Marchese *et al. Financial Innovation* (2024) 10:138
https://doi.org/10.1186/s40854-024-00657-9

Financial Innovation

METHODOLOGY Open Access

Robustifying and simplifying high-dimensional regression with applications to yearly stock return and telematics data

Malvina Marchese[1], María Dolores Martínez-Miranda[2*], Jens Perch Nielsen[1] and Michael Scholz[3,4]

*Correspondence: mmiranda@ugr.es

[1] Bayes (formerly Cass) Business School, City, University of London, 106 Bunhill Row, London EC1Y 8TZ, UK
[2] Department of Statistics and Operations Research, University of Granada, Campus Fuentenueva, 18071 Granada, Spain
[3] Department of Economics, University of Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt, Austria
[4] JOANNEUM RESEARCH Forschungsgesellschaft mbH, Leonhardstraße 59, 8010 Graz, Austria

Abstract

The availability of many variables with predictive power makes their selection in a regression context difficult. This study considers robust and understandable low-dimensional estimators as building blocks to improve overall predictive power by optimally combining these building blocks. Our new algorithm is based on generalized cross-validation and builds a predictive model step-by-step from a simple mean to more complex predictive combinations. Empirical applications to annual financial returns and actuarial telematics data show its usefulness in the financial and insurance industries.

Keywords: Forecasting, Non-linear prediction, Stock returns, Dimension reduction, Telematics

JEL Classification: C14, C53, C58, G17, G22

Introduction

When selecting a particular submodel for a high-dimensional statistical problem, the noise to signal ratio must be monitored. The challenge is to obtain the optimal bias reduction for any additional noise and stop adding more noise at the right time (see, among many others, the popular book of James et al. 2013). In this study, we introduce a simple approach for optimizing the model selection process in high-dimensional problems with limited data. We consider a situation with several potentially useful covariates and limited data to estimate a complicated nonlinear underlying model. We propose building a potentially complex nonparametric model structure using simple low-dimensional components. With many available covariates, one can build many one- or two-dimensional models and then construct an optimal weighted average for these submodels. In principle, this is different from linear regression, in which the linear components involved can only be interpreted together with the entire linear model. We propose a forward model-selection method that begins with a simple mean. Through cross-validation, we then examine the noise-to-signal ratio when replacing a fraction of that mean with each of our many low-dimensional competing submodels. We select the best replacement, as measured via cross-validation,

including the possibility of replacing with the mean itself (such that nothing is happening). If the mean itself is picked, then we stop the procedure. However, if any other submodel is picked, we continue the methodology of replacing a fraction of the mean with a fraction of other submodels until the method stops because the mean itself has been selected and nothing is happening. This method is described in detail in the following section. Our approach is an alternative to forward or backward model selection methods in linear regression (e.g., Steinberger and Leeb 2019) or robust linear regression via penalization (e.g., Filzmoser and Nordhausen 2021) and can be used for both standard and time-series regression problems. In the context of time-series regression, our method is similar to an ensemble method that combines alternative forecasts in an optimal manner. Our method is different from most other non-parametric approaches because it does not assume an overall complicated model structure based on all covariates, such as linear regression, generalized additive models, or very general generalized structured models (see Mammen and Nielsen 2003). Our final estimator is always a combination of estimators that are already relevant and useful estimators. Therefore, our final estimator is easier to understand. It might also be more robust because it does not add noise from estimating residuals, as linear regression or generalized additive models—or other big model methods—are designed to do.

Consider K possible candidates $\hat{m}_1, \ldots, \hat{m}_K$ when predicting Y (from low-dimensional predictive models). Then, the combined forecast is

$$\hat{Y} = \sum_{k=1}^{K} w_k \hat{m}_k$$

with weights $w_1, \ldots, w_K$ being chosen appropriately. We propose a step-wise (forward) procedure starting from the historical mean and adding "small" bits to the final model based on improvements of our validation criterion. Our new algorithm for prediction is based only on low-dimensional functions, and can therefore be used in high-dimensional situations where the number of observations is low compared to the number of covariates. We consider our method to be a simple interpretable and robust alternative to popular regularization and shrinkage methods such as Lasso (Tibshirani 1996), "adaptive" Lasso (Zou 2006), or ridge regression (Hoerl and Kennard 1970), hyperparameter optimization (Bischl et al. 2023), among many others also aiming at the challenge of analyzing sparse data with many covariates.

In "Materials and methods" section, we formalize our approach in a general manner that applies to both standard regression data and time-series regressions. Our validation strategy can be defined according to the nature of the data, with leave-one-out cross-validation being the simplest and most common choice for non-autocorrelated data. When adding low-dimensional components to the final predictor, we also consider limiting the number of covariates used to reduce the noise. In "An empirical application to time series data" section we provide a real data example for a time series regression for forecasting yearly stock returns. Our application adds to the results of previous studies by Nielsen and Sperlich (2003) and Kyriakou et al. (2020, 2021a, 2021b) and obtain better prediction accuracy. In "An empirical application

to cross-sectional actuarial telematics data" section we present a real-data regression example for an actuarial telematics dataset. We present a finite-sample simulation study in "Simulation studies" section where we consider two standard regression cases without any time-series dependency. The case we consider has only 200 observations and ten covariates and it turns out that our new approach performs as well for this case as linear regression, and performs better than ridge regression and Lasso, even when the linear regression is true. When the linear regression is not true and there is nonlinearity, our new simple non-parametric ensemble method performs much better than linear, Lasso, and ridge regression. This is also in line with the results of Scholz (2022), who show that Lasso and ridge regressions do not perform well for the nonlinear yearly stock prediction problem that we consider as our real-life data example in "An empirical application to time series data" section. The conclusions of this study, including a discussion on possible extensions, are presented in "Conclusion" section.

