Title

Complete subset regressions

Permalink

https://escholarship.org/uc/item/1st3n7z7

Journal

JOURNAL OF ECONOMETRICS, 177(2)

ISSN

0304-4076

Authors

Elliott, Graham
Gargano, Antonio
Timmermann, Allan

Publication Date

2013

DOI

10.1016/j.jeconom.2013.04.017

Peer reviewed

Journal of Econometrics 177 (2013) 357–373

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Complete subset regressions[☆]

Graham Elliott[a], Antonio Gargano[b], Allan Timmermann[a,*]

[a] UC San Diego, United States

[b] University of Melbourne, Australia

ARTICLE INFO

Article history:
Available online 17 April 2013

Keywords:
Subset regression
Forecast combination
Shrinkage

ABSTRACT

This paper proposes a new method for combining forecasts based on complete subset regressions. For a given set of potential predictor variables we combine forecasts from all possible linear regression models that keep the number of predictors fixed. We explore how the choice of model complexity, as measured by the number of included predictor variables, can be used to trade off the bias and variance of the forecast errors, generating a setup akin to the efficient frontier known from modern portfolio theory. In an application to predictability of stock returns, we find that combinations of subset regressions can produce more accurate forecasts than conventional approaches based on equal-weighted forecasts (which fail to account for the dimensionality of the underlying models), combinations of univariate forecasts, or forecasts generated by methods such as bagging, ridge regression or Bayesian Model Averaging.

1. Introduction

Methods for controlling estimation error in forecasting problems that involve small sample sizes and many potential predictor variables have been the subject of much recent research.[1] One lesson learned from this literature is that a strategy of including all possible variables is typically too profligate; given the relatively short data samples typically available to estimate the parameters of economic forecasting models, it is important to limit the number of parameters that have to be estimated or in other ways reduce the effect of parameter estimation error. This has led to the preponderance of forecast methods such as shrinkage or ridge regression (Hoerl and Kennard, 1970), model averaging (Bates and Granger, 1969; Raftery et al., 1997), bagging (Breiman, 1996), and the Lasso (Tibshirani, 1996), which accomplish this in different ways.

This paper proposes a new method for combining forecasts based on complete subset regressions. For a given set of potential predictor variables we combine forecasts from all possible linear regression models that keep the number of predictors fixed. For example, with K possible predictors, there are K unique univariate models and $n_{k,K} = K!/((K-k)!k!)$ different k-variate models for $k \leq K$. We refer to the set of models for a fixed value of k as a complete subset and propose to use equal-weighted combinations of the forecasts from all models within these subsets indexed by k. Moreover, we show that an optimal value of k can be determined from the covariance matrix of the potential regressors and so lends itself to being selected recursively in time.

Special cases of subset regression combinations have appeared in the empirical literature. For example, Rapach et al. (2010) consider equal-weighted combinations of all possible univariate equity premium models and find that they produce better forecasts of stock returns than a simple no-predictability model. This corresponds to setting $k = 1$ in our context. Papers such as Aiolfi and Favero (2003) consider equal-weighted combinations of forecasts of stock returns from all possible 2^K models. While their combination scheme is not directly nested by our approach, this can nevertheless be obtained from a combination of the individual subset regression forecasts.

From a theoretical perspective, we show that subset regression combinations are akin to a complex version of shrinkage which, in general, does not reduce to shrinking the Ordinary Least Squares (OLS) estimates coefficient by coefficient. Rather, the adjustment to the coefficients depends on all least squares estimates and is a function of both k, the number of variables included in the model, and K, the total number of potential predictors. Only in the special case where the covariance matrix of the predictors is orthonormal does subset regression reduce to ridge regression or, equivalently, to a Bayes estimator with a specific prior distribution. For this special case we derive the exact degree of shrinkage implied by different values of k and thus formalize how k, the number of parameters in the conditional mean equation, is equivalent to other measures of model complexity that have previously been proposed in the literature.

[☆] We thank the Editor, Herman van Dijk, and two anonymous referees for many constructive and helpful comments.

[*] Correspondence to: UC San Diego, Rady School of Management, 9500 Gilman Drive, La Jolla, CA 92093-0553, United States.
E-mail address: atimmermann@ucsd.edu (A. Timmermann).

[1] See, e.g., Stock and Watson (2006) for a review of the literature.

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.04.017

We also show that the weights implied by subset regression reflect omitted variable bias in a way that can be useful for forecasting. This holds particularly in situations with strongly positively correlated regressors since the subset regression estimates account for the omitted predictors.

An attractive property of the proposed method is that, unlike the ridge estimator and conventional Bayesian estimators, it does not impose the same amount of shrinkage on each coefficient. Unlike model selection methods, it also does not assign binary zero–one weights to the OLS coefficients. Other approaches that apply flexible weighting to individual predictors include bagging (Breiman, 1996) which applies differential shrinkage weights to each coefficient, the adaptive Lasso (Zou, 2006) which applies variable-specific weights to the individual predictors in a data-dependent adaptive manner, the Elastic Net (Zou and Hastie, 2005; Zou and Zhang, 2009) which introduces extra parameters to control the penalty for inclusion of additional variables, and Bayesian methods such as adaptive Monte Carlo (Lamnisos et al., 2012).

To illustrate the subset regression approach empirically we consider, like many previous studies, predictability of US stock returns. In particular, following Rapach et al. (2010), we study quarterly data on US stock returns in an application that has 12 potential predictor variables and so generates subset regressions with $k = 1, 2, \ldots, 12$ predictor variables. We find that subset regression combinations that use $k = 2$, 3, or 4 predictors produce the lowest out-of-sample mean squared error (MSE)-values. Moreover, these subset models generate superior predictive accuracy relative to the equal-weighted average computed across all possible models, a benchmark that is well-known to be difficult to beat, see Clemen (1989). We also find that the value of k in the subset regression approach can be chosen recursively (in pseudo "real time") in such a manner that the approach produces forecasts with lower out-of-sample MSE-values than those produced by recursive versions of Bayesian Model Averaging, ridge regression, Lasso, or bagging.

The outline of the paper is as follows. Section 2 introduces the subset regression approach and characterizes its theoretical properties, Section 3 presents a Monte Carlo simulation study, Section 4 conducts the empirical analysis of US stock returns, while Section 5 concludes.

2. Theoretical results

This section presents the setup for the analysis and derives theoretical results for the proposed complete subset regression method.

2.1. Setup

Suppose we are interested in predicting the univariate (scalar) variable y_{T+1} using a linear regression model based on observing K predictors $x_T \in \mathbb{R}^K$, and a history of data, $\{y_{t+1}, x_t\}_{t=0}^{T-1}$. Let $E[x_t x_t'] = \Sigma_X$ for all t and, without loss of generality, assume that $E[x_t] = 0$ for all t. To focus on regressions that include only a subset of the predictors, define β to be a $K \times 1$ vector with slope coefficients in the rows representing included regressors and zeros in the rows of the excluded variables. Let β_0 be the pseudo true value for β, the population value of the projection of y on X, where $y = (y_1, \ldots, y_T)$ is a $T \times 1$ vector and $X = (x_0, x_1, \ldots, x_{T-1})'$ stacks the x observations into a $T \times K$ matrix. Denote the generalized inverse of a matrix A by A^-. Let S_i be a $K \times K$ matrix with zeros everywhere except for ones in the diagonal cells corresponding to included variables, so that if the $[j, j]$ element of S_i is one, the jth regressor is included, while if this element is zero, the jth regressor is excluded. Sums over i are sums over all permutations of S_i.

We propose an estimation method that uses equal-weighted combinations of forecasts based on all possible models that include a particular subset of the predictor variables. Each subset is defined by the set of regression models that include a fixed (given) number of regressors, $k \leq K$. Specifically, we run the 'short' regression of y_t on a particular subset of the regressors, then average the results across all k dimensional subsets of the regressors to provide an estimator, $\hat{\beta}$, for forecasting, where $k \leq K$. With K regressors in the full model and k regressors chosen for each of the short models, there will be subset regressions to average over. In turn, each regressor gets included a total of $n_{k-1,K-1}$ times.

As an illustration, consider the univariate case, $k = 1$, which has $n_{1,K} = K$ short regressions, each with a single variable. Here all elements of $\hat{\beta}_i$ are zero except for the least squares estimate of y_t on x_{it} in the ith row. The equal-weighted combination of forecasts from the individual models is then

$$\hat{y}_{T+1} = \frac{1}{K}\sum_{i=1}^{K} x_T' \hat{\beta}_i. \tag{1}$$

Following common practice, our analysis assumes quadratic or mean square error (MSE) loss. For any estimator, we have

$$\begin{aligned} &E\left[\left(y_{T+1} - \hat{\beta}_T' x_T\right)^2\right] \\ &= E\left[\left(y_{T+1} - \beta_0' x_T + (\beta_0 - \hat{\beta}_T)' x_T\right)^2\right] \\ &= E\left[\left(\varepsilon_{T+1} + (\beta_0 - \hat{\beta}_T)' x_T\right)^2\right] \\ &= \sigma_\varepsilon^2 \left(1 + T^{-1}\sigma_\varepsilon^{-2} E\left[T(\hat{\beta}_T - \beta_0)' x_T x_T' (\hat{\beta}_T - \beta_0)\right]\right). \end{aligned} \tag{2}$$

Here ε_{T+1} is the residual from the population projection of y_{T+1} on x_T and σ_ε^2 is its variance. We concentrate on the last term since the first term does not depend on $\hat{\beta}$. Hence, we are interested in examining $\sigma_\varepsilon^{-2} E\left[(\hat{\beta}_T - \beta)' x_T x_T' (\hat{\beta}_T - \beta)\right]$.

2.2. Complete subset regressions

Subset regression coefficients can be computed as averages over least squares estimates of the subset regressions. When the covariates are correlated, the individual regressions will be affected by omitted variable bias. However, as we next show, the subset regression estimators are themselves approximately a weighted average of the components of the full regression OLS estimator, $\hat{\beta}_{\text{OLS}}$.

Theorem 1. *Assume that as the sample size gets large $\hat{\beta}_{\text{OLS}} \to^p \beta_0$ for some β_0 and $T^{-1}X'X \to^p \Sigma_X$. Then, for fixed K, the estimator for the complete subset regression, $\hat{\beta}_{k,K}$, can be written as*

$$\hat{\beta}_{k,K} = \Lambda_{k,K} \hat{\beta}_{\text{OLS}} + o_p(1),$$

where

$$\Lambda_{k,K} \equiv \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} \left(S_i' \Sigma_X S_i\right)^- (S_i' \Sigma_X).$$

A proof of this result is contained in the Appendix.

This result on the relationship between $\hat{\beta}_{k,K}$ and the OLS estimator makes use of high level assumptions that hold under very general conditions on the data; see White (2001, Chapter 3) for a set of sufficient conditions. Effectively, any assumptions on the model that result in the OLS estimators being consistent for their population values and asymptotically normal will suffice. For

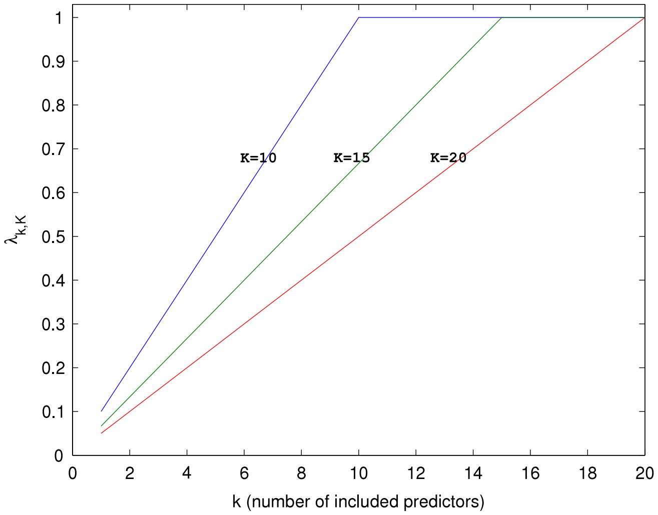

Fig. 1. Degree of shrinkage measured by $\lambda_{k,K} = 1 - \frac{n_{k,K-1}}{n_{k,K}}$ as a function of k (the number of included predictors) for three values of K (the number of possible predictors) assuming a diagonal covariance matrix.

example, the result allows $\{X_t\}$ to be dependent, mixing with a sufficiently small mixing coefficient, and even allows $E[X_t'X_t]$ to be heterogeneous over time, in which case Σ_X is the average variance covariance matrix, although, for simplicity, we assume that Σ_X is constant over time. Ruled out are unit roots in the x- variables, although predictor variables are routinely transformed to be stationary in forecast experiments.

In general, $\Lambda_{k,K}$ is not diagonal and hence the coefficients $\hat{\beta}_{k,K}$ are not (approximately) simple OLS coefficient-by-coefficient shrinkages. Rather, the subset regression coefficients are functions of all the OLS coefficients in the regression. Insight into how the method works as a shrinkage estimator can be gained from the special case when the covariates are orthonormal.[2] In this case, $\hat{\beta}_{k,K} = \lambda_{k,K}\hat{\beta}_{\text{OLS}}$, where $\lambda_{k,K} = 1 - (n_{k,K-1}/n_{k,K})$ is a scalar and so subset regression is numerically equal to ridge regression.[3]

To see this, note that for this special case $\hat{\beta}_{\text{OLS}} = X'y$ while each of the subset regression estimates can be written $\hat{\beta}_i = S_iX'y$. The complete subset regression estimator is then given by

$$\begin{aligned}\hat{\beta}_{k,K} &= \frac{1}{n_{k,K}}\sum_{i=1}^{n_{k,K}}\hat{\beta}_i\\ &= \frac{1}{n_{k,K}}\sum_{i=1}^{n_{k,K}}S_iX'y\\ &= \left(\frac{1}{n_{k,K}}\sum_{i=1}^{n_{k,K}}S_i\right)\hat{\beta}_{\text{OLS}}.\end{aligned}$$

The result now follows by noting that the elements of $\sum_{i=1}^{n_{k,K}}S_i$ are zero for the off-diagonal terms, and equal the number of times the regressor is included in the subset regressions for the diagonal terms. In turn the diagonal terms equal $n_{k,K}$ minus the number of times a regressor is excluded, which gives the result, noting that the solution is the same for each diagonal.

