Complete Subset Regressions*

Graham Elliott
UC San Diego

Antonio Gargano
Bocconi University, visiting UCSD

Allan Timmermann
UC San Diego

November 7, 2012

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.

Key words: subset regression, forecast combination, shrinkage.

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

1 Introduction

Methods for controlling estimation error in forecasting problems that involve small sample sizes and many potential predictor variables has been the subject of much recent research.[1] One lesson learned from this literature is that a strategy of including all possible variables is 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, Madigan, and Hoeting (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 corresponding to a given 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 be selected recursively in time.

Special cases of subset regression combinations have appeared in the empirical literature. For example, Rapach, Strauss and Zhou (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 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 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 also show that the weights implied by subset regression reflects 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 the usual application of 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

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

coefficients. Other approaches that apply flexible weighting to individual predictors include bagging 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) and 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, Griffin and Steel (2012)).

To illustrate the subset regression approach empirically we consider, like many previous studies, predictability of U.S. stock returns. In particular, following Rapach et al. (2010), we study quarterly data on U.S. stock returns in an application that has 12 potential predictor variables and so generates subset regressions with $k = 1, 2, ..., 12$ 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 regression model based on 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, i.e., the population value of the projection of y on X, where $y = (y_1, ..., y_T)$ is a $T \times 1$ vector and $X = (x_0', x_1', ..., x_{T-1}')'$ 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, zeros for the excluded 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 $n_{k,K} = K!/(k!(K-k)!)$ 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 $n_{1,K} = K!/(1!(K-1)! = K$ 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 i^{th} 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. 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 a weighted average of the full regression OLS estimator:

Theorem 1 *Assume that as the sample size gets large $\hat{\beta}_{OLS} \to^p \beta_0$ 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*

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

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 (2000, 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

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 heterogenous 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, 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, $\hat{\beta}_{k,K} = \lambda_{k,K}\hat{\beta}_{OLS}$, 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 $\hat{\beta}_{OLS} = X'y$ while each of the subset regression estimates can be written $\hat{\beta}_i = S_i X'y$ where S_i is a $K \times K$ diagonal vector with ones (zeros) on the diagonal for each included (excluded) regressor, and zeros off the diagonal. The complete subset regression estimator is then given by

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

The result now follows by noting that the elements of $\sum_{i=1}^{n_{k,K}} S_i$ are zero for the off-diagonals 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, Figure 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 is k relative to K, the greater the amount of shrinkage. Effectively the Theorem 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

[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 asymptocially.

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 $\hat{\beta}_{OLS}$. The weights depend not only on $\{k, K\}$ but on all elements in Σ_X. For example, if $K = 3$ and $k = 1$, we have

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

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 $\hat{\beta}_{OLS}$. Apart from the division by 1/3, the own coefficient is given a relative weight of one while the remaining coefficients are those we expect from omitted variable bias formulas. Clearly the effect of dividing by $n_{1,3} = 3$ is to shrink all coefficients, including the 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 $h-$periods 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:

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

$$\sigma_\varepsilon^{-2} E\left[T(\hat{\beta}_T - \beta)'\Sigma_X(\hat{\beta}_T - \beta)\right] \approx \sum_{j=1}^{K} \zeta_j + b'(\Lambda_{k,K} - I)'\Sigma_X(\Lambda_{k,K} - I)b, \tag{4}$$

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 are the elements of b, 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

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

which depends on $\{K, k, b'b\}$. For fixed values of $b'b$ and K, as k increases, $\lambda_{k,K}$ gets bigger 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 Figure 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 is $b'b$. 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, bigger 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 Figure 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 Figure 2 uses data from the application in the empirical section 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 the risk for OLS regression is K while for this case the risk of the subset regression method reduces to

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

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

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

For small values of K this holds for nearly all possible correlations. To illustrate this, Figure 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 $K \leq 6$, 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.

Figure 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 $b = 1$, 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 $\min(0, MSE^{ridge} - MSE^{subset})$, with dark red indicating that $MSE^{ridge} > MSE^{subset}$, while, conversely, yellow and blue indicate areas where $MSE^{ridge} < MSE^{subset}$. 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 regression. 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.