Materials and methods

This section presents the mathematical formulation and implementation of the proposed approach. The "Prediction framework" section begins with a general definition of loss measures incorporating both the loss measure used in this study and the popular penalty loss measures that the Lasso and ridge regressions—which we compare our new method to in "Simulation studies" section—are based on.

Prediction framework

We assume the general prediction framework described in Hastie et al. (2017) as follows:

$$\min_{f\in\mathcal{H}}\left\{L\big(y_{t+h},f(Z_t)\big)+p(f,\tau)\right\},\quad t=1,\ldots,T, \tag{1}$$

where y_{t+h} is the variable to be predicted h periods ahead, Z_t is the set of (all available) D covariates, $\mathcal{H}$ is a space of possible functions f that combine the data to form the prediction, p is a penalty on f, τ is a set of hyperparameters (for example, λ in the Lasso, typically chosen via a version of cross-validation), and L is a loss function that defines the optimal forecast. Typical loss functions are the L_2 and L_1 norms.

When choosing a predictor in a high-dimensional situation where the number of covariates D is large, the sparsity of the data becomes a crucial problem. Therefore, it is necessary to avoid the estimation of high-dimensional objects and their inherent variability. Our proposal is a forward algorithm, where at each step, combinations of only low-dimensional (and potentially nonlinear) functions, $f_k \in \mathcal{H}$ $(1 \le k \le K)$, are considered based on a subset of d_k covariates, $Z_{k,t} \subseteq Z_t$. For example, in the data application we describe later, we have seven potential covariates; however, we only consider one- and two-dimensional functions based on them.

Our proposal

Our proposal consists of three elements: low-dimensional predictors, a method for evaluating their predictive power, and a criterion for defining combinations of low-dimensional predictors with good predictive power. The reason for this is to introduce

robustness into the forecasting approach. We want to avoid estimators such as the Lasso or Elastic Net (Rapach and Zhou 2020), or even linear regression, where entering many covariates may introduce noise that is difficult to handle in the validation analyses. We use the same validation measure (out-of-sample R^2) as Rapach and Zhou (2020), Yae and Luo (2023), and Kelly et al. (2022). However, although these three studies cite Campbell and Thompson (2008) for the introduction of this measure, we cite Nielsen and Sperlich (2003) (see also McGibney and Smith 1993; Anh et al. 2017). We know from Scholz (2022) that machine learning methods, such as Lasso and Elastic Net, do not help improve forecasting when combined with our simple low-dimensional forecasting candidates. Our combined forecast out-of-sample R^2 values generally seem higher than the out-of-sample R^2 values obtained in Rapach and Zhou (2020) and Kelly et al. (2022). Although these validation scores are not directly comparable because the datasets used are not exactly the same, they do indicate that combining low-dimensional forecasts might be the most efficient way forward. In addition, the simple model selection procedure of this study may compete very well with more complicated machine learning combinations, whether these machine learning techniques are applied directly to the covariates, as in Rapach and Zhou (2020) and Kelly et al. (2022) or to the low-dimensional forecasts suggested in this study, see Scholz (2022). See also Leeb and Steinberger (2021) for another approach to simple linear models used directly on the covariates when forecasting stock returns and Anatolyev (2019) for a simple and friendly guide to maintaining simplicity in regression studies with complex data. We do not think that the promising results of this study can be attributed to p-hacking (Brodeur et al. 2020) because we select simplicity over complexity, and we are very careful in our validation technique. It could be interesting, but beyond the scope of this study, to generalize our approach to allow for a hierarchical structure of combined forecasts, as in Spiliotis et al. (2021), or to robustify the output of the methodology by combining quantile regression rather than standard regression when forecasting, as in Belloni et al. (2019).

For the first element we formulate K predictive regression models, based on $d_k << D$ $(k = 1, \ldots, K)$ covariates, of the type

$$y_{t+h} = f_k(Z_{k,t}) + \xi_{k,t}, \tag{2}$$

where

$$f_k(z) = \mathbb{E}(y_{t+h} \,|\, Z_{k,t} = z_{k,t}),\ z_{k,t} \in \mathbb{R}^{d_k}, \tag{3}$$

is an unknown function and $\xi_{k,t}$ is an error term. The predictors $\hat{f}_k$ for the unconditional means f_k can be computed by assuming a linear structure using ordinary least squares (OLS) or more flexible nonparametric techniques such as kernel smoothing (Wand and Jones 1994). It is important that these underlying low-dimensional objects, the estimated f_k's, fit the job at hand. In real data problems, both linearities and nonlinearities should be considered. Nonparametric smoothing methods, such as the popular local linear estimation, can adapt to both situations; therefore, in our empirical studies, we consider local linear smoothers with an optimal data-driven bandwidth choice. Other types of non-parametric estimators of the underlying f_k are also possible. One simple

approach would be to allow for local neighborhood bandwidths rather than the constant bandwidths used in this study. Many other nonparametric, semiparametric, or even parametric choices of f_k are possible, as long as they fit the problem at hand.