Several points follow from this result. First, the amount of shrinkage implied by $\lambda_{k,K}$ is a function of both k and K. As an illustration, Fig. 1 plots $\lambda_{k,K}$ as a function of k for the orthonormal case. Higher curves represent smaller values of K, where $K = \{10, 15, 20\}$. For any value of K, $\lambda_{k,K}$ is a linear function of k that increases to one. In fact, setting $k = K$ corresponds to simply running OLS with all variables included. Further, as K increases, the slope of the $\lambda_{k,K}$ line gets reduced, so the amount of shrinkage is decreasing for any k, the larger is K, the total number of potential predictors. Essentially, the smaller k is relative to K, the greater the amount of shrinkage. Effectively, the result relates shrinkage provided by model averaging to shrinkage on the coefficients, whereas a typical Bayesian approach would separate the two.

Second, in general $\Lambda_{k,K}$ reduces to the ridge estimator, either approximately or exactly, only when the regressors are uncorrelated. When this does not hold, subset regression coefficients will not be simple regressor-by-regressor shrinkages of the OLS estimates, and instead depend on the full covariance matrix of all regressors. Specifically, $\Lambda_{k,K}$ is not diagonal and each element of $\hat{\beta}$ is approximately a weighted sum of all of the elements in $\hat{\beta}_{\text{OLS}}$. The weights depend not only on $\{k, K\}$ but on all elements in Σ_X, denoted Σ_{ij}. For example, if $K = 3$ and $k = 1$, we have

$$\Lambda_{1,3} = \frac{1}{3}\begin{pmatrix} 1 & \dfrac{\Sigma_{12}}{\Sigma_{11}} & \dfrac{\Sigma_{13}}{\Sigma_{11}} \\ \dfrac{\Sigma_{12}}{\Sigma_{22}} & 1 & \dfrac{\Sigma_{23}}{\Sigma_{22}} \\ \dfrac{\Sigma_{13}}{\Sigma_{33}} & \dfrac{\Sigma_{23}}{\Sigma_{33}} & 1 \end{pmatrix}. \tag{3}$$

Each row of $\Lambda_{1,3}$ is the result of including a particular subset regression in the average. For example, the first row gives the first element of $\hat{\beta}_{1,3}$ as a weighted sum of the OLS regressors $\hat{\beta}_{\text{OLS}}$. Apart from the multiplication by 1/3, its own coefficient is given a relative weight of one while the remaining coefficients are those we expect from omitted variable bias formulas. The effect of dividing by $n_{1,3} = 3$ is to shrink all coefficients, including its own coefficient, towards zero.

For $k > 1$, each regressor gets included more often in the regressions. This increases their effect on $\Lambda_{k,K}$ through a higher inclusion frequency, but decreases their effect through the omitted variable bias. Since the direct effect is larger than the omitted variable bias, an increased k generally reduces the amount of shrinkage. Of course, in the limit as $k = K$, there is no shrinkage and the method is identical to OLS.

While we focus on one-period forecasts in our analysis, the results readily go through for arbitrary horizons provided that the direct approach to forecasting is used, i.e., current values of y are projected on h-period lagged values of the predictors. Conversely, the iterated approach to forecasting requires modeling a VAR comprising both y and all x-variables and so is more involved.

2.3. Risk

We next examine the risk of the subset regression estimator. Forecasting is an estimation problem and risk is the expected loss as a function of the true (but unknown) model parameters. Under MSE loss, risk amounts to the expected loss. In common with all biased methods, for values of β_0 far from zero, the risk is large and so it is appropriate not to shrink coefficients towards zero. Shrinkage methods only add value when β_0 is near zero. To capture such a situation, we assume that β_0 is local to zero. Specifically, we assume that $\beta_0 = T^{-1/2}\sigma_\varepsilon b$ for some fixed vector b.

Under general, dependent data generating processes, the risk is difficult to derive. However, if we restrict the setup to i.i.d. data $\{y_{t+1}, x_t\}$, we get the following result.

[2] We refer to subset regressions as similar to shrinkage although for some configurations of the variance covariance matrix of the predictors and some OLS estimates, subset regression will not actually shrink the coefficient estimates.

[3] Equivalently, this case corresponds to a Bayes estimator under normality with prior $N(\mu, \gamma_{k,K}^{-1}\sigma_\varepsilon^2)$, $\hat{\beta} = (X'X + \gamma_{k,K}I)^{-1}(X'y + \gamma_{k,K}\mu)$, prior mean $\mu = 0$, and $\gamma_{k,K} = (1-\lambda_{k,K})/\lambda_{k,K}$. If the assumption on the regressors is weakened to $\Sigma_X = I_K$, the same result holds asymptotically.

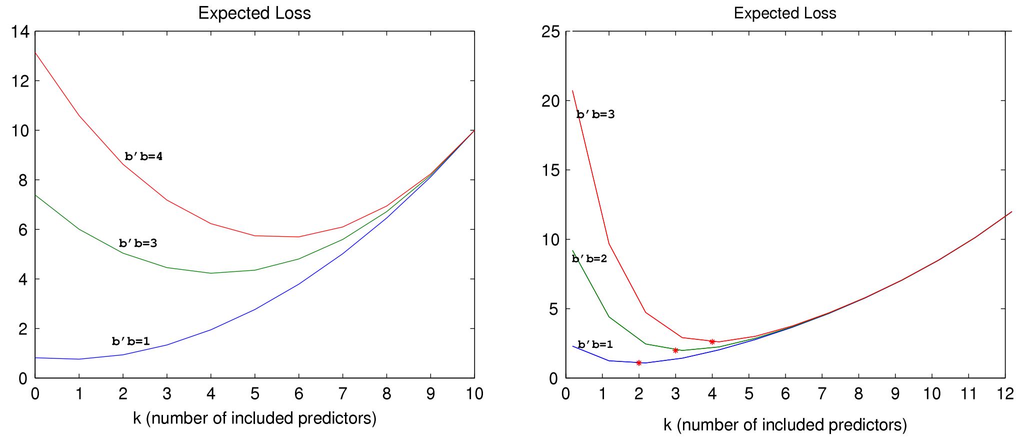

Fig. 2. Expected loss (risk) for different values of the local-to-zero parameters (b) for $\Sigma = I$ (left) and $\Sigma = \hat{\Sigma}_X$ (right), assuming $K = 10$ possible predictor variables.

Theorem 2. *Assume that the data* $\{y_{t+1}, x_t\}$ *are i.i.d.,* $\beta_0 = T^{-1/2}\sigma_\varepsilon b$, $E[(\hat{\beta} - \beta_0)^2 | x_{T+1}] = E[(\hat{\beta} - \beta_0)^2]$, *and* $T^{1/2}(\hat{\beta}_{\mathrm{OLS}} - \beta) \to^d N(0, \Sigma_X^{-1})$. *Then, in large samples,*

$$\sigma_\varepsilon^{-2} E\left[T(\hat{\beta}_T - \beta)'\Sigma_X(\hat{\beta}_T - \beta)\right]$$

$$\approx \sum_{j=1}^{K} \zeta_j + b'(\Lambda_{k,K} - I)'\Sigma_X(\Lambda_{k,K} - I)b, \qquad (4)$$

where ζ_j *are the eigenvalues of* $\Lambda'_{k,K}\Sigma_X\Lambda_{k,K}\Sigma_X^{-1}$.

The expected loss depends on many aspects of the problem. First, it is a function of the variance covariance matrix through both Σ_X and $\Lambda_{k,K}$. Second, it depends on the dimension of the problem, K, and of the subset regression, k. Third, it is a function of the elements of b. Different trade-offs can be explored by varying these parameters. Some will be very attractive when compared to OLS, while others might not be. As in the simple orthogonal case, the larger the elements of b are, the worse the complete subset regression methods will perform.

For different choices of $\{K, k, \Sigma_X, b\}$, we can compute the expected loss frontier as a function of k. If $\Sigma_X = I$, so the regressors are mutually orthogonal, (4) reduces to

$$\sigma_\varepsilon^{-2} E\left[(\hat{\beta}_T - \beta)'\Sigma_X(\hat{\beta}_T - \beta)\right] = \lambda_{k,K}^2 K + (1 - \lambda_{k,K})^2 b'b, \qquad (5)$$

which depends on $\{K, k, b'b\}$. For fixed values of $b'b$ and K, as k increases, $\lambda_{k,K}$ gets larger and the increase in the first term in (5) is offset by the decrease in the second term in this equation. The extent of this offset depends on the relative sizes of K and $b'b$. As an illustration of this, the left window in Fig. 2 plots the value of the expected loss (5) as a function of k, for $K = 10$ and $b'b = (1, 3, 4)$. Each line corresponds to a separate value of $b'b$ with larger intercept on the x axis, the greater $b'b$ is. Setting $k = K = 10$ yields OLS loss, so all lines converge at that point. A variety of shapes are possible. If $b'b$ is quite small, so that the regressors are not that useful for forecasting, then a large amount of shrinkage, and hence a small value of k, works best. Conversely, if $b'b$ is large, larger values of k become optimal.

In practice, different choices of k can be motivated by theoretical considerations. As always with shrinkage estimation, the smaller b is expected to be, the more useful it is to apply strong shrinkage. As we discuss above, the amount of shrinkage tends to be greater, the smaller one chooses k. Since $\{k, K\}$ are known and Σ_X can be estimated by $T^{-1}X'X$, (4) can be used to produce curves such as those in the left window of Fig. 2 but relevant for the application at hand. One can then choose k as the point at which expected loss is lowest given reasonable choices for b. As an illustration of this point, the right window of Fig. 2 uses data from the application in Section 4 to estimate Σ_X and shows expected loss curves for $b'b = 1, 2,$ or 3. Although the expected loss curve varies quite a bit across different values of $b'b$, an interior optimal value – corresponding to a minimal expected loss – around $k = 2, 3,$ or 4 is obtained in all three cases.

2.4. Comparison with OLS and ridge

It is informative to compare the risk for subset regressions to that of models estimated by OLS. In some cases, this comparison can be done analytically. For example, this can be done for general K when Σ_X has ones on the diagonal and ρ elsewhere and $k = 1$, corresponding to combinations of univariate models. First, note that when $b = 1_K$, a K-vector of ones, the risk for OLS regression is K while for this case the risk of the subset regression method reduces to

$$E[(y_{T+1} - \hat{\beta}'_{1,K}x_T)^2] = \frac{1}{K}\left(1 + (K-1)\rho^2\right)$$

$$+ (\rho - 1)^2\left(\frac{K-1}{K}\right)^2 (K + K(K-1)\rho). \qquad (6)$$

Hence, subset regressions produce lower risk than OLS for any (K, ρ) pair for which

$$\frac{1}{K}\left(1 + (K-1)\rho^2\right) + (\rho - 1)^2\left(\frac{K-1}{K}\right)^2$$

$$\times\ (K + K(K-1)\rho) < K.$$

For small values of K this holds for nearly all possible correlations. To illustrate this, Fig. 3 plots the ratio of the subset regression MSE to the OLS MSE as a function of ρ, the correlation between the predictors, and k, the number of predictors included. The figure assumes that $T = 100$. Whenever the plotted value falls below one, the subset regression approach dominates OLS regression in the sense that it produces lower risk. For any $K \le 6$, subset regression always (for any ρ for which Σ_X is positive definite) has a lower risk than OLS based on the complete set of regressors. For $K > 6$, we find that there is a small region with small values of ρ and $k = 1$ for which the reverse is true, but otherwise subset regression continues to perform better than OLS.

The figure thus illustrates that a simple equal-weighted average of univariate forecasts can produce better forecasts than the conventional multivariate model that includes all predictors, even in situations where the univariate models are misspecified due to omitted variable bias.

It is also of interest to compare the subset regression approach to methods such as ridge regression which apply the same amount

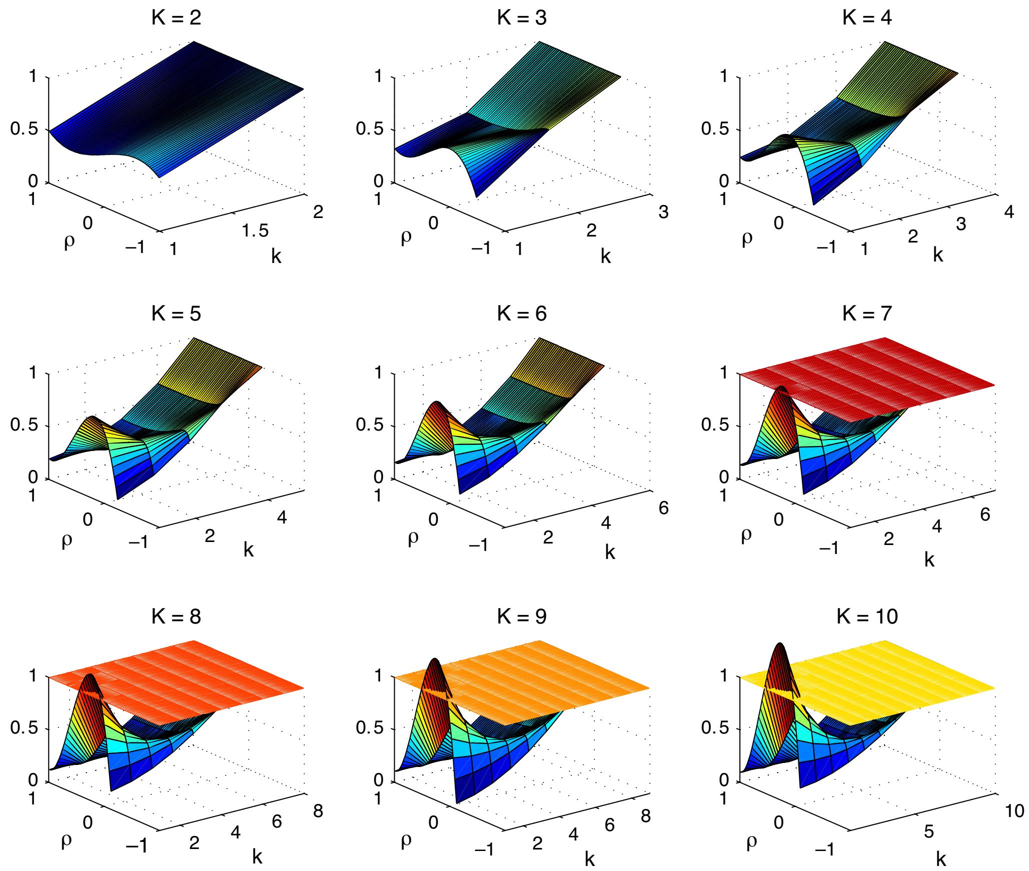

Fig. 3. Relative performance of OLS versus subset regression. The figure shows the MSE loss under subset regression relative to OLS (i.e. $\text{MSE}^{\text{subset}}/\text{MSE}^{\text{OLS}}$) as a function of ρ, the correlation between predictors, k, the number of included predictors, and K, the total number of predictors. Values below unity mean that the subset regression risk is lower than the OLS risk, whereas values above unity imply that OLS is better.

of shrinkage to each regression coefficient. To this end, Fig. 4 uses heat maps to compare the expected loss of the subset regressions to that of the ridge regression approach for different values of the limit of the shrinkage parameter, γ/T. The figure assumes that there are $K = 8$ predictor variables, sets $b = 1_K$, a vector of ones, and lets Σ_X have ones on the diagonal and ρ on all off-diagonal cells. The correlation between predictor variables, ρ, varies along the horizontal axis, while the shrinkage parameter, γ, varies along the vertical axis. We use colors to indicate the value for $\min(0, \text{MSE}^{\text{ridge}} - \text{MSE}^{\text{subset}})$, with dark red indicating that $\text{MSE}^{\text{ridge}} > \text{MSE}^{\text{subset}}$, while, conversely, yellow and blue indicate areas where $\text{MSE}^{\text{ridge}} < \text{MSE}^{\text{subset}}$. Each window corresponds to a different value of k. Suppose that, moving along the vertical axis corresponding to a particular value of ρ, there is no red color. This shows that, for this particular value of ρ, ridge regressions always produce a lower expected loss than the corresponding subset regressions. Conversely, if, for a given value of ρ, the area is red for all values of γ, subset regressions dominate all ridge regressions, regardless of the chosen shrinkage parameter.

