valkanov, and S. Yan (2004): "On Predicting Stock Returns with Nearly Integrated Explanatory Variables," *Journal of Business* **77**, 937-966.

Valkanov, R. (2003): "Long-Horizon Regressions: Theoretical Results and Applications," *Journal of Financial Economics* **68**: 201-232.

Wheelock, D. C. and M. E. Wohar (2009): "Can the Term Spread Predict Output Growth and Recessions? A Survey of the Literature," *Federal Reserve Bank of St. Louis Review* 91(5) Part **1**, 419-440.

West, K. (1996): "Asymptotic Inference about Predictive Ability." *Econometrica* **64**(5): 1067-1084.

Zhu, M. (2013): "Jackknife for Bias Reduction in Predictive Regressions," *Journal of Financial Econometrics* **11**(1): 193-220.

* Journal of Business and Economic Statistics 30(4), 481-493.

Contents

List of Contributors

Marco Aiolfi *Quantitative Management LLC*

Cristina Amado *Aarhus Universty and CREATES, University of Minho and NIPE*

Heather M. Anderson *Monash University*

Laurent A. F. Callot *Free University Amsterdam and CREATES*

Jennifer L. Castle *Oxford University, Institute for New Economic Thinking, Oxford Martin School*

Jin Seo Cho *Yonsei University*

James Davidson *University of Exeter Business School*

Robert Engle *New York University, Stern School of Business*

Andreea G. Halunga *University of Exeter Business School*

David F. Hendry *Oxford University, Institute for New Economic Thinking, Oxford Martin School*

Eric Hillebrand *Aarhus University and CREATES*

Isao Ishida *Osaka University and CSFI*

Katarina Juselius *University of Copenhagen*

Mikael Juselius *Bank for International Settlements*

Anders Bredahl Kock *Aarhus University and CREATES*

Robinson Kruse *Leibniz University Hannover and CREATES*

Helinä Laakkonen *Bank of Finland*

Tae-Hwy Lee *University of California at Riverside*

Helmut Lütkepohl *Freie Universität Berlin and DIW*

Marcelo C. Medeiros *Pontifical Catholic University of Rio de Janeiro*

Eduardo F. Mendes *University of New South Wales*

Marius Rodriguez *Federal Reserve Bank of San Francisco*

Rickard Sandberg *Stockholm School of Economics*

Bård Støve *University of Bergen*

Allan Timmermann *University of California, San Diego, Rady School of Management and CREATES*

Dag Tjøstheim *University of Bergen*

Farshid Vahid *Monash University*

Halbert White *University of San Diego, San Diego*