To evaluate the predictive power of a model of type (2), we consider the validated R-squared of Nielsen and Sperlich (2003) (see also Kyriakou et al. 2021b) defined as

$$R_V^2 = 1 - \frac{L(Y,\hat{f}_{-t})}{L(Y,\bar{Y}_{-t})}, \tag{4}$$

with $Y = (y_{1+h},\ldots,y_T)^\top$, loss function L, and where $\hat{f}_{-t}$ and $\bar{Y}_{-t}$ are the estimators of the conditional mean function f_k (2) and the unconditional historical mean of Y, respectively, computed without the information contained in Y_t. This involves defining a general validation set, which we estimate with all but the t observation and the $2l$ observations around it. Leave-one-out cross-validation or the more general K-fold cross-validation are common choices for models with uncorrelated errors. For time series forecasting, and because of inherent serial correlation, forecasters tend to prefer out-of-sample evaluation. One problem with out-of-sample evaluation is that it only evaluates once, whereas cross-validation involves several evaluations, which may be more convenient, especially for small sample sizes. Other approaches for the case of correlated errors include cross-validation excluding $(2l+1)$ observations (with $l > 1$) as defined above, as well as other versions, such as the so-called h-block cross-validation (h observations preceding and following the observation are omitted from the test set). However, most of these time-series alternatives have problems that require additional corrections. Hence, a simple leave-one-out cross-validation may be a better option in many practical situations. See Bergmeir et al. (2018) for a discussion on the validity of cross-validation for autoregressive time-series predictions and some recommendations in practice.

The validated R-squared value in (4) measures the predictive power of a given model against a benchmark, that is, the historical mean (a classic benchmark for financial time series). The positive values of R_V^2 for a given predictor $\hat{f}$ indicate that it outperforms the corresponding historical mean forecast. Considering this, our algorithm chooses predictors that maximize R_V^2, which is equivalent to the minimization of the loss function $L(Y,\hat{f}_{-t})$.

The last element of our proposal is a combination of predictors to increase predictive power. For predictors $\hat{f}_k$ $(k = 1,\ldots,K)$, we define a combination of types

$$\hat{f}^{comb} = \sum_{k=1}^{K} w_k \hat{f}_k, \tag{5}$$

for certain weights $w_1,\ldots,w_K$. It is well known in forecasting literature that forecast combinations often lead to better forecast accuracy (Clemen 1989). These and other methods were recently analyzed by Scholz (2022), who show that single predictors can perform better in terms of the validated R-squared than the combination $\hat{f}^{comb}$.

Based on the previous definitions, we propose an algorithm to construct optimal combinations of predictors. Starting with the historical mean $\bar{Y}$, we combine it linearly with a portion of $\hat{f}_i$. The idea is to think like an investor; that is, using the most promising candidates (the models with $R_V^2 > 0$), allowing for leverage (weights could sum up to a

value larger than one), use only small "bits" of the candidate at hand (say a 10% weight), validate its impact immediately and discard it if no further improvement in predictive power is achieved. For a fixed $\alpha \in (0, 1)$ (e.g., $\alpha = 0.1$), we calculate

$$\hat{f}_{FW} = \bar{Y} + \alpha(\hat{f}_i - \bar{Y}) \tag{6}$$

where $i \in \{1, \ldots, K\}$ is selected such that the validated R-squared of $\hat{f}_{FW}$ is the maximal. Then, we iterate in the same manner as long as the predictive power improves,

$$\hat{f}_{FW}^{new} = \hat{f}_{FW}^{old} + \alpha(\hat{f}_i - \bar{Y}) \tag{7}$$

where $i \in \{1, \ldots, K\}$ is chosen again such that the validated R-squared of $\hat{f}_{FW}^{new}$ is the maximal.

Based on the definitions above, our forward Algorithm 1 is described by the following steps.

Algorithm 1 **Require:** A validation criterion, $\alpha \in (0, 1)$ and the historical mean $\bar{Y}$.

Step 1: **Provide** K predictors $\hat{f}_i$, $1 \leq i \leq K$, based on low-dimensional sets of covariates $Z_{k,t}$. If $\hat{f}_i$ does not have a better validation than $\bar{Y}$, **stop**, and **return** $\bar{Y}$.

Step 2: **Construct** $\hat{f}_{FW} = \bar{Y} + \alpha(\hat{f}_i - \bar{Y})$ with i such that $\hat{f}_{FW}$ has the best validation with respect to the chosen criterion. If $\hat{f}_{FW}$ does not improve compared with the best predictor $\hat{f}_s$, then **stop** and **return** $\hat{f}_s$.

Step 3: **Construct** $\hat{f}_{FW}^{new} = \hat{f}_{FW}^{old} + \alpha(\hat{f}_i - \bar{Y})$ with i such that $\hat{f}_{FW}^{new}$ has the best validation with respect to the chosen criterion.

Step 4: **Repeat** Step 3 as long as the validation of $\hat{f}_{FW}^{new}$ improves. **Return** $\hat{f}_{FW}^{new}$.