Fig. 4 shows that when $k = 1$, ridge regressions mostly produce lower MSE-values than subset regressions for $\rho < 0.6$. Conversely, for $\rho > 0.85$, the univariate subset regressions uniformly dominate all ridge results. If $k = 2$, subset regressions uniformly dominate when $\rho > 0.6$, while if $k = 4$, subset regressions always dominate when $\rho < 0.5$.

2.5. Discussion

The method presented above, along with the analytical results, relies on the total number of regressors, K, being somewhat smaller than T, the number of observations available. This necessarily limits the possible values for K, given that for many applications, especially in macroeconomics, T is not particularly large. Model instabilities may further exacerbate this concern since they could limit the amount of past data available for model combination. In such situations, using an equal-weighted average forecast can provide robust out-of-sample predictive performance and so helps to motivate our approach of not using estimated combination weights. Moreover, empirical work has largely failed to come up with alternative weighting schemes that systematically beat equal-weighting, so we find the simplicity of this weighting scheme attractive. However, it is of interest to consider extensions to very large values of K or to alternative weighting schemes. We next discuss these issues.

2.5.1. Computational issues

In cases where K is very large and so $n_{k,K}$ is too large to allow all models in a given subset to be considered, one can employ fewer than all possible models in each subset. Specifically, if $n_{k,K}$ is very large, one can randomly draw a smaller number of models and average across these. Uniform probability weighting of the models within each subset is the easiest approach and is natural to consider here since we use equal weights in the model averaging stage.

Alternatively, the probability that a model is included could depend on the properties of that model, an approach that will be computationally costlier since it requires evaluation of the models. Methods exist that employ some of the model information to decide on inclusion without requiring statistics for all models to be computed. MCMC algorithms developed in the Bayesian model combination and selection literature can be used, particularly

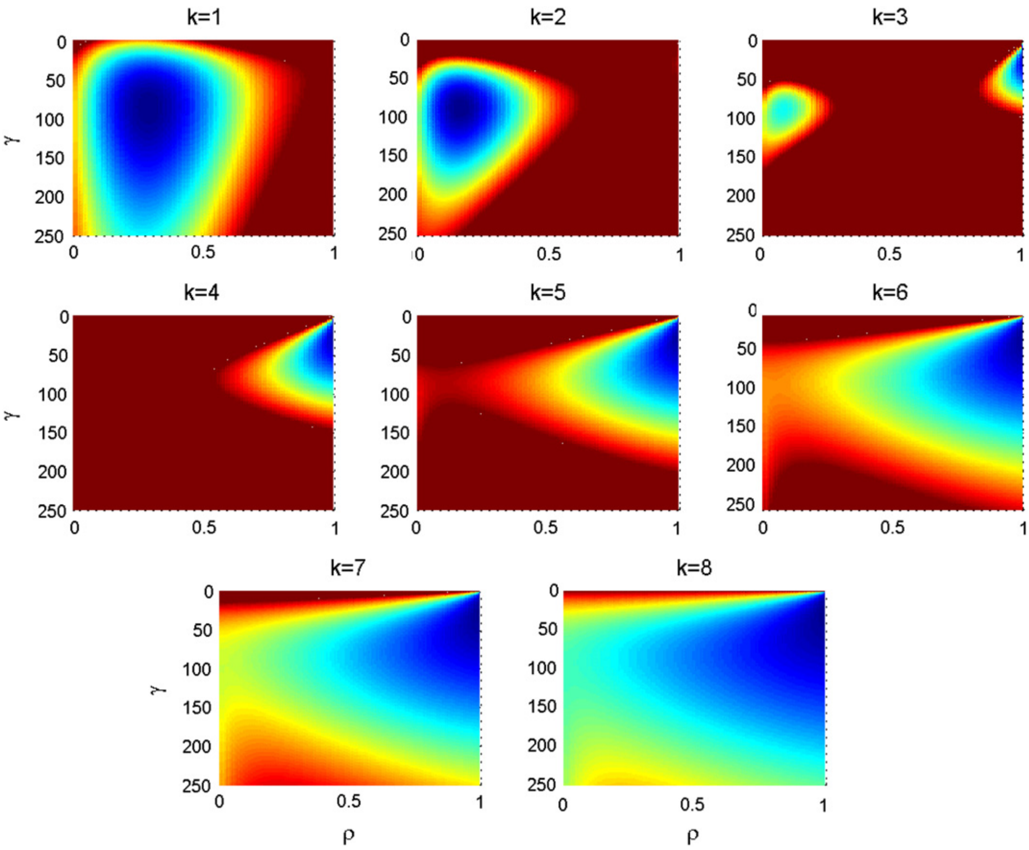

Fig. 4. Relative performance of ridge versus subset regression. This figure shows $\min(0, \text{MSE}^{\text{ridge}}/\text{MSE}^{\text{subset}})$ as a function of ρ, the correlation between the predictor variables on the x-axis, and γ, the shrinkage parameter used by the ridge approach on the y-axis. Dark red color shows areas where the subset regression produces a lower MSE than the ridge approach, while yellow and blue colors indicate areas where the subset approach produces the highest MSE values. Each box corresponds to a different value of k, the number of predictors included in the forecast model. The graph assumes that b is a vector of ones and $K = 8$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

if it is desired that model weights should reflect posteriors. A possibility along these lines that allows K to become very large is the transdimensional Markov chains that simultaneously cover both the parameter and model spaces. More generally, reversible jump MCMC techniques (reviewed by Sisson, 2005) or stochastic search algorithms such as the shotgun stochastic search algorithm of Hans et al. (2007) can be adopted.

2.5.2. Weighting schemes

Our approach uses equal-weighted combinations of forecasts within each subset. However, alternative weighting schemes could be used and we will also consider approximate Bayesian Model Averaging (BMA) weights that are based on the individual models' values of the Schwarz information criterion (SIC). In keeping with the empirical evidence on optimal MSE weights, we do not attempt to use Bates and Granger (1969) weights; the large literature on forecast combination under MSE loss does not suggest methods that we expect to work better than equal weights.

Outside of minimizing the risk criterion considered here, there exist other combination methods that rely on alternative characterizations of risk. Liang et al. (2011) consider linear models with serially independent homoskedastic normal errors and estimate combination weights through a procedure designed to minimize the trace of the MSE of the parameter vector estimates. Note that this objective is different from minimizing the forecast error MSE which weights the sampling error of the parameter vector differently from that invoked by the trace.

The optimal prediction pool approach of Geweke and Amisano (2011) combines models so as to maximize the log predictive score. This requires computation of the density for each model and not just an estimate of the conditional mean. Although this approach has many theoretically appealing properties and does not need to assume that the true model is included in the set over which the model search is conducted, it is unclear how well it would work in settings that combine a very large set of models.

2.5.3. Model instability

Economic time series often undergo change. As a consequence, the parameters of the underlying forecast models may be subject to change and the best forecast model could also change over time. To deal with this, Groen et al. (2013) consider an approach that accounts for breaks in the individual models' parameters as well as breaks in the error variance of the overall combination. Similarly, Billio et al. (2012) propose a variety of combination strategies that allow for time-varying weights, and Koop and Korobilis (2012) consider dynamic model averaging methods. While model instability is ignored here, it can be partially incorporated either by explicitly modeling the break process or by using ad hoc approaches such as rolling-window estimators.

3. Monte Carlo simulation

To better understand how the subset combination approach works, we first consider a Monte Carlo simulation experiment that allows us to study both the absolute forecast performance of the subset regression approach as well as its performance relative to alternative methods.

3.1. Simulation setup

Our Monte Carlo design assumes a simple linear regression model:

$$Y_{t+1} = \sum_{k=1}^{K} \beta_k x_{kt} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2). \tag{7}$$

We assume a sample size of $T = 100$ observations and consider one-step-ahead forecasts of Y_{T+1}. The covariance matrix of the X-variables $\Sigma_X = \text{Cov}(X_1, \ldots, X_K)$ takes the simple form

$$\Sigma_X = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \ddots & & \vdots \\ & & \ddots & & \\ \vdots & & \ddots & 1 & \rho \\ \rho & & \cdots & \rho & 1 \end{pmatrix},$$

where $\rho \in \{0, 0.25, 0.5, 0.75, 0.95\}$. Small values of ρ correspond to small values of the predictive R^2, while the R^2 increases as ρ is raised. Data are assumed to be i.i.d., and we include up to $K = 8$ predictors. Two designs are considered for the regression parameter: $b = 1_K$ and $b = (1\,1\,1\,1\,0\,0\,0\,0)$. In the first experiment, all predictors are relevant and matter equally; in the second experiment only the first four predictors matter to the outcome.

3.2. Comparison with other approaches

We are interested not only in how well the subset combination approach performs in absolute terms, but also in how it compares with other approaches. Many alternative ways to combine or shrink forecasts from different models have been considered in the literature. Among the most prominent ones are ridge regression (Hoerl and Kennard, 1970), the Lasso (Tibshirani, 1996), the Elastic Net (Zou, 2006), bagging (Breiman, 1996), and Bayesian Model Averaging (Raftery et al., 1997). Given the availability of these alternatives, it becomes important to compare the subset regression combination approach to such methods. We briefly discuss each of the methods and explain how we implement them.

3.2.1. Ridge regression

The only parameter that has to be chosen under the ridge approach is γ, which regulates the amount of shrinkage imposed on the regression coefficients. Given a value of γ, the forecasts are given by

$$\hat{y}^{\text{RIDGE}}_{T+1|T} = x_T' \hat{\beta}_\gamma, \tag{8}$$

where

$$\hat{\beta}_\gamma = \underset{\beta}{\text{argmin}} \left(\sum_{t=1}^{T-1} (y_{t+1} - x_t'\beta)^2 + \gamma \sum_{j=1}^{K} \beta_j^2 \right). \tag{9}$$

Note that, by construction, as $\gamma \to \infty$, $\hat{y}^{\text{RIDGE}}_{T+1} \to \frac{1}{T-1} \sum_{j=2}^{T} y_j$, so the ridge forecast simply converges to the sample mean. Following Inoue and Kilian (2008), we consider a range of shrinkage values $\gamma \in \{0.5, 1, 2, 3, 4, 5, 10, 20, 50, 100, 150, 200\}$.

3.2.2. Lasso

Least absolute shrinkage and selection operator, Lasso (Tibshirani, 1996), retains the features of both model selection and ridge regression: it shrinks some coefficients and sets others to zero. Lasso forecasts are computed as

$$\hat{y}^{\text{LASSO}}_{T+1|T} = x_T' \hat{\beta}_\psi, \tag{10}$$

where

$$\hat{\beta}_\psi = \underset{\beta}{\text{argmin}} \left(\sum_{t=1}^{T-1} (y_{t+1} - x_t'\beta)^2 \right), \tag{11}$$

$$\text{s.t.} \sum_{j=1}^{K} |\beta_j| \leq \psi.$$

Here the parameter ψ controls for the amount of shrinkage. For sufficiently large values of ψ the constraint is not binding and the Lasso estimator reduces to OLS. Given the absolute value operator $|\cdot|$, the constraint is not linear and a closed form solution is not available. $\hat{\beta}_\psi$ is therefore computed following the algorithm described in Section 6 of Tibshirani (1996). Because the forecasts depend on ψ, we consider a grid of values $\psi \in \{1, 2, 3, 4, 5, 10, 20, 50, 100\}$.

3.2.3. Elastic Net

Various authors have recently proposed flexible generalizations of the Lasso such as the adaptive Lasso of Zou (2006) or the Elastic Net of Zou and Hastie (2005) and Zou and Zhang (2009). We focus on the Elastic Net, which is a useful compromise between ridge and Lasso and has previously been used in economic studies such as Korobilis (2013). Ridge regressions shrink the coefficients of correlated predictors towards each other. Conversely, Lasso is indifferent to very correlated predictors and tends to simply pick one and ignore the rest. Elastic Net forecasts avoid these extreme solutions and are computed as

$$\hat{y}^{\text{NET}}_{T+1|T} = x_T' \hat{\beta}_{\alpha,\psi}, \tag{12}$$

where

$$\hat{\beta}_{\alpha,\psi} = \underset{\beta}{\text{argmin}} \left(\sum_{t=1}^{T-1} (y_{t+1} - x_t'\beta)^2 + \psi \left\{ \sum_{j=1}^{K} (1-\alpha)\beta_j^2 + \alpha|\beta_j| \right\} \right).$$

The Elastic Net penalty term includes as special cases the ridge penalty ($\alpha = 0$) and the Lasso penalty ($\alpha = 1$). $\hat{\beta}_{\alpha,\psi}$ is computed using the coordinate descent algorithm developed in Friedman and Tibshirani (2010). We set $\alpha = 0.5$, while for ψ we consider a grid of values $\psi \in \{1, 2, 3, 4, 5, 10, 20, 50, 100\}$.