Figure 4 shows that when $k = 1$, ridge regressions mostly produce lower MSE-values than subset regressions for $\rho < 0.6$. Conversely, for $\rho > 0.85$, this univariate subset regression 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 $\rho < 0.5$. This means that, for $k = 2$, 3, or 4, one of the subset regressions will always dominate the best ridge regression as they produce the lowest MSE loss.

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 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 so 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 evaluating the models. However, methods exist that employ some of the model information to decide on inclusion without requiring statistics for all models to be computed.

Sophisticated MCMC algorithms developed in the Bayesian model combination and selection literature can be used, particularly 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 space. 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 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 also 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 computing 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, Paap and Ravazzolo (2012) 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, Casarin, Ravazzolo and van Dijk (2012) propose a variety of combination strategies that allow for time-varying weights, while Koop and Korobilis (2012) consider dynamic model averaging methods. While this issue 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:

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

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 $\Sigma_X = Cov(X_1, ..., X_K)$ takes the simple form

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

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 Bayesian model averaging (Raftery, Madigan and Hoeting, 1997), Bagging (Breiman, 1996), ridge regression (Hoerl and Kennard, 1970), and the Lasso (Tibshirani, 1996). 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 obtained by

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

where

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

Note that, by construction, as $\gamma \to \infty$, $\hat{y}_{T+1}^{RIDGE} \to \frac{1}{T-1} \sum_{j=2}^{T} y_j$, 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

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

where

$$\hat{\beta}_\psi = argmin \quad \left(\sum_{t=1}^{T-1} (y_{t+1} - x_t'\beta)^2 \right), \tag{11}$$
$$s.t. \sum_{j=1}^{K} |\beta_j| \leq \psi.$$

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 $|.|$, 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

In Tibshirami's classical implementation Lasso has a single shrinkage parameter. In this sense our approach is more similar to that of recent flexible generalizations such as the adaptive Lasso of Zou (2006) or the Elastic Net of Zou and Hastie (2005) and Zou and Zhang (2009). Following authors such as Korobilis (2013), we consider the Elastic Net which is a useful compromise between ridge and lasso. 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

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

where

$$\hat{\beta}_{\alpha,\psi} = argmin \left(\sum_{t=1}^{T-1} (y_{t+1} - x'_t\beta)^2 + \psi \underbrace{\sum_{j=1}^{K} (1-\alpha)\beta_j^2 + \alpha|\beta_j|}_{P_\alpha(\beta)} \right)$$

P_α, the elastic net penalization, is the compromise between 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, Hastie and Tibshirani (2009). 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 1,000 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 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 $c \in \{.3853, .6745, 1.2816, 1.4395, 1.6449, 1.9600, 2.2414, 2.5758, 2.8070, 3.0233, 3.2905, 3.4808, 3.8906, 4.4172, 5.3267\}$. The final Bagging forecasts are obtained by averaging across the bootstrap draws

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

3.2.5 Bayesian model averaging

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

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

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

where $\hat{b}_j$ is the posterior mean of the predictive likelihood 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 equations (15) and (16), we follow the specification suggested by Fernandez, Ley and Steel (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, $k = \sum_{i=1}^{K} \gamma_i$. If θ is kept fixed, k has a $Binomial(K, \theta)$ distribution with expected value $E[k] = m = \theta K$, 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 $s_2 = s$. 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, Ley and Steel (2001a), and adopt a combination of a "non-informative" improper prior on the common intercept and scale and a g-prior (Zellner, 1986) on the regression coefficients leading to the prior density $p(\alpha, \beta_j, \sigma|M_j) \propto \sigma^{-1} f_N^q(\beta|0, \sigma^2(gZ_j'Z_j)^{-1})$, 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 $b_j = \frac{1}{T}\sum_{i=1}^{T} y_i + z_j'\beta_j/(g+1)$.

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, something 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, Ley and Steel (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. 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 $\beta = 1_K$ (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

[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).

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 combination (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 Figure 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 $\beta = (1\ 1\ 1\ 1\ 0\ 0\ 0\ 0)'$, 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 to 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. 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 $M_k = [m_1, m_2, ..., m_{kC_K}]$, each model receiving a weight of c^{-1}.

The stochastic search algorithms select models according to a fitness function, $f(.)$ such as the model posterior. The included 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 $p = min\big(1, \frac{f(m^{t+1})}{f(m^t)}\big)$. If the candidate model offers a better fit ($f(m^{t+1}) > f(m^t)$), 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. However, 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 $p = min\big(1, \sum_{N_{m^{t+1}}} f(m_j) \; / \; \sum_{N_{m^t}} f(m_i)\big)$.