To simplify the model choice, it may be appropriate to limit the number of candidates. Thus, in the practical application and simulation study, we also include a variant of Algorithm 1, in which the use of five maximal predictors $\hat{f}_i$ is allowed. We denote the predictor based on Algorithm 1 as **forward** and the variant as **forward5**.

Results and discussion

This section provides a finite sample simulation study in "Simulation studies" section, empirical applications to time-series data in "An empirical application to time series data" section, and cross-sectional data in "An empirical application to cross-sectional actuarial telematics data" section. The finite-sample study in "Simulation studies" section shows that linear, Lasso, and ridge regression are competitive with our new method when the comparison is on their home turf: the linear model is true. However, linear regression, Lasso, and ridge regression are not competitive with our new method when the linear model is not true. Therefore, it is clear that our new method is superior in performance to linear, Lasso, and ridge regression. The first empirical application revisits a well-studied dataset of Robert Shiller that is often used for forecasting yearly stock returns. It is a time series of more than 100 yearly data points with a number of relevant covariates, such as dividend yield, earnings, inflation, and interest rates. This dataset is in line with the objective of our study, in which we want to consider small datasets with many covariates. The combination of covariates results in 28 low-dimensional

competing predictors that could all be used for forecasting. Our new approach uses only three of these predictors and the mean to forecast yearly stock returns. The second empirical application uses telematics data on individual driving patterns to predict traffic accident claims costs, which is a major problem in motor insurance. The dataset consists of around 500 individuals (of age 18–35) and includes information on the policyholder (like age, age of car, or years holding a driver's license) and the policyholder's driving style and driving patterns ("telematic covariates" like annual distance driven or percentage of kilometers driven above the speed limit or at night). The combination of available covariates again results in 28 low-dimensional competing predictors that could potentially be used to predict traffic accident claims costs. With our new approach, we use only five of these predictors and the mean.

Simulation studies

In this section, we compare the performances of the **forward** and **forward5** methods proposed in Algorithm 1 with those of the five related methods on the simulated data. The first method is standard linear regression based on OLS, denoted by **linreg**. The second and third methods are regularization techniques that ideally shrink the weights w_k of the predictors $\hat{f}_k$ without predictive power to zero. We use only two of the many different approaches and their variants: Lasso (Tibshirani 1996) and ridge regression (Hoerl and Kennard 1970), denoted by **lasso** and **ridge**. To compute the Lasso and ridge regressions, we use the `R` package `glmnet`. The fourth method is a fully nonparametric local-linear smoother based on the quartic (product) kernel, whose bandwidths are chosen with cross-validation, as described in "Materials and methods" section. We denote this method by **loclin**. The final method is a simple average $\bar{y}$ denoted by the **mean**. In all these methods, the full set of available covariates is used in the prediction procedure.

We consider two cases: a linear data-generating process (dgp) and a nonlinear process. Both simulations are similar in the choice of a maximum of ten covariates, of which four are relevant and six are irrelevant. We evaluate the predictions based on two measures: (i) the cross-validation (CV) as discussed in "Materials and methods" section, that is, $L(y,\hat{f}_{-t}) = ||y - \hat{f}_{-t}||_2^2$ and (ii) the out-of-sample mean squared forecast error (MSFE), that is, $L(y^{oos},\hat{f}_{-t}^{oos}) = ||y^{oos} - \hat{f}_{-t}^{oos}||_2^2$ based on additional observations from the same dgp that have not been used in the estimation step. In each case, we generate $T = 200$ observations for the estimation and $T^{oos} = 50$ observations for the out-of-sample validation in a total of 500 iterations.

Case 1

We generate data using the model:

$$y = 1.0 + 0.8x_1 + 0.6x_2 + 0.4x_3 + 0.2x_4 + \varepsilon$$

with $\varepsilon, x_i \sim \mathcal{N}(0,1), i = 1,\ldots,10$. As the model is linear in this case, we use OLS for the estimation in **forward** and **forward5** to reduce the computational burden. Note that the dgp is linear, such that using a local-linear smoother would also estimate the true linear model without any bias, and thus deliver similar results.

Box plots of the CV (left) and MSFE (right) are shown in Fig. 1. The median CV for the **forward5** is clearly smaller than that of the other competitors. The second is **forward**, followed by **linreg**. For MSFE, the performance is similar for our **forward** and

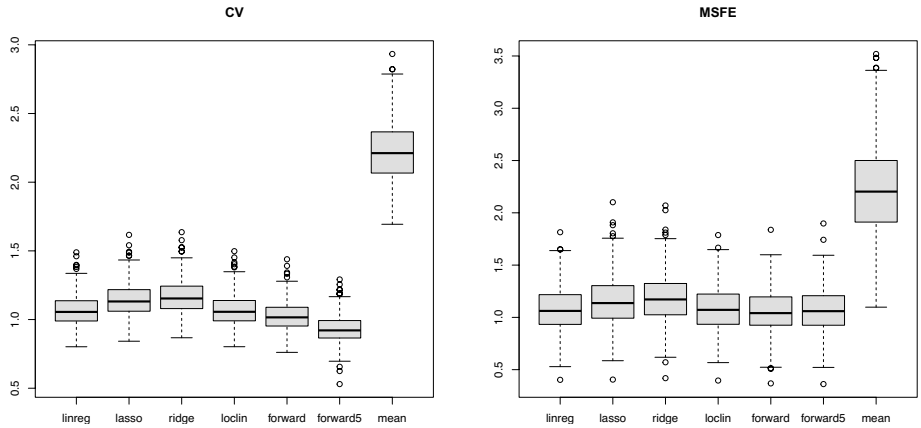

Fig. 1 Results of the simulation study for the linear data generating process of Case 1. Left: CV, right: MSFE