3.2.4. Bagging

Our implementation of bagging is based on 1000 bootstrapped samples of the original data arranged in the $\{y_{t+1:T}, X_{t:T-1}\}$ tuple. We preserve the autocorrelation structure of the predictors by applying the circular block bootstrap of Politis and Romano (1992) with block size chosen optimally according to Politis and White (2004).[4] Contemporaneous dependence across observations is preserved by using the same blocks for all variables. For each bootstrapped sample $\{y^b_{t+1:T}, X^b_{t:T-1}\}$, an estimate of β, $\hat{\beta}^b$ is obtained and forecasts are computed as

$$\hat{y}^b_{T+1|T} = (x_T' S_T)\hat{\beta}^b. \tag{13}$$

Here S_T is the stochastic selection matrix whose (i, i) elements equal the indicator function $I(|t_i| > c)$. A predictor is added only if its t-statistic is significant at the threshold implied by c. The larger

[4] To ensure robustness, we also implemented the parametric bootstrap, but found that the results are not sensitive to this choice.

the value of c, the higher the threshold and so the more parsimonious the final model will be. To control for this effect, we follow Inoue and Kilian and consider different values $c \in \{0.3853, 0.6745, 1.2816, 1.4395, 1.6449, 1.9600, 2.2414, 2.5758, 2.8070, 3.0233, 3.2905, 3.4808, 3.8906, 4.4172, 5.3267\}$. The final bagging forecasts are obtained by averaging across the bootstrap draws

$$\hat{y}^{\text{BAGG}}_{T+1|T} = \frac{1}{B}\sum_{b=1}^{B} \hat{y}^{b}_{T+1|T}. \tag{14}$$

3.2.5. Bayesian Model Averaging

Bayesian Model Averaging predictions are obtained by weighting each model's forecast by its posterior probability:

$$\hat{y}^{\text{BMA}}_{T+1|T} = \sum_{j=1}^{2^K} \hat{y}_j p(M_j|y_{1:T}), \tag{15}$$

where $\hat{y}_j$ is the posterior mean and $p(M_j|y_{1:T})$ is the posterior probability of the jth model, which follows from Bayes' theorem

$$p(M_j|y_{1:T}) = \frac{f(y_{1:T}|M_j)p(M_j)}{\sum_{j=1}^{2^K} f(y_{1:T}|M_j)p(M_j)}. \tag{16}$$

To obtain the predictive likelihood, $f(y_{T+1}|y_T, M_j)$, the marginal likelihood, $f(y_{1:T}|M_j)$, and the model priors, $p(M_j)$, in Eqs. (15) and (16), we follow the specification suggested by Fernandez et al. (2001a,b) and Ley and Steel (2009). Let γ_i be an indicator variable which takes a value of one if the predictor is included in the regression and is zero otherwise. Let θ be the probability of inclusion, so the prior probability of the jth model is $P(M_j) = \theta^{k_j}(1-\theta)^{K-k_j}$, where k_j is the number of predictors in the jth model. A prior for θ is obtained indirectly through a prior on the model size, $m_k = \sum_{i=1}^{K} \gamma_i$. If θ is kept fixed, m_k has a Binomial(K, θ) distribution with expected value $E[m_k] = m = \theta K$, from which it follows that $\theta = m/K$.[5]

As in Ley and Steel (2009), we also allow θ to be random and follow a beta distribution with shape parameters $s_1 = 1$ and s_2. Ley and Steel (2009) show that under this specification, k will follow a binomial-beta distribution. As in the fixed θ scenario, a prior on s_2 is obtained indirectly by solving the equation for the expected model size, $s_2 = (K - m)/m$.

The marginal and predictive likelihoods have closed form expressions only when using conjugate priors. We follow Fernandez et al. (2001a), and adopt a combination of a "non-informative" improper prior on the common intercept α and scale σ_ε and a g-prior (Zellner, 1986) on the regression coefficients β_j, leading to the prior density $p(\alpha, \beta_j, \sigma_\varepsilon|M_j) \propto \sigma_\varepsilon^{-1} f_N^q(\beta|0, \sigma_\varepsilon^2(gZ_j'Z_j)^{-1})$, where Z_j are the demeaned regressors that are included in the jth model. Under this specification $y_{T+1}|y_T, M_j$ follows a t-distribution with location parameter $\hat{y}_j = \frac{1}{T}\sum_{i=1}^{T} y_i + x_j'\beta_j/(g+1)$.

To sum up, we need to specify a value for the prior model size, m, and the g-prior. In the empirical exercise we set m equal to 0.1 and 1 to keep the models from including too many predictors which we know is likely to hurt the performance of the return forecasts, see, e.g., Goyal and Welch (2008). In the Monte Carlo simulations we set m to one half and one third of K. We follow Fernandez et al. (2001a) and set g to $1/T$ or $1/K^2$. In the empirical exercise we add $g = 1$ to ensure stronger shrinkage since, as $g \to \infty$, $\hat{b}_j$ converges to the prevailing mean.

3.3. Simulation results

Table 1 shows results from the simulation experiment, using 25,000 simulations and $T = 100$. We report performance in terms of the R^2-value, which is inversely related to the MSE-value, but conveys the same message and is slightly easier to interpret. First, consider the performance of the subset regression approach when $b = 1_K$ (left panel). Since the R^2 is positive for the (infeasible) model that uses the correct parameter values, negative R^2-values show that parameter estimation error dominates whatever evidence of predictability the model identifies. This case only occurs for the subset regressions when $\rho = 0$ and $k = 8$, corresponding to the "kitchen sink" approach that includes all predictors and so does not average across multiple models. For small values of ρ the best subset regressions use three or four predictors. As ρ increases, the number of variables included in the best-performing subset regressions tends to decrease and the best performance is obtained for $k = 1$ or $k = 2$. In general, the difference between the best and worst subset combinations (usually the kitchen sink, $k = 8$) tends to be greater, the smaller the value of ρ. This is likely to reflect the greater importance of estimation error in situations where the predictive signal is weaker, parameter estimation error matters more and affects the larger models (large k) more than the smaller models (small k).

The ridge regression results most closely resemble those from the subset regressions. Compared with subset regression, ridge regression performs quite well, although, consistent with Fig. 4, the best subset regression produces better performance than the best ridge regression in all cases. In turn, the best subset and ridge regressions generally perform better than the best Lasso, bagging and BMA approaches.

Similar conclusions emerge when we set $b = (1\,1\,1\,1\,0\,0\,0\,0)'$, the results for which are shown in the right panel of Table 1. This case represents a setup with a smaller degree of predictability over the outcome variable, and so lower R^2-values are obtained. Unsurprisingly, for this case the best subset regressions use a smaller value of k than in the previous case where all predictors had an effect on the outcome. The subset regressions that include relatively few predictors, e.g., $k = 2$, 3, or 4, continue to perform particularly well, whereas performance clearly deteriorates for the models that include more predictors.

3.4. Subset combinations with large K

Computing forecasts for all models within a given subset is not feasible when $n_{k,K}$ is large. To explore the consequence of this limitation, we next use simulations that evaluate some of the alternative solutions discussed in Section 2.5.1. First, we set $K = 15$, a number small enough that we can use the complete subset method for all values of $k \leq 15$. We report the outcome of three alternative approaches that combine forecasts over (i) randomly selected models; models selected by stochastic search using either (ii) a Markov chain or (iii) the shotgun approach of Hans et al. (2007). The Markov chain and shotgun approaches differ in how they explore the model space.

The simulations were implemented as follows. Let $c \leq n_{k,K}$ be the number of included models, while $\alpha \in (0, 1)$ is the fraction of the $n_{k,K}$ models that is combined so $c = \alpha \times n_{k,K}$. Also define $\bar{c}$ and $\underline{c}$ as upper and lower bounds on c so that if $\alpha \times n_{k,K} > \bar{c}$, only $\bar{c}$ models are combined while if $\alpha \times n_{k,K} \leq \underline{c}$, we set $c = n_{k,K}$. Our simulations set $\alpha = 0.25$, $\underline{c} = 100$, and $\bar{c} = 5000$.

Under the random approach c models are drawn without replacement from the model space $M_k = [m_1, m_2, \ldots, m_{n_{k,K}}]$, each model receiving a weight of c^{-1}.

The stochastic search algorithms select models according to a fitness function, $f(\cdot)$, such as the model posterior. The included

[5] This approach avoids using uniform priors over the model space, which can lead to undesirable properties, particularly when regressors are correlated, see George and Foster (2000).

Table 1

Monte Carlo simulation results. This table reports the R^2 from a linear prediction model $y_{t+1} = x_t'\beta + \epsilon_{t+1}$, with X containing eight predictors. The X-variables and ϵ are assumed to be normally distributed and i.i.d., while $\beta_i = b_i \frac{\sigma_\epsilon}{\sqrt{T}}$. The covariance matrix of the predictor variables has ones on the diagonal and ρ in all off-diagonal cells, so ρ controls the degree of correlation among the predictors. All forecasting methods only use information up to time T to produce predictions $\hat{y}_{T+1}^{(j)}$, where j refers to the simulation number. The prevailing mean forecast is $\bar{y}_{T+1}^{(j)} = \frac{1}{T}\sum_{t=1}^{T} y_t^{(j)}$. The reported out-of-sample R^2 is computed as $R^2 = \left(1 - \frac{\sum_{j=1}^{25.000}(y_{T+1}^{(j)} - \hat{y}_{T+1}^{(j)})^2}{\sum_{j=1}^{25.000}(y_{T+1}^{(j)} - \bar{y}_{T+1}^{(j)})^2}\right) \times 100$ and is reported in parentheses. The results are based on 25,000 simulations and a sample size of $T = 100$ observations.

	$b = [1\,1\,1\,1\;\;1\,1\,1\,1]$								$b = [1\,1\,1\,1\,0\,0\,0\,0]$							
	Subset regression															
	R^2								R^2							
k	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
ρ																
0	1.608	2.725	3.356	3.498	3.139	2.256	0.816	−1.227	0.690	1.000	0.930	0.474	−0.380	−1.653	−3.378	−5.593
0.25	10.361	14.291	15.782	16.092	15.698	14.796	13.469	11.736	3.213	4.380	4.651	4.364	3.632	2.488	0.926	−1.081
0.5	21.025	24.081	24.618	24.349	23.645	22.606	21.255	19.578	6.334	7.131	7.033	6.495	5.593	4.330	2.685	0.624
0.75	31.821	32.845	32.741	32.275	31.553	30.589	29.374	27.886	10.378	10.579	10.268	9.641	8.710	7.462	5.871	3.903
0.95	37.557	37.475	37.176	36.686	35.994	35.089	33.952	32.556	12.529	12.359	11.921	11.234	10.251	8.979	7.373	5.404
	Ridge regression															
	R^2								R^2							
γ	30	60	90	120	150	180	210	240	30	60	90	120	150	180	210	240
ρ																
0	2.194	3.206	3.486	3.497	3.400	3.261	3.110	2.960	−1.732	−0.262	0.405	0.732	0.897	0.978	1.011	1.018
0.25	14.997	15.827	15.897	15.653	15.267	14.819	14.348	13.875	2.751	3.951	4.393	4.537	4.544	4.483	4.387	4.274
0.5	23.460	24.295	24.309	24.234	23.916	23.520	23.080	22.617	5.373	6.510	6.900	7.028	7.031	6.985	6.899	6.793
0.75	32.238	32.669	32.613	32.381	32.055	31.672	31.253	30.809	9.614	10.295	10.459	10.472	10.420	10.333	10.224	10.100
0.95	37.474	37.429	37.253	37.005	36.704	36.364	35.993	35.599	12.394	12.493	12.495	12.453	12.384	12.298	12.194	12.085
	Elastic Net															
	R^2								R^2							
ψ	1	15	30	45	60	75	90	100	1	15	30	45	60	75	90	100
ρ																
0	0.000	2.215	2.082	0.724	−0.299	−0.820	−1.054	−1.128	0.000	0.533	−1.242	−3.254	−4.526	−5.136	−5.401	−5.483
0.25	0.000	10.814	14.449	13.964	12.929	12.273	11.964	11.865	0.000	3.632	3.176	1.417	0.094	−0.583	−0.874	−0.966
0.5	0.000	18.317	22.830	22.482	21.229	20.313	19.888	19.746	0.000	5.892	5.993	3.860	2.145	1.246	0.875	0.758
0.75	0.000	26.437	31.395	31.278	29.894	28.746	28.215	28.062	0.000	9.041	9.790	7.871	5.777	4.635	4.176	4.049
0.95	0.000	32.703	36.968	36.968	35.481	33.763	32.984	32.773	0.000	11.365	12.334	11.032	8.243	6.453	5.764	5.594
	Lasso															
	R^2								R^2							
ψ	1	15	30	45	60	75	90	100	1	15	30	45	60	75	90	100
ρ																
0	0.000	2.162	1.811	0.536	−0.386	−0.853	−1.064	−1.133	0.000	0.383	−1.500	−3.429	−4.608	−5.164	−5.410	−5.488
0.25	0.000	10.954	14.176	13.742	12.810	12.225	11.944	11.855	0.000	3.597	2.934	1.232	0.004	−0.617	−0.887	−0.971
0.5	0.000	18.454	22.492	22.236	21.084	20.257	19.866	19.736	0.000	5.872	5.721	3.641	2.023	1.203	0.862	0.752
0.75	0.000	26.743	31.013	30.999	29.710	28.671	28.192	28.052	0.000	9.049	9.533	7.569	5.600	4.574	4.159	4.041
0.95	0.000	32.745	36.643	36.745	35.212	33.605	32.930	32.751	0.000	11.385	12.189	10.704	7.885	6.302	5.725	5.579
	Bagging															
	R^2								R^2							
c	0.38	1.28	1.64	2.24	2.80	3.29	3.89	5.32	0.38	1.28	1.64	2.24	2.80	3.29	3.89	5.32
ρ																
0	−1.219	−0.005	0.766	1.694	1.746	1.416	0.901	0.159	−5.558	−3.610	−2.345	−0.607	0.207	0.418	0.375	0.103
0.25	11.729	12.479	12.720	11.855	9.244	6.366	3.320	0.373	−1.061	0.654	1.709	2.690	2.429	1.758	0.927	0.093
0.5	19.581	20.546	20.848	19.403	14.738	9.704	4.771	0.455	0.656	2.689	3.926	4.766	3.852	2.566	1.228	0.102
0.75	27.902	29.061	29.417	26.747	19.246	11.942	5.392	0.428	3.937	6.086	7.364	7.669	5.613	3.451	1.509	0.108
0.95	32.617	34.233	34.550	29.972	19.913	11.435	4.696	0.310	5.467	8.094	9.615	9.596	6.614	3.853	1.579	0.119
	Bayesian Model Averaging															
	R^2								R^2							
m	4				6				2				4			
g	$1/T$		$1/K^2$		$1/T$		$1/K^2$		$1/T$		$1/K^2$		$1/T$		$1/K^2$	
θ	Fix	Random	Fix	Random	Fix	Random	Fix	Random	Fix	Random	Fix	Random	Fix	Random	Fix	Random
ρ																
0	1.281	0.932	1.321	1.034	1.111	0.976	1.069	1.077	0.180	0.241	0.154	0.236	−0.225	0.219	−0.336	0.191
0.25	12.445	11.136	12.669	11.540	13.158	11.342	13.245	11.746	2.655	2.152	2.757	2.294	2.926	2.297	2.941	2.431
0.5	21.229	20.392	21.385	20.650	21.607	20.512	21.637	20.762	5.124	4.548	5.228	4.699	5.358	4.691	5.362	4.829
0.75	30.354	29.825	30.434	29.966	30.389	29.884	30.348	30.019	9.095	8.628	9.142	8.733	9.032	8.715	8.977	8.792
0.95	36.586	36.856	36.511	36.817	35.921	36.830	35.761	36.779	11.935	11.635	11.943	11.695	11.635	11.679	11.524	11.693