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 the three 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 alternative approaches, this section provides an empirical application to US stock returns. This application is well suited for our analysis in part because predictability of stock returns has been the subject of an extensive literature in finance, recently summarized by Rapach and Zhou (2012), in part 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, see, e.g., the discussion in Goyal and Welch (2008).

Specifically, we investigate if 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. Models with time-varying coefficients have been considered for stock market data by Kalli and Griffin (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. Tests of in-sample performance have higher power than out-of-sample tests and so from a purely statistical point of view have some advantages. 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; *Treasure 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 equity premium, 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 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

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

where $r_{2:t+1}$ is the equity premium variable 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, more specifically

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

[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.

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

4.2 Bias-variance trade-off

Figure 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 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.

Figure 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. Thus, 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 $x_t^S = x_t S$ and $X_{1:T}^S = X_{1:T} S$, and note that

$$var(\hat{r}_{t+1}) = var(\hat{\alpha} + x_t^S \hat{\beta}) = [\iota'\iota + x_t^S (X_{1:T}' X_{1:T})^{-1} x_t^{S\prime}]\hat{\sigma}_\epsilon, \tag{19}$$

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

The upper right window in Figure 6 shows gains from pooling the models due to the reduction in the (squared) bias. 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 Figure 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 Figure 6 establishes a (squared) bias-variance trade-off that resembles the well-known mean-variance efficient frontier known from modern portfolio theory in finance. Specifically, 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 insights into the forecast performance of the various models, Figure 7 plots the out-of-sample R^2 (top window) and 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 4,096 models, a benchmark frequently used in the forecast combination literature. 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), while the first six k-variate combinations produce better performance than the combination of all models, i.e., the "thick" forecast modelling approach described in Aiolfi & Favero (2003). 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.

Figure 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. Figure 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] This finding is not surprising given that about 60% of the models contain five, six or seven predictors so that the combinations get dominated by forecast models with five, six and seven variables included.[9] In fact, when models are randomly selected, the probability of picking a 6-variate model is about 0.225 against 0.002 for the univariate or eleven-variate models. Indeed the combinations of the six-variate models has very similar performance to the total combination.

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 as the single best k-variate model, the identity of which, however, 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.[10] 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

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

[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] This fraction is given by $\binom{12}{5} + \binom{12}{6} + \binom{12}{7}/2^{12}$.

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

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 $k = 12$.[11]

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 at which point 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 produces 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, Leybourne, and Newbold (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" so as 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 hyper parameter which would have given the best performance. Figure 9 shows the recursively chosen values for the hyperparameters. The subset 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

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

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 is very poor by this measure, as are 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. As can be seen, there is evidence of slight improvements in the out-of-sample R^2 values for some subsets, but the values are very similar under the two combination schemes. This suggests that 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%.

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 $V_t[r_{m,t+1}]$, 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

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

where γ 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, $V_t[r_{m,t+1}]$, 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.

The investor's ex-post realized utility is

$$u_{t+1} = r_{f,t+1} + \omega^*_t(r_{m,t+1} - r_{f,t+1}) - 0.5\gamma\omega^{*2}_t VOL_{t+1}. \tag{21}$$

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 which 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 translate 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 under 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 forecasting situations 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 can 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 U.S. 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.

6 Appendix

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

6.1 Proof of Theorem 1

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

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

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

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

$S_i'T^{-1}X'XS_i$ can be rearranged so that the upper $k \times k$ block is $T^{-1}X^{*\prime}X^*$ where X^* contains the k regressors included in the i^{th} regression. Since $T^{-1}X'X \to^p \Sigma_X$, then $T^{-1}X^{*\prime}X^* \to^p \Sigma_X^*$ (which is the variance covariance matrix of the included regressors) by the definition of convergence in probability for matrices. Rearranging the term $(S_i'T^{-1}X'XS_i)^- - (S_i'\Sigma_X S_i)^-$ 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. ■

6.2 Proof of Theorem 2

Proof. From the results of Theorem 1, we have

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

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

Now

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

and so

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

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

References

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

[2] Amihud, Y., 2002, Illiquidity and Stock Returns: Cross-section and Time-Series Effects. Journal of Financial Markets 5, 31-56.