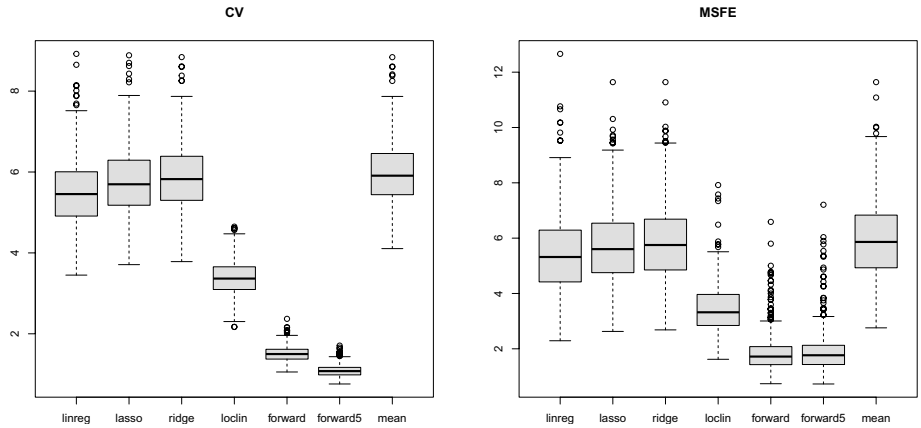

Fig. 2 Results of the simulation study for the non-linear data generating process of Case 2. Left: CV, right: MSFE

forward5 methods and the standard local linear regression. As expected, **loclin** and **linreg** perform comparably well. Interestingly, both shrinkage methods perform worse than the other model-based approaches. The **mean** produces the worst results. For the MSFE measure, **forward5**, **forward**, **loclin**, and **linreg** perform similarly. **Lasso** and **ridge** are again slightly worse. Again, the **mean** is the worst predictor under the considered approaches.

Case 2

In this case, the dgp is given by:

$$y = -x_1^2 - 2\sin\left(\frac{\pi}{2}x_2\right) + x_3x_4 + \varepsilon$$

where $\varepsilon, x_i \sim \mathcal{N}(0,1), i = 1,\ldots,10$. Note again that only the variables $x_1,\ldots,x_4$ are used; that is, there are six irrelevant variables. In **forward** and **forward5**, we now use a local linear smoother instead of OLS as in Case 1.

Box plots of the CV (left) and MSFE (right) are shown in Fig. 2. **forward5** again performs best in terms of the lowest value in the CV measure followed by **forward**. In

addition, the fully non-parametric approach, which is usually ideally suited for such nonlinear dgp's is outperformed by our new approaches. The **linreg**, **lasso**, and **ridge** approaches cannot account for the nonlinearities in the dgp and perform much worse than the other model-based approaches. Even the **mean** shows CV values similar to those of the purely linear methods. When considering the MSFE, a similar ranking of the methods can be obtained. Again, **forward** and **foreward5** outperform their competitors.

We have carried out simulations in dimensions up to 200 for our new method and allowed it to compete with linear, Lasso, and ridge regression on the home turf of these three methods: when the problem is linear. Our proposed method outperforms the others—even when the true model was indeed linear—when the measure is cross-validation, which is the most used measure of machine learning methods. However, if we also look at the out-of-sample performance MSFE, that is, we simulate a sample to estimate a model and use another simulated sample for evaluation; then, it is not clear whether the Lasso, ridge, or our proposal would be preferable. The simulations yield mixed messages. Linear regression without variable reduction is not competitive for such high dimensions, even when estimating the true linear regression model. The noise of the standard linear regression in high dimensions is too large. What is clear is that our new method is very competitive with Lasso and ridge, even on their home turf. It should also be added that, unsurprisingly, when nonlinearities are present, Lasso and ridge are not competitive with our proposed method.

An empirical application to time series data

We apply the methods described in "Materials and methods" section to annual US stock market data from 1872 to 2019. We use a revised and updated version of the series described in Shiller's Chapter 26 (Shiller 1989) which consists of the Standard and Poor's (S &P) Composite Stock Price Index, dividends and earnings accruing to the index, a 1-year interest rate, a long government bond yield, and the consumer price index (http://www.econ.yale.edu/~shiller/data.htm). We replace the original risk-free rate series (which was discontinued in 2013) with an annual yield based on the 6-month Treasury bill rate (https://fred.stlouisfed.org/series/TB6MS) secondary market. This new series is only available from 1958 onwards. Therefore, we regress the Treasury bill rate on the original commercial paper rate from Shiller's data and instrument the risk-free rate from

Table 1 US market data (1872–2019)

	Max	Min	Mean	Sd	Skew	Exc.kurt
S &P stock price index	2789.80	3.25	277.58	558.13	2.43	5.50
Dividend accruing to index	53.75	0.18	6.04	10.56	2.45	6.00
Earnings accruing to index	132.39	0.16	13.96	26.31	2.43	5.35
Dividend-by-price	9.88	1.17	4.31	1.71	0.46	0.25
Earnings-by-price	17.75	1.72	7.28	2.75	1.05	1.39
Short-term interest rate	14.93	0.07	3.97	2.50	0.96	2.34
Long-term interest rate	14.59	1.88	4.53	2.27	1.81	3.63
Inflation	20.69	− 15.65	2.23	5.96	0.26	1.60
Spread	3.64	− 3.71	0.56	1.32	− 0.05	0.02

1872 to 1957 with the corresponding predicted values. For further details, see Kyriakou et al. (2020) and Mammen et al. (2019).