Table 2
Monte Carlo simulation results for large values of K. This table reports the out-of-sample R^2-value from a linear prediction model $y_{t+1} = x_t'\beta + \varepsilon_{t+1}$, with X containing $K = 15$ (Panel A) or $K = 20$ (Panel B) predictors. The X-variables and ε are assumed to be normally distributed and i.i.d., while $\beta_i = \sigma_\varepsilon/\sqrt{T}$. The covariance matrix of the predictor variables has ones on the diagonal and $\rho = 0.5$ in all off-diagonal cells. Forecasting methods only use information up to time T to produce predictions $\hat{y}_{T+1}^{(j)}$, where j refers to the simulation number and $T = 100$.

Out-of-sample R^2								
k	Panel A				k	Panel B		
	Complete subset	Random	Markov chain	Shotgun		Random	Markov chain	Shotgun
1	43.409	43.409	43.409	43.409	1	51.909	51.909	51.909
2	50.652	49.757	49.925	49.84	2	62.485	61.362	61.696
3	52.901	52.826	52.733	52.81	3	65.61	65.273	65.217
4	53.768	53.759	53.765	53.023	4	67.657	67.898	66.395
5	54.095	54.056	54.095	53.697	5	68.687	68.761	68.25
6	54.147	54.174	54.178	53.897	6	69.363	69.424	69.507
7	54.03	54.044	53.861	53.536	7	69.839	69.837	70.236
8	53.786	53.801	53.758	53.364	8	70.017	70.003	70.206
9	53.435	53.427	53.258	52.753	9	70.272	70.24	69.832
10	52.982	53.023	52.915	52.857	10	70.37	70.381	70.336
11	52.429	52.454	52.372	52.971	11	70.39	70.291	69.815
12	51.77	51.631	51.545	51.114	12	70.3	70.381	70.3
13	50.999	50.972	50.987	50.362	13	70.221	70.4	70.091
14	50.104	50.104	50.104	50.104	14	70.103	70.073	69.966
15	49.073	49.073	49.073	49.073	15	69.863	69.734	69.63
					16	69.565	69.542	69.886
					17	69.011	68.992	69.101
					18	68.732	68.424	68.934
					19	68.332	68.332	68.332
					20	67.778	67.778	67.778

models, as well as their weights, depend on the chain's path, with models never visited receiving a zero weight, while visited models receive a weight proportional to the number of visits divided by the length of the chain.

Specifically, the Markov chain moves from model m^t to the next candidate model, m^{t+1}, based on a uniform probability draw from the set of models $N_{m^t} \subset M_k$, where N_{m^t} represents the set of models containing at least $k-1$ of the variables originally in m^t. The transition probability of the chain is $p = \min\left(1, \frac{f(m^{t+1})}{f(m^t)}\right)$. If the candidate model offers a better fit, so $f(m^{t+1}) > f(m^t)$, the chain jumps to m^{t+1} for sure; if this condition fails, the chain may still move to $f(m^{t+1})$ since this prevents the chain from being trapped in local solutions. The worse the relative fit of the candidate model, the lower the probability of such a move.

Under the shotgun approach, the candidate model, m^{t+1}, following from an initial model m^t, is drawn from N_{m^t} with a probability proportional to its fit, so that the jth candidate model has probability $p(m_j) = f(m_j)/\sum_{N_{m^t}} f(m_i)$. Here the transition probability is $p = \min\left(1, \sum_{N_{m^{t+1}}} f(m_j)/\sum_{N_{m^t}} f(m_i)\right)$.

Panel A of Table 2 reports results in the form of out-of-sample R^2-values. These values are very similar across each of the columns, suggesting that very little is lost in terms of performance of the combined forecast as a result of using only a portion of all models within a given subset.

We next increase K to 20. In this case, some of the $n_{k,K}$ values are too large to allow us to evaluate the complete subset combination and so we only present results for cases (i)–(iii). Panel B in Table 2 shows that, once again, there is little to distinguish between models selected randomly versus models selected by the Markov chain or shotgun approaches. These findings suggest that our subset combination approach can be implemented without much loss when K is large.

4. Empirical application: stock return predictions

To illustrate the complete subset regression approach to forecast combination and to compare its performance against that of other approaches, this section provides an empirical application to US stock returns. This application is well suited for our analysis both because predictability of stock returns has been the subject of an extensive literature in finance, recently summarized by Rapach and Zhou (forthcoming), and because there is a great deal of uncertainty about which, if any, predictors help forecast stock returns. Clearly this is a case where estimation error matters a great deal.

Specifically, we investigate whether there is any improvement in the subset regression forecasts that combine k-variate models for $k \geq 2$ relative to using a simple equal-weighted combination of univariate models ($k = 1$), as proposed in Rapach et al. (2010), or relative to other combination schemes such as those described in the previous section.

Predictability of US stock returns by means of combinations of models based on different sets of predictors has been considered by studies such as Avramov (2002), Cremers (2002), and Rapach et al. (2010). For example, Avramov (2002) uses BMA on all possible combinations of models with 16 predictors to forecast monthly returns.

Our analysis again ignores model instability. Models with time-varying coefficients have been considered for stock market data by Griffin and Kalli (2012) and Dangl and Halling (2012), while Pettenuzzo and Timmermann (2011) consider forecast combination in the presence of model instability.

Diebold (2012) discusses the merits of out-of-sample versus in-sample tests of predictive accuracy. From the perspective of inference on model validity, in-sample performance tests provide higher power than out-of-sample tests. However, in applications such as ours where the interest lies in testing whether a method could have been used in real time to generate forecasts that led to better economic decisions (portfolio holdings) that improved economic utility, an out-of-sample perspective seems appropriate.

4.1. Data

Data are taken from Goyal and Welch (2008), updated to 2010, and are recorded at the quarterly horizon over the period

1947Q1–2010Q4. The list of predictors comprises 12 variables for a total of $2^{12} = 4096$ possible models.[6]

The 12 variables are the Dividend Price Ratio (dp), the difference between the log of the 12-month moving sum of dividends and the log of the S&P 500 index; Dividend Yield (dy), the difference between the log of the 12-month moving sum of dividends and the lagged log S&P 500 index; Earnings Price Ratio (ep), the difference between the log of the 12-month moving sum of earnings and the log S&P 500 index; Book to Market (bm), the ratio of the book value to market value for the Dow Jones Industrial Average; Net Equity Expansion (ntis), the ratio of the 12-month moving sum of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks; Treasury Bill (tbl), the 3-Month Treasury Bill (secondary market) rate; Long Term Rate of Returns (ltr), the long-term rate of return on US Bonds; Term Spread (tms), the difference between the long term yield on government bonds and the Treasury Bill rate; Default Yield Spread (dfy), the difference between yields on AAA- and BAA-rated bonds; Default Return Spread (dfr), the difference between long-term corporate bond and long-term government bond returns; Inflation (infl), the (log) growth of the Consumer Price Index (All Urban Consumers); and Investment to Capital Ratio (ik), the ratio of aggregate investments to aggregate capital for the whole economy.

The excess return, our dependent variable, is the difference between the continuously compounded return on the S&P 500 index (including dividends) and the 3-month Treasury Bill rate. As in Rapach et al. (2010) and Goyal and Welch (2008), we adopt a recursively expanding estimation scheme. The initial estimation sample goes from 1947Q1 to 1964Q4, yielding a first forecast for 1965Q1, while the last forecast is for 2010Q4. Each quarter's parameters are (re)estimated using all available information up to that point. This pseudo out-of-sample forecasting exercise simulates the practice of a real time forecaster. As in the theoretical analysis, forecasts are generated from the following predictive regression:

$$r_{2:t+1} = \alpha + (X_{1:t}S)\beta + \epsilon_{2:t+1}, \qquad (17)$$

where $r_{2:t+1}$ is the excess return defined above, $X_{1:t}$ is the full regressor matrix, $\epsilon_{2:t+1}$ is a vector of error terms, α and β are unknown parameters estimated by OLS, and S is a diagonal selector matrix whose unity elements determine which variables get included in the model. For example, the "kitchen sink" model containing all predictors is obtained by setting $S = I_{12}$, while the constant 'null' model is obtained by setting S equal to a 12×12 matrix of zeros. Following the analysis in Section 2, our focus is on the combination of k-variate models,

$$\hat{r}^k_{t+1} = \frac{1}{n_{k,K}} \sum_{j=1}^{n_{k,K}} (\hat{\alpha}_j + x'_t S_j \hat{\beta}_j) \quad \text{s.t. } \mathrm{tr}(S_j) = k, \qquad (18)$$

where $tr(\circ)$ is the trace operator.

4.2. Bias–variance trade-off

Fig. 5 plots time-series of out-of-sample forecasts of returns for the different k-variate subset regression combinations. The forecasts display similar patterns except that as k increases, the variance of the combined forecasts also increases. The least volatile forecasts are generated by the constant model ($k = 0$), while the most volatile forecasts arise when we use the model that contains all regressors ($k = K = 12$). Neither of these cases perform any forecast combination. As we shall subsequently see, forecasts from the best k-variate combinations are in turn more volatile than those from combinations of univariate models ($k = 1$) but less volatile than those from the other k-variate combinations. The extent to which volatility of the forecast reduces or enhances forecast performance depends, of course, on how strongly this variation is correlated with the outcome—a point we further address below.

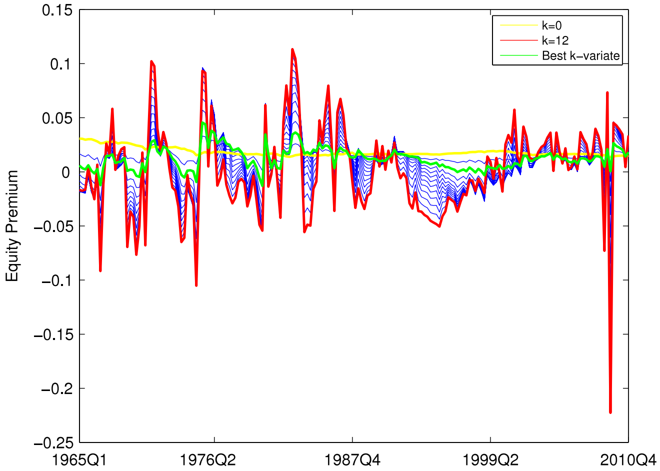

Fig. 5. Out-of-sample forecasts of monthly stock returns for different k-variate subset combinations.

Fig. 6 provides insight into the relation between the variance and bias of the forecasts. Along the x-axis, the upper left window lists the number of predictors included in each model, k, while the y-axis lists the time-series variance associated with a given model. For example, for $k = 1$ the circles show the variance for each of the 12 univariate forecasting models, while for $k = 2$ the circles show the forecast variance for each of the 66 bivariate models. The upper left graph shows that the variance of the forecast is increasing in the number of variables included in the forecast models. To see why, define $x^S_t = x_t S$ and $X^S_{1:T} = X_{1:T}S$, and note that

$$\mathrm{var}(\hat{r}_{t+1}) = \mathrm{var}(\hat{\alpha} + x^S_t \hat{\beta}) = [\iota'\iota + x^S_t (X'_{1:T}X_{1:T})^{-1} x^{S\prime}_t]\hat{\sigma}_\epsilon, \qquad (19)$$

which is increasing in $\hat{\sigma}_\epsilon$, the estimated standard deviation of the residuals, and in the column dimension of ι', $x^{S\prime}_t$ and X^S. Therefore, the larger the dimension of the pooled models, the higher the forecast variance.

The upper right window in Fig. 6 shows how the squared bias is reduced by pooling the models. Specifically, the combination of the three-variate models has the lowest bias. The constant model produces the most (upward) biased forecasts. At the other end of the spectrum, the "kitchen sink" model with all variables included generates the most biased forecasts because of its occasional extreme negative forecasts (see Fig. 5). Except for the models based on dp, dy and ep, the individual univariate models generate a large bias.