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

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

[5] Billio, M., Casarin, R., Ravazzolo, F. and H.K. van Dijk, 2012, Combining Predictive Density Using Bayesian Filtering with Applications to US Economics Data. Ca' Foscari University of Venice working paper No. 16.

[6] Breiman, L., 1996, Bagging predictors. Machine Learning 36, 105-139.

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

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

[9] Clemen, R.T., 1989, Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559-581.

[10] Cremers, K., 2002. Stock return predictability: A Bayesian Model Selection perspective. Review of Financial Studies 15, 1223–1249.

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

[12] 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.

[13] Fernandez, C., E. Ley and F.J.J. Steel, 2001a, Benchmark Priors for Bayesian Model Averaging. Journal of Econometrics 100, 381-427.

[14] Fernandez, C., E. Ley and F.J.J. Steel, 2001b, Model Uncertainty in Cross-Country Growth Regressions. Journal of Applied Econometrics 16, 563-576.

[15] Geweke, J. and G. Amisano, 2011, Optimal Prediction Pools, Journal of Econometrics 164, 130-141.

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

[17] Groen, J.J., Paap, R., and F. Ravazzolo, 2012, Real-time Inflation Forecasting in a Changing World, Journal of Business and Economic Statistics, forthcoming.

[18] Hans, C., Dobra, A., and West, M., 2007, Shotgun Stochastic Search for Large p Regression, Journal of American Statistical Association 478, 507–516.

[19] Griffin, J.E. and M. Kalli, Time-varying Sparsity in Dynamic Regression Models.

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

[21] Hoerl, A.E., and R.W. Kennard, 1970, Ridge regression: Biased estimation for Nonorthogonal Problems, Technometrics 12, 55-67.

[22] Inoue, A., and L. Killian, 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.

[23] Koop, G., 2003, Bayesian Econometrics, NewYork: John Wiley.

[24] Koop, G. and D. Korobilis, 2012, Forecasting Inflation using Dynamic Model Averaging. International Economic Review 53(3), 867-886.

[25] Korobilis, D., 2013, Hierarchical Shrinkage Priors for Dynamic Regressions with Many Predictors. International Journal of Forecasting 29, 43-59.

[26] Lamnisos, D., J.E. Griffin, and M.F.J. Steel, 2012, Adaptive Monte Carlo for Bayesian Variable Selection in Regression Models. Forthcoming in Journal of Computational and Graphical Statistics.

[27] Ley, E. and M.F.J. Steel, 2009, On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression. Journal of Applied Econometrics 24, 651-674.

[28] Ley, E. and M.F.J. Steel, 2012, Mixtures of g-Priors for Bayesian Model Averaging with Economic Applications. Journal of Econometrics, forthcoming.

[29] Liang, H., G. Zou, A.T.K. Wan, and X. Zhang, 2011, Optimal Weight Choice for Frequentist Model Average Estimators, Journal of the American Statistical Association 106, 1053-1066.

[30] Pettenuzzo, D., and A. Timmermann, 2011, Predictability of Stock Returns and Asset Allocation under Structural Breaks. Journal of Econometrics 164, 60-78.

[31] Politis, D., and J.P. Romano, 1992, A circular block-resampling procedure for stationary data, in Exploring the limits of bootstrap, New York: John Wiley, 263-270

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

[33] Raftery, A., D. Madigan, and J. Hoeting, 1997, Bayesian model averaging for linear regression models. Journal of the American Statistical Association 97, 179-191.

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

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

[36] Stock, J., and M.W. Watson, 2006, Forecasting with many predictors. Pages 515-554 in Elliott, G., C.W.J. Granger, and A. Timmermann (eds.) Handbook of Economic Forecasting. North Holland.

[37] Tibshirani, R., 1996, Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society B 58, 267-288.

[38] White, H., 2000, Asymptotic Theory for Econometricians, revised edition. Academic Press, San Diego.

[39] 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, p. 233–243.

[40] Zou, H., 2006, The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101, 1418-1429.

[41] Zou, H. and T. Hastie, 2005, Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society: Series B, 67, 301-320.

[42] Zou, H. and H.H. Zhang, 2009, On the Adaptive Elastic-Net with a Diverging Number of Parameters, Annals of Statistics 37, 1733-1751