Our analysis focuses on the nonlinear predictive relationships between 1-year stock returns in excess of a reference rate (or benchmark) and a set of explanatory variables. We use the data analyzed recently by Kyriakou et al. (2021a). Table 1 summarizes the available variables using their basic descriptive statistics. In our analysis, we focus on 1-year-ahead forecasts, but we can also consider longer horizons, $T > 1$; however, in that case, we should account for overlapping observations and related econometric problems (see Kyriakou et al. 2020 for more details). The approach considered here estimates a fixed forecast function for dynamic time-varying covariates. The time-varying forecasting dynamics of our approach are derived only from the time-varying covariates. This differs from approaches such as Fan et al. (2022), in which time dependency is modelled explicitly and applied to portfolio optimization.

We investigate stock returns $S_t = (P_t + D_t)/P_{t-1}$, $t = 1, \ldots, T$, where D_t denotes the (nominal) dividends paid during year t and P_t is the (nominal) stock price at the end of year t, in excess (log scale) of the given benchmark $B_{t-1}^{(A)}$

$$y_t = \ln \frac{S_t}{B_{t-1}^{(A)}}, \tag{8}$$

where $A \in \{R, L, E, C\}$ with, respectively,

$$B_t^{(R)} = 1 + \frac{R_t}{100}, \quad B_t^{(L)} = 1 + \frac{L_t}{100}, \quad B_t^{(E)} = 1 + \frac{E_t}{P_t}, \quad B_t^{(C)} = \frac{CPI_t}{CPI_{t-1}}, \tag{9}$$

R_t is the short-term interest rate and L_t is the long-term interest rate rate, E_t is earnings accrued to the index in year t, and CPI_t the consumer price index for year t. We are interested in 1-year-ahead forecasts y_t, that is, $h = 1$, via Eq. (2) using low-dimensional combinations of popular time-lagged predictive variables $Z_{t-1} \in \mathbb{R}^q$ and $1 \leq q \leq 2$, including the: (1) dividend-by-price ratio $d_{t-1} = D_{t-1}/P_{t-1}$; (2) earnings-by-price ratio $e_{t-1} = E_{t-1}/P_{t-1}$; (3) short-term interest rate $r_{t-1} = R_{t-1}/100$ (4) long-term interest rate $l_{t-1} = L_{t-1}/100$, (5) inflation $\pi_{t-1} = (CPI_{t-1} - CPI_{t-2})/CPI_{t-2}$; (6) term spread $s_{t-1} = l_{t-1} - r_{t-1}$; and vii) excess stock return y_{t-1}. This gives us $K = 28$ different (potentially highly correlated) predictors, $\hat{f}_1, \ldots, \hat{f}_K$ that can be ranked based on their predictive power measured by the validated R-squared in (4).

Table 2 Predictive power (in percent) of the 28 low-dimensional models

Model		R_V^2	Model		R_V^2	Model		R_V^2	Model		R_V^2
1	y	− 1.4	8	y, d	− 1.6	15	d, r	0.8	22	e, s	5.4
2	d	− 0.2	9	y, e	− 3.4	16	d, l	− 1.2	23	r, l	5.2
3	e	− 1.5	10	y, r	− 0.9	17	d, π	9.5	24	r, π	9.5
4	r	0.8	11	y, l	− 2.5	18	d, s	7.9	25	r, s	7.4
5	l	− 0.8	12	y, π	9.7	19	e, r	− 0.4	26	l, π	9.9
6	π	10.3	13	y, s	4.8	20	e, l	− 2.5	27	l, s	7.4
7	s	7.2	14	d, e	− 1.9	21	e, π	10.9	28	π, s	**15.4**

Model with best predictive power highlighted in bold

As stated above, we focus on all one- and two-dimensional non-parametric models for predicting real excess stock returns (using the benchmark $B^{(C)}$), which gives us 28 different predictors $\hat{f}_i, 1 \leq i \leq 28$, all estimated with the local linear smoother using the quartic (product) kernel. The smoothing parameters (bandwidths) were chosen by leave-one-out cross-validation, that is, by maximizing the in-sample performance measure R_V^2 introduced in "Materials and methods" section. Table 2 lists the predictive powers of the candidates. For each candidate, we indicate the variables involved (one or two variables maximum) and the value of the R_V^2 that provides the predictive power of the candidate model. We can see that the best candidate model has an R_V^2 of 15.4% and is based on the inflation rate (π) and term spread (s) as predictive variables.