Putting together the forecast variance and bias results, the bottom window of Fig. 6 establishes a (squared) bias–variance trade-off. This resembles the well-known mean–variance efficient frontier known from modern portfolio theory in finance, albeit with the role of the bias and variance reversed. In our example, the (squared) bias is largest for models with either very few or very many predictors, while the variance increases monotonically in k.

4.3. Performance of subset regressions

To gain insight into the forecast performance of the various models, Fig. 7 plots the out-of-sample R^2 (top window) and

[6] Data are available at http://www.hec.unil.ch/agoyal/. Variable definitions and data sources are described in more detail in Goyal and Welch (2008). To avoid multicollinearity when estimating some of the multivariate models, we exclude the log dividend earnings ratio and the long term yield. By construction, the log dividend earnings ratio is equal to the difference between the log dividend price ratio and the log earnings price ratio, while the long term yield is equal to the sum of the term spread and the Treasury Bill rate.

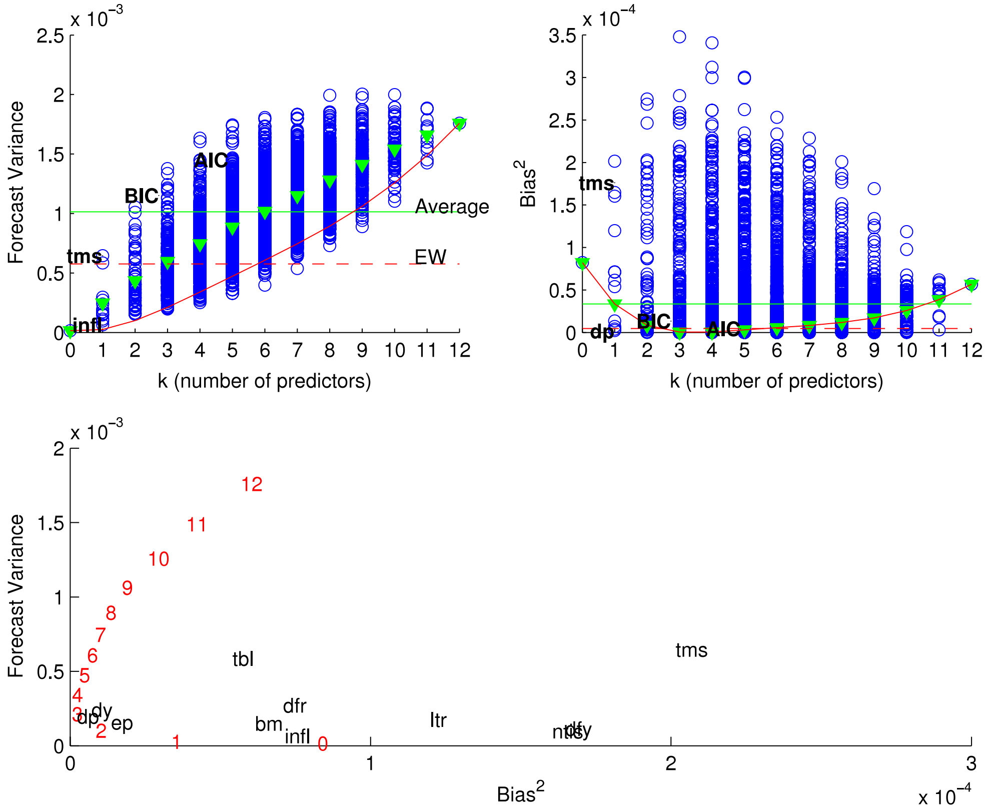

Fig. 6. Bias–variance trade-off. Each circle represents a single regression model, grouped according to the number of predictors the model contains. Triangles represent average values computed across all models with a given number of predictors, k, i.e., for a given subset. The horizontal line marked 'Average' shows the performance averaged across all 4096 models while the dotted horizontal line marked 'EW' refers to the performance of the equal-weighted forecast combination based on all models. The full curved line tracks the subset combination of the k-variate models. The best and worst univariate models are displayed as text strings; AIC and BIC refer to the models recursively selected by these information criteria. The bottom figure displays the scatter plot of the squared bias against the variance for each of the k-variate subset combinations (with k denoted in red) as well as for the individual univariate models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the MSE-value (bottom window) for the individual k-variate forecasting models along with those for the subset regression combinations.[7] The lower x-axis shows the number of predictors included in each model, while the upper x-axis in the top window lists the total number of k-variate models, i.e., $n_{k,12}$. For $1 \leq k \leq 6$, the k-variate combinations generate lower MSE-values than the equal-weighted average forecast computed across all 4096 models, the "thick" forecast modeling approach used by Aiolfi and Favero (2003). They also perform better than the constant equity premium model ($k = 0$), a benchmark considered difficult to beat in the finance literature, see Goyal and Welch (2008).

Interestingly, the two and three-variate combinations generate out-of-sample R^2-values that are 1% higher than the univariate combination approach used by Rapach et al. (2010). This may not seem like such a large difference but, as emphasized by Campbell and Thompson (2008), even small differences in out-of-sample R^2 can translate into economically large gains in investor utility.

Fig. 6 showed that the forecast results do not depend simply on the number of pooled forecasts. For example, there are 66 two-variate as well as ten-variate models, but the corresponding equal-weighted combinations produce very different outcomes. This is not surprising given that the worst two-variate model is better than the best ten-variate model. To control for the mere effect of the number of models included in the combination, we also combine models that are randomly selected across different values of k. Fig. 8 plots the out-of-sample MSE- and R^2-values as a function of the number of models in the combined forecast. Less than 100 models, i.e. about 2% of the total, need to be pooled in order to approximate the behavior of the forecasts obtained by combining all models.[8]

The benefit of subset combination is evident from three observations. First, the k-variate subset combinations have similar, if not better (for $k = 1, 2, 3, 10$ and 11), performance compared with the single best k-variate model, the identity of which is difficult to establish ex ante. Second, for $k \leq 10$ the k-variate combinations produce better results than models selected by recursively applying information criteria such as the AIC or the BIC. This happens despite the fact that these subset combinations contain, on average, the same or a larger number of predictors.[9] Third, while some univariate models, the ones containing dp, dy, dfr, and ik, produce better results than the equal-weighted combination of all models, in contrast no single predictor model does better than the three best-performing k-variate subset combinations.

[7] The out-of-sample R^2-value is computed as

$$R^2 = 1 - \frac{\sum_{\tau=T_0}^{T-1} (r_{\tau+1} - \hat{r}_{\tau+1|\tau})^2}{\sum_{\tau=T_0}^{T-1} (r_{\tau+1} - \hat{r}^{bmk}_{\tau+1|\tau})^2},$$

where T_0 is the start of the evaluation period and T is the final data point.

[8] This finding becomes very relevant in situations where it is infeasible to estimate all 2^K models, e.g., when $K > 20$, since the number of models is exponentially related to the number of predictors.

[9] On average, the BIC and AIC criteria select 2.73 and 4.88 predictors, respectively.

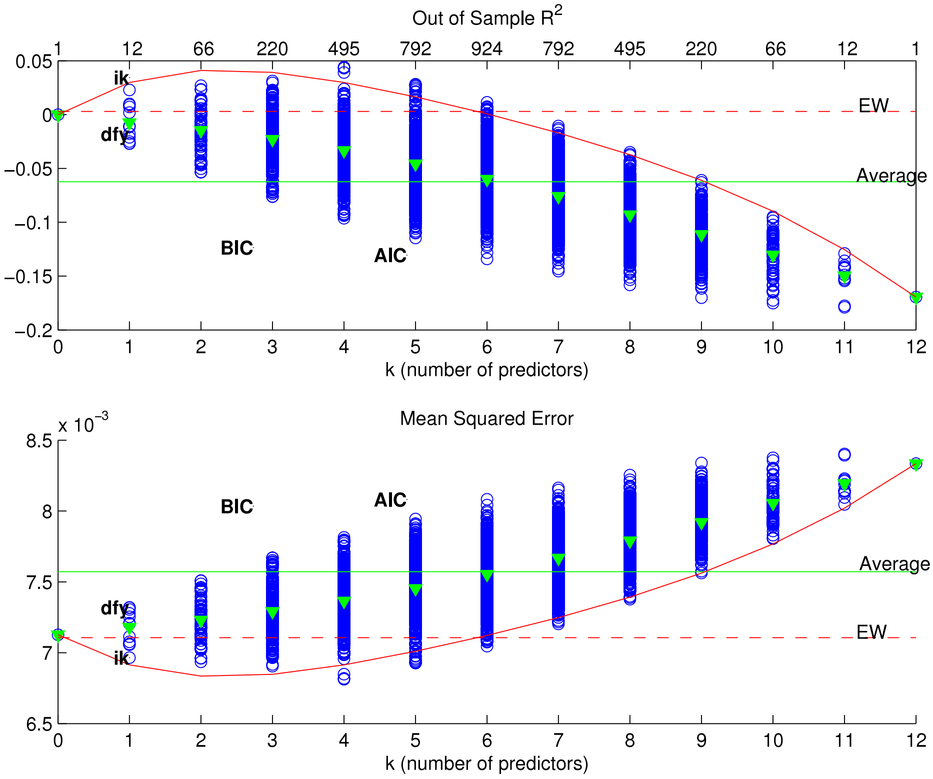

Fig. 7. Out-of-sample forecast performance. Each circle represents a single regression model, grouped according to the number of predictors the model contains. For a given value of k, the number of possible k-variate models, $\binom{12}{k} = \frac{12!}{k!(12-k)!}$, is reported on the upper x-axis at the top of the diagram. Triangles represent average values computed across all models with a given number of predictors, k. The horizontal line marked 'Average' shows the performance averaged across all 4096 models while the dotted horizontal line marked 'EW' refers to the performance of the equal-weighted forecast combination based on all models. The full curved line tracks the subset combination of the k-variate models. The best and worst univariate models are displayed as text strings above $k = 1$; AIC and BIC refer to the models recursively selected by these information criteria.

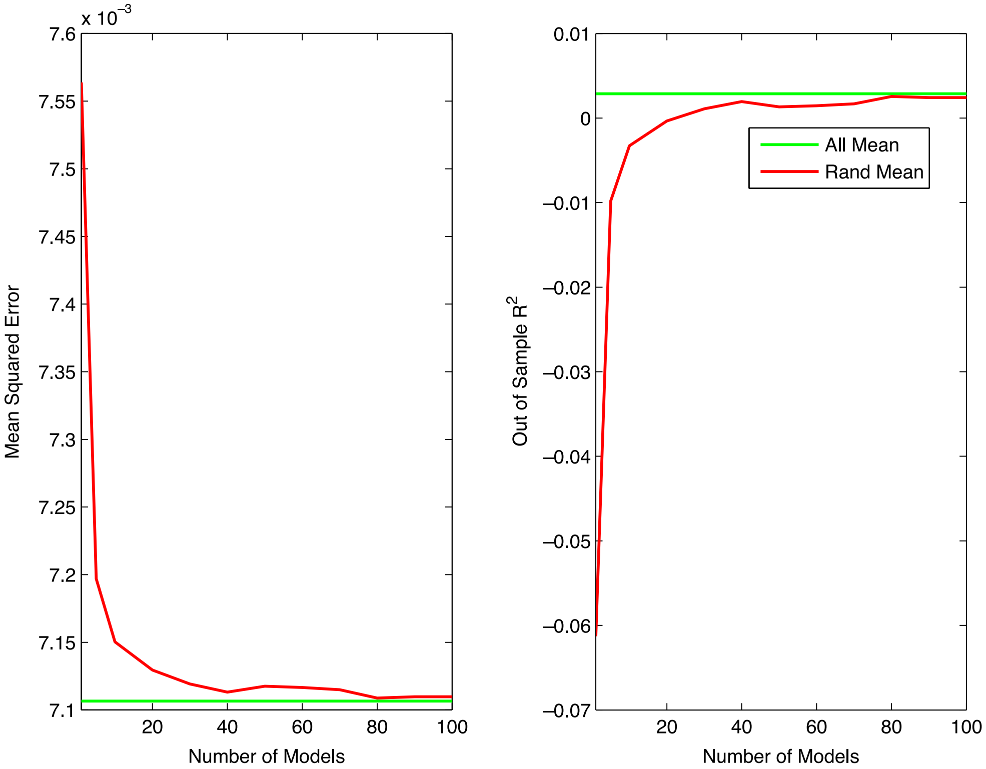

Fig. 8. Performance of pools of randomly selected models. At each point in time, n models are randomly selected (without replacement), their forecasts pooled, and the forecast performance recorded. This procedure is repeated 1000 times. The solid line tracks the median value across these trials. For comparison, the horizontal line shows the performance of the combination of all 4096 models.

Table 3

Out-of-sample forecast performance for US stock returns. Panel A displays the out-of-sample forecast performance for the 12 univariate models, Panel B for the subset regression, Panel C for Lasso and Elastic Net, Panel D for ridge regression, Panel E for bagging and Panel F for Bayesian Model Averaging. The p-values associated with the out-of-sample R^2 are based on the one-sided test of Clark and West (2007), and the encompassing test of Harvey et al. (1998). All forecasts of quarterly stock returns are computed recursively and cover the period 1965Q1–2010Q4. MSE is the out-of-sample mean squared error, R^2 is the out-of-sample R^2, CER is the certainty equivalent return. Except for the p-values and MSE, all entries are in percentages.

Panel A: Univariate

Variable	MSE	R^2	CER	p-val$_{\text{CW}}$	p-val$_{\text{HLN}}$
dp	0.708	0.708	0.247	0.039	0.041
dy	0.706	0.986	0.405	0.030	0.031
ep	0.720	−1.066	−0.016	0.297	0.298
bm	0.725	−1.767	−1.272	0.427	0.428
ntis	0.728	−2.115	−0.594	0.630	0.629
tbl	0.731	−2.502	0.164	0.046	0.048
ltr	0.721	−1.150	−0.100	0.305	0.306
tms	0.732	−2.672	−1.305	0.056	0.058
dfy	0.732	−2.699	−1.119	0.717	0.716
dfr	0.706	0.906	−0.110	0.110	0.112
infl	0.711	0.192	0.269	0.307	0.308
ik	0.696	2.281	1.054	0.010	0.011

Panel B: Subset regression

k	MSE	R^2	CER	p-val$_{\text{CW}}$	p-val$_{\text{HLN}}$
0	0.713	0.000	0.000		
1	0.691	2.991	1.364	0.002	0.002
2	0.684	4.097	1.984	0.004	0.004
3	0.685	3.923	2.088	0.006	0.007
4	0.691	2.985	1.897	0.009	0.010
5	0.701	1.643	1.572	0.014	0.015
6	0.712	0.073	1.200	0.020	0.021
7	0.725	−1.696	0.805	0.027	0.028
8	0.739	−3.716	0.371	0.035	0.037
9	0.756	−6.096	−0.139	0.046	0.047
10	0.777	−8.979	−0.778	0.058	0.059
11	0.802	12.535	−1.610	0.072	0.074
12	0.833	16.948	−2.697	0.090	0.092

Panel C.1: Lasso

ψ	MSE	R^2	CER	p-val$_{\text{CW}}$	p-val$_{\text{HLN}}$
1	0.713	0.000	−0.000	0.255	0.256
2	0.712	0.055	0.247	0.392	0.393
3	0.712	0.073	0.261	0.370	0.371
5	0.717	−0.662	0.401	0.502	0.502
20	0.733	−2.829	0.341	0.140	0.142
50	0.782	−9.721	−0.978	0.056	0.058
Panel C.2: Elastic-Net ($\alpha = 0.5$)					
1	0.713	0.000	−0.000	0.255	0.256
2	0.712	0.044	0.197	0.401	0.402
3	0.711	0.217	0.325	0.301	0.303
5	0.714	−0.174	0.454	0.403	0.404
20	0.726	−1.891	0.563	0.110	0.112
50	0.780	−9.432	−0.912	0.055	0.056

Panel D: Ridge regression

γ	MSE	R^2	CER	p-val$_{\text{CW}}$	p-val$_{\text{HLN}}$
0.5	0.824	−15.630	−2.254	0.084	0.086
1	0.817	−14.671	−1.965	0.080	0.082
2	0.807	−13.268	−1.582	0.074	0.076
3	0.800	−12.227	−1.322	0.070	0.072
4	0.794	−11.389	−1.125	0.068	0.070
5	0.789	−10.684	−0.965	0.065	0.067
10	0.771	−8.185	−0.426	0.057	0.059
20	0.750	−5.289	0.177	0.047	0.049
50	0.722	−1.314	1.004	0.032	0.034
100	0.704	1.203	1.515	0.024	0.025
150	0.697	2.266	1.707	0.020	0.021
200	0.693	2.793	1.778	0.017	0.019

Panel E: Bagging

c	MSE	R^2	CER	p-val$_{\text{CW}}$	p-val$_{\text{HLN}}$
0.3853	0.754	−5.754	−0.948	0.156	0.158
0.6745	0.745	−4.557	−0.550	0.137	0.139
1.2816	0.717	−0.613	0.736	0.082	0.084
1.4395	0.710	0.379	0.968	0.065	0.067
1.6449	0.705	1.137	1.109	0.055	0.057
1.96	0.700	1.725	1.131	0.044	0.045
2.2414	0.703	1.368	0.892	0.059	0.061
2.5758	0.706	0.979	0.608	0.080	0.082
2.807	0.706	0.945	0.493	0.069	0.071
3.0233	0.709	0.579	0.324	0.116	0.118
3.2905	0.710	0.418	0.203	0.131	0.133
3.4808	0.711	0.222	0.145	0.218	0.219
3.8906	0.712	0.105	0.058	0.268	0.269
4.4172	0.712	0.072	0.033	0.223	0.224
5.3267	0.713	0.021	0.003	0.349	0.350

Panel F: Bayesian Model Averaging

$m\|g\|\theta$	MSE	R^2	CER	p-val$_{\text{CW}}$	p-val$_{\text{HLN}}$
$0.1\|\frac{1}{n}\|\text{fix}$	0.723	−1.477	0.199	0.495	0.495
$0.1\|\frac{1}{k^2}\|\text{fix}$	0.723	−1.423	0.166	0.519	0.518
$0.1\|1\|\text{fix}$	0.709	0.578	0.389	0.112	0.114
$0.1\|\frac{1}{n}\|\text{rnd}$	0.724	−1.601	0.162	0.510	0.510
$0.1\|\frac{1}{k^2}\|\text{rnd}$	0.723	−1.506	0.134	0.532	0.532
$0.1\|1\|\text{rnd}$	0.709	0.578	0.390	0.113	0.115
$1\|\frac{1}{n}\|\text{fix}$	0.741	−4.029	0.212	0.392	0.393
$1\|\frac{1}{k^2}\|\text{fix}$	0.737	−3.478	0.214	0.401	0.402
$1\|1\|\text{fix}$	0.699	1.882	1.441	0.052	0.054
$1\|\frac{1}{n}\|\text{rnd}$	0.746	−4.686	0.129	0.462	0.462
$1\|\frac{1}{k^2}\|\text{rnd}$	0.739	−3.703	0.147	0.464	0.464
$1\|1\|\text{rnd}$	0.703	1.338	1.396	0.072	0.074

4.4. Performance comparisons

Table 3 presents out-of-sample R^2-values. First, consider the univariate models shown in Panel A. Only five of the twelve variables generate positive out-of-sample R^2-values, the highest such value being 2.28% for the investment–capital ratio. Panel B shows that all subset regressions with $k \leq 6$ generate positive out-of-sample R^2-values, the largest values occurring for $k = 2$ or $k = 3$, which lead to an R^2 around 4%. As k grows larger, the out-of-sample forecasting performance quickly deteriorates, with values below −10% when $k = 11$ or $k = 12$.[10]

Turning to the alternative approaches described earlier, Panel C shows that the Lasso forecasts are only capable of producing small positive R^2-values for $\psi \leq 3$ and generate large negative R^2-values for the largest values of ψ. Panel D shows that the ridge regressions generate large negative R^2-values when the shrinkage parameter, γ, is small, corresponding to the inclusion of many predictors. Better performance is reached for higher values of γ, but even the best value of γ only leads to an R^2 of 2.8%. The bagging approach (panel E) suffers from similar deficiencies when c is small, leading to large prediction models, but improves for values of c around two for which an R^2 of 1.7% is reached. Turning to the BMA results, we also consider a value of $g = 1$, in addition to the previous values of $g = 1/k^2$ and $g = 1/n$. This value of g induces less concentrated weights on a few models, which turns out to be advantageous here. Indeed, the Bayesian Model Averaging forecasts produce positive R^2-values in three out of four cases when $g = 1$ and otherwise mostly produce negative R^2-values.

To compare model performance more formally, we use the test proposed by Clark and West (2007), treating the simple prevailing mean forecast as our benchmark. This test re-centers the difference in mean squared forecast errors to account for the higher variability associated with forecasts from larger models. The test results show that three of the univariate models (corresponding to dp, dy, and ik) produce better forecasting performance than the benchmark at the 5% significance level. For the bagging method, forecasting performance superior to the benchmark is obtained only when c is around two, while the BMA fails to dominate the benchmark. The ridge regressions produce significantly improved forecasts for $\gamma \geq 20$, while the subset regressions do so for all but the largest models, i.e., as long as $k \leq 9$. Notably, the rejections are much stronger for many of the subset regressions, with p-values below 1% as long as $k \leq 5$. Similar results are obtained when the encompassing test of Harvey et al. (1998) is adopted.

4.4.1. Recursive selection of hyperparameters

Our results so far show that the choice of hyperparameter can matter a great deal for the performance of many of the combination approaches. It is therefore important to establish whether such hyperparameters can be chosen recursively, in "real time" to deliver good forecasting performance. To this end, we conduct an experiment that, at each point in time, uses the data up to this point (but not thereafter) to select the value of the hyperparameter which would have given the best performance. Fig. 9 shows the recursively chosen values for the hyperparameters. The subset

[10] Very similar results were obtained when we expanded our list of predictor variables to include a liquidity measure as proposed by Amihud (2002).

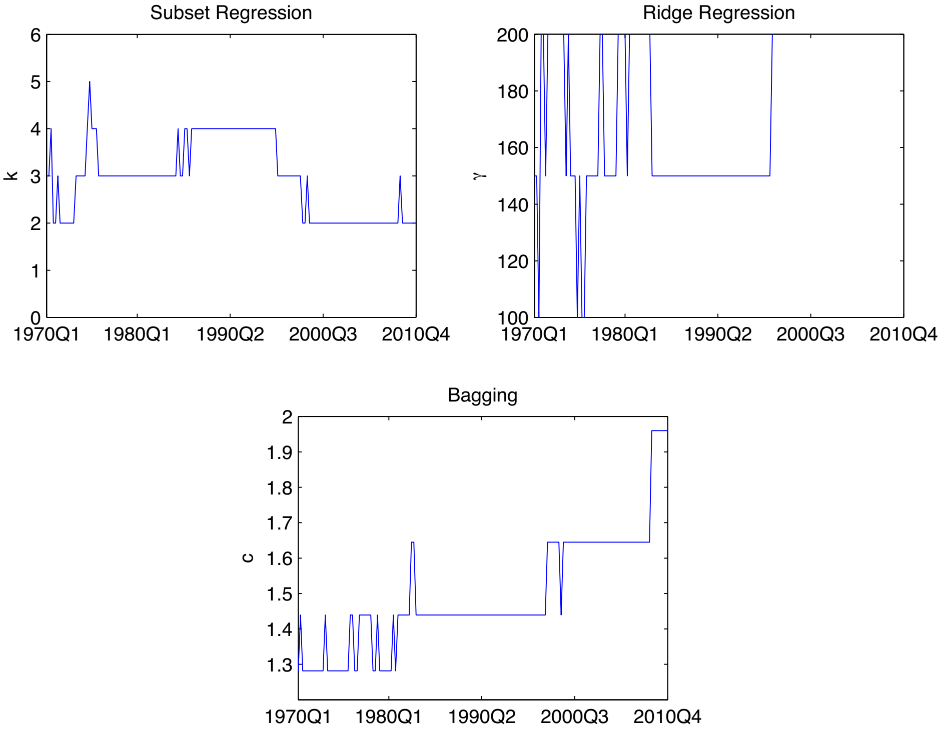

Fig. 9. Recursive choice of parameter values. For each approach the graphs show the value of the design parameter that, at each point in time, gets selected based on its recursive performance evaluated up to that point in time. Parameter selection is performed from the discrete grid of values described in the text of the paper.

Table 4
Out-of-sample forecast performance with recursively selected hyperparameters. This table displays the out-of-sample forecast performance when the model parameters are chosen recursively in a pseudo out-of-sample experiment with an expanding estimation window. p-values are based on Clark and West (2007). The forecast evaluation period is 1970Q1–2010Q4. MSE is the out-of-sample mean squared error, R^2 is the out-of-sample R^2, CER is the certainty equivalent return. Except for the p-values and MSE, all entries are in percentages.

	MSE	R^2	CER	p-val$_{\text{CW}}$	p-val$_{\text{HLN}}$
Univariate	0.826	−9.805	−2.435	0.740	0.739
Subset	0.741	1.515	1.294	0.074	0.076
Lasso	0.769	−2.137	−0.119	0.407	0.408
Elastic Net	0.787	−4.652	−0.593	0.745	0.743
Ridge	0.747	0.704	1.093	0.076	0.079
Bagging	0.750	0.328	0.712	0.075	0.078
BMA	0.764	−1.570	0.540	0.355	0.356

Table 5
Out-of-sample forecast performance under different weighting schemes. This table displays the out-of-sample forecast performance of complete subset regressions when models are combined using equal weights (EW) versus weights that are proportional to the values of the models' Bayes information criterion (BIC). MSE is the out-of-sample mean squared error, R^2 is the out-of-sample R^2, CER is the certainty equivalent return. The forecast evaluation period is 1970Q1–2010Q4.

k	MSE$_{\text{BIC}}$	MSE$_{\text{EW}}$	R^2_{BIC}	R^2_{EW}
1	0.692	0.691	2.967	2.991
2	0.683	0.684	4.135	4.097
3	0.684	0.685	4.010	3.923
4	0.691	0.691	3.091	2.985
5	0.700	0.701	1.750	1.643
6	0.711	0.712	0.179	0.073
7	0.724	0.725	−1.586	−1.696
8	0.738	0.739	−3.510	−3.716
9	0.755	0.756	−5.975	−6.096
10	0.776	0.777	−8.868	−8.979
11	0.801	0.802	−12.458	−12.535
12	0.833	0.833	−16.948	−16.948

regression approach always chooses $k = 2$ or $k = 3$, with $k = 2$ being chosen almost exclusively from 1990 onwards. The value for γ chosen under the ridge approach fluctuates between 100 and 200. The critical value, c, in the bagging approach fluctuates between 1.2 and 2.2, while ϕ fluctuates between zero and 100 under the BMA approach.

Table 4 shows the resulting forecast performance numbers from this exercise. The univariate regression approach performs very poorly by this measure, as do the Lasso, Elastic Net and BMA approaches, all of which generate negative R^2-values. Bagging produces an R^2 of 0.3%, while the ridge approach generates an R^2-value around 0.7%. The best approach, however, is the subset regression method which generates an R^2-value of 1.5%. Using the Clark–West p-values, the subset, ridge, and bagging forecasts all improve on the prevailing mean forecast at the 10% significance level.

4.4.2. Performance with BIC weights

Our approach uses equal-weighted combinations of forecasts from models within the same subset. As discussed in Section 2.5, many alternative weighting schemes have been proposed in the combination literature. One such approach is to simply let each model's weight be proportional to the exponential of its Schwarz information criterion value. Within each subset, the number of parameters is the same across models and so the models with high likelihood will obtain larger weights than models with low likelihood by this procedure.

Table 5 presents results for this combination scheme. For direct comparison, we also show results for the equal-weighted subset combination. There is evidence of slight improvement in the out-of-sample R^2-values for some subsets, but the values are very similar under the two combination schemes. Although minor improvements might be achievable by straying away from equal-weights, the convenience and simplicity of this weighting scheme justifies its use in our approach.