Using our new algorithm and setting $\alpha = 0.1$, the predictive power can be increased by 8% to $R_V^2 = 16.6$. The single rounds are listed in Table 3. The first round corresponds to steps 1–2 of our forward algorithm (see Algorithm 1 in "Our proposal" section) and only considers the model candidate with the best predicted power, that is a two-dimensional model based on π and s as predictive variables (see Table 2). Step 3 of the algorithm is iterated until no improvement is achieved. The optimal model (28) is chosen in the first seven consecutive rounds, reaching almost 100 % predictive power with this 70 % input. In rounds 8 and 11, model 21 is selected, followed by model 18 in rounds 9, 10, and 12. The final forecast is given by (7), which yields the following combination of predictors

$$\hat{f}_{FW} = -0.2\bar{Y} + 0.3\hat{f}_{18} + 0.2\hat{f}_{21} + 0.7\hat{f}_{28}$$

where 12 rounds are used and the model weights sum to 1.2. Note that already the model produced in round 9 (using a total weight of 0.9) has a larger predictive power than the best individual model.

An empirical application to cross-sectional actuarial telematics data

We apply the new method described in "Materials and methods" section to a cross-sectional dataset obtained from a Spanish insurance company (Bolancé et al. 2022). It consists of 488 car insurance policyholders who reported at least one claim in 2011,

Table 3 Predictive power (in percent) of the round-based forward algorithm fixing $\alpha = 0.1$

Round	Model	Covariates	R_V^2
1	28	π, s	3.1
2	28	π, s	5.9
3	28	π, s	8.4
4	28	π, s	10.5
5	28	π, s	12.2
6	28	π, s	13.5
7	28	π, s	14.5
8	21	e, π	15.2
9	18	d, s	**15.8**
10	18	d, s	**16.2**
11	21	e, π	**16.5**
12	18	d, s	**16.6**

Steps in Algorithm 1 with improved predictive power compared to the single best predictor $\hat{f}_{28}$ highlighted in bold

Table 4 Spanish claim costs data

	Max	Min	Mean	Sd	Skew	Exc. kurt
co	17.03	0.02	1.55	2.03	3.36	18.27
ag	34.07	20.59	27.01	3.25	0.09	1.99
ad	14.69	2.00	6.44	2.83	0.76	2.95
ac	20.47	2.11	8.90	4.15	0.79	3.11
tk	35.10	1.22	8.36	4.53	1.30	6.27
nk	42.83	0.04	7.52	6.51	1.86	7.84
uk	80.66	3.81	27.07	14.12	0.96	3.92
sk	48.00	0.12	7.21	7.10	1.90	7.47

Table 5 Predictive power (in percent) of the 28 low-dimensional models

Model		R_V^2	Model		R_V^2	Model		R_V^2	Model		R_V^2
1	*ag*	− 0.33	8	*ag, ad*	− 0.64	15	*ad, tk*	− 0.60	22	*ac, sk*	− 0.52
2	*ad*	− 0.20	9	*ag, ac*	− 0.78	16	*ad, nk*	0.08	23	*tk, nk*	− 0.26
3	*ac*	− 0.41	10	*ag, tk*	− 0.72	17	*ad, uk*	− 0.83	24	*tk, uk*	− 0.83
4	*tk*	− 0.18	11	*ag, nk*	− 0.42	18	*ad, sk*	− 0.41	25	*tk, sk*	− 0.45
5	*nk*	**0.15**	12	*ag, uk*	− 0.88	19	*ac, tk*	− 0.81	26	*nk, uk*	− 0.65
6	*uk*	− 0.51	13	*ag, sk*	− 0.30	20	*ac, nk*	− 0.45	27	*nk, sk*	− 0.25
7	*sk*	− 0.04	14	*ad, ac*	− 0.70	21	*ac, uk*	− 0.98	28	*uk, sk*	− 0.58

Model with best predictive power highlighted in bold

covering, among others, the following information: cost per policyholder in thousands of euros (*co*), age in years (*ag*), number of years holding a driver's license (*ad*), age of car in years (*ac*), annual distance driven in thousands of kilometers (*tk*), percentage of kilometers driven at night (*nk*), percentage of kilometers driven on urban roads (*uk*), and percentage of kilometers driven above the speed limit (*sk*). We use the same data analyzed recently by Bolancé et al. (2023) (see also Bolancé et al. and 2018). Table 4 lists the variables and their basic descriptive statistics.

We analyze the logarithm of claim costs, $y = \log(co)$, and focus on all one- and two-dimensional models. This provides 28 different predictors $\hat{f}_i$ and $1 \leq i \leq 28$, all estimated with the local linear smoother using the quartic (product) kernel. Smoothing parameters (bandwidths) were chosen using leave-one-out cross-validation. Table 5 summarizes the predictive power of the models as measured by the R_V^2. Only 2 of the 28 models show a validated R-squared larger than zero, indicating predictive power. The best individual model has an R_V^2 of 0.15% and is based on the percentage of kilometers driven at night. Using our new algorithm (setting $\alpha = 0.1$), the predictive power can be increased by a factor of 3.3 to 0.5%. Note that the relative signal to noise ratio is much lower in this study than in the previous annual financial returns study in "An empirical application to time series data" section. The single rounds are presented in Table 6. The final model is given by

$$\hat{f}_{FW} = -0.4\bar{y} + 0.3\hat{f}_4 + 0.3\hat{f}_5 + 0.3\hat{f}_7 + 0.1\hat{f}_{13} + 0.4\hat{f}_{16}$$