4.5. Economic value of forecasts

To assess the economic value of our return forecasts, we consider the value of the predictions from the perspective of a mean-variance investor who chooses portfolio weights to maximize expected utility subject to the constraint that the weight on stocks lies in the interval [0, 1.5], thus ruling out short sales and leverage above 50%.[11]

Specifically, we assume that the investor optimally allocates wealth to the aggregate stock market given estimates of the first two conditional moments of the return distribution, $E_t[r_{t+1}]-r^f_{t+1}$ and $V_t[r_{t+1}]$, where r_{t+1} is the market return and r^f_{t+1} is the risk-free rate (T-bill rate). Under mean-variance preferences, this gives rise to an optimal allocation to stocks, ω^*_t, of

$$\omega^*_t = \frac{E_t[r_{t+1}]-r^f_{t+1}}{\varphi V_t[r_{t+1}]}, \tag{20}$$

where φ captures the investor's risk aversion. We set $\varphi = 3$ in our analysis, similar to the value adopted in finance studies. Following standard methods in the literature on volatility modeling, we use a GARCH(1, 1) specification to capture time-variation in volatility, $V_t[r_{t+1}]$, but results based on a realized volatility measure are very similar. Since our focus is on predicting mean returns, we keep the volatility specification constant across all models.

Following Marquering and Verbeek (2004), the investor's ex-post realized utility is

$$u_{t+1} = r_{f,t+1} + \omega^*_t(r_{t+1} - r_{f,t+1}) - 0.5\varphi\omega^{*2}_t \mathrm{Var}_{t+1}, \tag{21}$$

where Var_{t+1} is the realized variance based on squared daily returns during month $t+1$. Finally, we compare the investor's average utility, $\bar{u} = \frac{1}{T-1}\sum_{i=1}^{T-1} u_{t+i}$, under the modeling approaches that allow for time-varying expected returns against the corresponding value under the benchmark prevailing mean model. We report results in the form of the annualized certainty equivalent return (CER), i.e., the return that would leave an investor indifferent between using the prevailing mean forecasts versus the forecasts produced by one of the other approaches. Positive values indicate that the prevailing mean method underperforms, while negative values indicate that it performs better than the alternative forecasts.

Table 3 shows that the better statistical performance of the subset and ridge regression methods translates into positive CER-values. For the subset regressions with $k = 2$ or 3 predictors, a CER-value around 2% is achieved, whereas for the ridge regressions, values around 1.5%–1.7% are achieved for the largest values of γ. Interestingly, the BMA approach delivers consistently good performance on this criterion, always generating higher CER-values than the prevailing mean model.

Moreover, Table 4 shows that when the methods are implemented recursively, the prevailing mean approach delivers higher average utility than the univariate, Lasso and Elastic Net methods. Conversely, according to this utility-based approach, the bagging and BMA methods deliver CER-values around 0.5% higher than the prevailing mean, while the ridge and subset regression approaches better the prevailing mean by more than one percent per annum.

5. Conclusion

We propose a new forecast combination approach that averages forecasts across complete subset regressions with the same number of predictor variables and thus the same degree of model complexity. In many cases the trade-off between model complexity and model fit is such that subset combinations perform well for a relatively small number of included predictors. Moreover, we find that subset regression combinations often do better than the simple equal-weighted combinations which include all models, small and large, and hence do not penalize sufficiently for including variables with weak predictive power. In many cases subset regression combinations amount to a form of shrinkage, but one that is more general than the conventional variable-by-variable shrinkage implied by ridge regression.

Empirically, in an analysis of US stock returns, we find that the subset regression approach appears to perform quite well when compared to competing approaches such as ridge regression, bagging, Lasso or Bayesian Model Averaging.

Appendix

This Appendix provides details of the technical results in the paper.

A.1. Proof of Theorem 1

Proof. The proof follows from aggregating over the finite number $n_{k,K}$ of subset regression estimators $\hat{\beta}_i = (S_i'X'XS_i)^-(S_i'X'y) = \left(S_i'\Sigma_X S_i\right)^-(S_i'\Sigma_X)\hat{\beta}_{\mathrm{OLS}} + o_p(1)$. First, note that

$$\begin{aligned}
\hat{\beta}_i &= (S_i'X'XS_i)^-(S_i'X'y)\\
&= (S_i'X'XS_i)^-(S_i'X'X)\hat{\beta}_{\mathrm{OLS}}\\
&= \left(S_i'\Sigma_X S_i\right)^-(S_i'\Sigma_X)\hat{\beta}_{\mathrm{OLS}}\\
&\quad + \left[(S_i'X'XS_i)^-(S_i'X'X) - \left(S_i'\Sigma_X S_i\right)^-(S_i'\Sigma_X)\right]\hat{\beta}_{\mathrm{OLS}}.
\end{aligned}$$

Since $\hat{\beta}_{\mathrm{OLS}} \to^p \beta$ and $T^{-1}X'X \to \Sigma_X$, we have

$$\begin{aligned}
&(S_i'X'XS_i)^-(S_i'X'X) - \left(S_i'\Sigma_X S_i\right)^-(S_i'\Sigma_X)\\
&\quad = (S_i'T^{-1}X'XS_i)^-(S_i'\Sigma_X) - \left(S_i'\Sigma_X S_i\right)^-(S_i'\Sigma_X) + o_p(1)\\
&\quad = \left[(S_i'T^{-1}X'XS_i)^- - \left(S_i'\Sigma_X S_i\right)^-\right](S_i'\Sigma_X) + o_p(1).
\end{aligned}$$

$S_i'T^{-1}X'XS_i$ can be rearranged so that the upper $k \times k$ block is $T^{-1}X_i^{*\prime}X_i^*$, where X_i^* contains the k regressors included in the ith regression. Since $T^{-1}X'X \to^p \Sigma_X$, then $T^{-1}X_i^{*\prime}X_i^* \to^p \Sigma^*_{X_i}$ (which is the variance covariance matrix of the included regressors) by the definition of convergence in probability for matrices. Rearranging the term $(S_i'T^{-1}X'XS_i)^- - \left(S_i'\Sigma_X S_i\right)^-$ in this way yields an upper $k\times k$ block that is $o_p(1)$ with the remaining blocks equal to zero. The final regressor is a sum over these individual regressors, yielding the result. □

A.2. Proof of Theorem 2

Proof. From the results of Theorem 1, we have

$$\begin{aligned}
&\sigma_\varepsilon^{-2}E\left[T(\hat{\beta}_T-\beta_0)'x_Tx_T'(\hat{\beta}_T-\beta_0)\right]\\
&\quad = \sigma_\varepsilon^{-2}E\left[T(\hat{\beta}_T-\beta_0)'\Sigma_X(\hat{\beta}_T-\beta_0)\right]\\
&\qquad + \sigma_\varepsilon^{-2}E\left[T(\hat{\beta}_{T,\mathrm{OLS}}-\beta_0)'\Lambda'(x_Tx_T'-\Sigma_X)\right.\\
&\qquad \times \left.\Lambda(\hat{\beta}_{T,\mathrm{OLS}}-\beta_0)\right] + o_p(1)\\
&\quad = \sigma_\varepsilon^{-2}E\left[T(\hat{\beta}_T-\beta_0)'\Sigma_X(\hat{\beta}_T-\beta_0)\right] + o_p(1),
\end{aligned}$$

[11] Utility-based measures of forecast performance have been widely used in studies of stock return predictability; see Pesaran and Timmermann (1995) for an early example.

where the second term is zero by the law of iterated expectations as we assume $E[(\hat{\beta}_{\text{OLS}} - \beta_0)^2|x_T] = E[(\hat{\beta}_{\text{OLS}} - \beta_0)^2]$ and $E[x_T x_T' - \Sigma_X] = 0$.

Now

$$\begin{aligned} &T^{1/2}\sigma_\varepsilon^{-1}\Sigma_X^{1/2}(\hat{\beta}_T - \beta_0) \\ &\quad = T^{1/2}\sigma_\varepsilon^{-1}\Lambda(\hat{\beta}_{T,\text{OLS}} - \beta_0) + T^{1/2}\sigma_\varepsilon^{-1}(\Lambda - I)\beta_0 + o_p(1) \\ &\quad = T^{1/2}\sigma_\varepsilon^{-1}\Lambda(\hat{\beta}_{T,\text{OLS}} - \beta_0) + (\Lambda - I)b + o_p(1), \end{aligned}$$

and so

$$\begin{aligned} &\sigma_\varepsilon^{-2}E\left[T(\hat{\beta}_T - \beta_0)'\Sigma_X(\hat{\beta}_T - \beta_0)\right] \\ &\quad = \sigma_\varepsilon^{-2}E\left[T(\hat{\beta}_{T,\text{OLS}} - \beta_0)'\Lambda'\Sigma_X\Lambda(\hat{\beta}_{T,\text{OLS}} - \beta_0)\right] \\ &\qquad + b'(\Lambda - I)'\Sigma_X(\Lambda - I)b \\ &\qquad + 2b'(\Lambda - I)'\Sigma_X\Lambda\left(\sigma_\varepsilon^{-1}\left[ET^{1/2}(\hat{\beta}_{T,\text{OLS}} - \beta)\right]\right) + o_p(1). \end{aligned}$$

Since $T^{1/2}(\hat{\beta}_{T,\text{OLS}} - \beta) \to^d N(0, \Sigma_X)$, the third term is zero in large enough samples and $\sigma_\varepsilon^{-2}T(\hat{\beta}_{T,\text{OLS}} - \beta_0)'\Lambda'\Sigma_X\Lambda(\hat{\beta}_{T,\text{OLS}} - \beta_0) \to^d Z'\Lambda'\Sigma_X\Lambda Z$ with $Z \sim N(0, \Sigma_X)$ and $E\left[Z'\Lambda'\Sigma_X\Lambda Z\right] = \sum_{j=1}^K \zeta_j$. □

References

Aiolfi, M., Favero, C.A., 2003. Model uncertainty: thick modelling and the predictability of stock returns. Journal of Forecasting 24, 233–254.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5, 31–56.

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

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Operations Research Quarterly 20, 451–468.

Billio, M., Casarin, R., Ravazzolo, F., van Dijk, H.K., 2012. Combining predictive density using Bayesian filtering with applications to US economics data. Ca'Foscari University of Venice Working Paper No. 16.

Breiman, L., 1996. Bagging predictors. Machine Learning 36, 105–139.

Campbell, J.Y., Thompson, 2008. Predicting the equity premium out of sample: can anything beat the historical average? Review of Financial Studies 21, 1201–2355.

Clark, T.E., West, K.D., 2007. Approximately normal estimator for equal predictive accuracy in nested models. Journal of Econometrics 127, 291–311.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5, 559–581.

Cremers, K., 2002. Stock return predictability: a Bayesian model selection perspective. Review of Financial Studies 15, 1223–1249.

Dangl, T., Halling, M., 2012. Predictive regressions with time-varying coefficients. Journal of Financial Economics 106, 157–181.

Diebold, F.X., 2012. Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold–Mariano Tests. Manuscript, University of Pennsylvania.

Fernandez, C., Ley, E., Steel, F.J.J., 2001a. Benchmark priors for Bayesian Model Averaging. Journal of Econometrics 100, 381–427.

Fernandez, C., Ley, E., Steel, F.J.J., 2001b. Model uncertainty in cross-country growth regressions. Journal of Applied Econometrics 16, 563–576.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.

George, E.I., Foster, D.P., 2000. Calibration and empirical Bayes variable selection. Biometrika 87 (4), 731–747.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164, 130–141.

Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508.

Griffin, J.E., Kalli, M., 2012. Time-varying sparsity in dynamic regression models. Working Paper.

Groen, J.J., Paap, R., Ravazzolo, F., 2013. Real-time inflation forecasting in a changing world. Journal of Business and Economic Statistics 31, 29–44.

Hans, C., Dobra, A., West, M., 2007. Shotgun stochastic search for large p regression. Journal of American Statistical Association 478, 507–516.

Harvey, D.I., Leybourne, S.J., Newbold, P., 1998. Tests for forecast encompassing. Journal of Business and Economic Statistics 16, 254–259.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Inoue, A., Kilian, L., 2008. How useful is bagging in forecasting economic time series? a case study of US consumer price inflation. Journal of the American Statistical Association 103, 511–522.

Koop, G., Korobilis, D., 2012. Forecasting inflation using dynamic model averaging. International Economic Review 53 (3), 867–886.

Korobilis, D., 2013. Hierarchical shrinkage priors for dynamic regressions with many predictors. International Journal of Forecasting 29, 43–59.

Lamnisos, D., Griffin, J.E., Steel, M.F.J., 2012. Adaptive Monte Carlo for Bayesian variable selection in regression models. Journal of Computational and Graphical Statistics.

Ley, E., Steel, M.F.J., 2009. On the effect of prior assumptions in Bayesian Model Averaging with applications to growth regression. Journal of Applied Econometrics 24, 651–674.

Liang, H., Zou, G., Wan, A.T.K., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. Journal of the American Statistical Association 106, 1053–1066.

Marquering, W., Verbeek, M., 2004. The economic value of predicting stock index returns and volatility. Journal of Financial and Quantitative Analysis 39, 407–429.

Pesaran, M.H., Timmermann, A., 1995. Predictability of stock returns: robustness and economic significance. Journal of Finance 50 (4), 1201–1228.

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

Politis, D., Romano, J.P., 1992. A circular block-resampling procedure for stationary data. In: Exploring the Limits of Bootstrap. John Wiley, New York, pp. 263–270.

Politis, D., White, H., 2004. Automatic block-length selection for the dependent bootstrap. Econometric Reviews 23, 53–70.

Raftery, A., Madigan, D., Hoeting, J., 1997. Bayesian Model Averaging for linear regression models. Journal of the American Statistical Association 97, 179–191.

Rapach, D., Zhou, G., 2012. Forecasting stock returns. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Vol. 2, Elsevier (forthcoming).

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

Sisson, S.A., 2005. Transdimensional Markov chains: a decade of progress and future perspectives. Journal of American Statistical Association 100, 1077–1089.

Stock, J., Watson, M.W., 2006. Forecasting with many predictors. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. North-Holland, Amsterdam, pp. 515–554.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B 58, 267–288.

White, H., 2001. Asymptotic Theory for Econometricians, Revised edition. Academic Press, San Diego.

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P.K., Zellner, A. (Eds.), Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti. North-Holland, Amsterdam, pp. 233–243.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society: Series B 67, 301–320.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics 37, 1733–1751.