Table 6 Predictive power (in percent) of the round-based forward algorithm fixing $\alpha = 0.1$

Round	Model	Covariates	R_V^2
1	16	*ad*, *nk*	0.127
2	16	*ad*, *nk*	**0.228**
3	16	*ad*, *nk*	**0.303**
4	16	*ad*, *nk*	**0.351**
5	13	*ag*, *sk*	**0.385**
6	5	*nk*	**0.410**
7	4	*tk*	**0.432**
8	7	*sk*	**0.453**
9	5	*nk*	**0.467**
10	4	*tk*	**0.480**
11	7	*sk*	**0.493**
12	7	*sk*	**0.498**
13	4	*tk*	**0.503**
14	5	*nk*	**0.504**

Steps in Algorithm 1 with improved predictive power compared to the single best predictor $\hat{f}_5$ highlighted in bold

where 14 rounds are used, and the model weights sum to 1.4. Note that the second round produces a model with improved predictive power compared with the single best model. Interestingly, models based on telematics variables (annual distance driven in thousands of kilometers, percentage of kilometers driven at night, and percentage of kilometers driven above the speed limit) and their interactions with the age of the driver and the number of years of holding a driver's license were selected. Bolancé et al. (2023) report a mean squared prediction error for their preferred model of 1.303. When calculating the mean squared prediction error for our $\hat{f}_{FW}$ using the 488 observations in the dataset, we find a smaller value of 1.220, indicating better predictive performance than the single-index model used in Bolancé et al. (2023).

Conclusion

This study introduces a new regression methodology applicable to both standard regression and time-series regression settings. Our new approach uses low-dimensional estimators as building blocks for a larger system instead of going directly to a large system, as its main competitors do. The new approach is particularly useful when many covariates but only a few observations of the dependent variable are available. Our data and simulation sections illustrate that our new approach is superior to its main competitors in our chosen finite sample studies (inspired by typical datasets of interest), and also seems superior in our important real-life data examples taken from the finance and insurance industries.

Our new methodology can also be considered an addition to the toolbox of machine learning methodology in insurance and finance; see, for example, Asimit et al. (2020). It is, of course, not only the mean values that can be modelled via our new approach to

machine learning, but also volatility or other moments that can be modelled in a similar way; see Mammen et al. (2019) for the second moment case, see also Gong et al. (2022). Further developments in the second moments might facilitate improvements to ARCH- or GARCH type structures, along the same lines as in Wu and Karmakar (2023), but with the transparency and robustness of the methodology of this study. Therefore, our new approach could be an alternative or supporting methodology for other volatility forecasts; see Zahid et al. (2022) among many others, for forecasting clusters of volatility based on GARCH-type models. Our technique can also be envisioned for higher-frequency data making it relevant, for example, for trading data; see, for example, Frattini et al. (2022). It would also be interesting to provide a non-supervised learning version (e.g., classification) of our supervised learning approach (regression) to provide an alternative methodology to practical problems, such as those in Brunhuemer et al. (2022). We hope that our new approach can help determine an optimal level of complexity within a much wider range of applications than the two financial applications indicated in this study. Note that any sophisticated modern alternative technique to our approach can be enlisted as a function to be included in our approach when validating our method for the optimal model. Our approach is therefore very flexible and can perhaps be imagined to work together with other modern techniques and applications; for example, see Kou et al. (2021, 2024a, 2024b) and Xu et al. (2024).

Abbreviations

CV	Cross-validation
dgp	Data generating process
MSFE	Out-of-sample mean squared forecast error
OLS	Ordinary least squares
oos	Out of sample

Acknowledgements
Not applicable.

Author contributions
The four authors have contributed equally.

Funding
M. D. Martínez-Miranda acknowledges financial support from Ministerio de Ciencia, Innovación y Universidades (PID2020- 116587GB-I00). M. Scholz acknowledges financial support from Austrian National Bank (Jubiläumsfondsprojekt 18901).

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests

Received: 28 April 2023 Accepted: 10 July 2024
Published online: 02 October 2024

References

Anatolyev S (2019) Many instruments and/or regressors: a friendly guide. J Econ Surv 33:689–726. https://doi.org/10.1111/joes.12295
Anh L, Dong L, Kreinovich V, Thach N (2017) Econometrics for financial applications. Studies in computational intelligence. Springer, Berlin

Asimit V, Kyriakou I, Nielsen JP (2020) Special issue "machine learning in insurance". Risks 8:54
Belloni A, Chernozhukov V, Chetverikov D, Fernández-Val I (2019) Conditional quantile processes based on series or many regressors. J Econom 213:4–29 (**Annals: In Honor of Roger Koenker**)
Bergmeir C, Hyndman RJ, Koo B (2018) A note on the validity of cross-validation for evaluating autoregressive time series predictions. Comput Stat Data Anal 120:70–83
Bischl B, Binder M, Lang M, Pielok T, Richter J, Coors S, Thomas J, Ullmann T, Becker M, Boulesteix A-L, Deng D, Lindauer M (2023) Hyperparameter optimization: foundations, algorithms, best practices, and open challenges. Wiley Interdiscip Rev Data Min Knowl Discov 13:e1484. https://doi.org/10.1002/widm.1484
Bolancé C, Cao R, Guillén M (2018) Flexible maximum conditional likelihood estimation for single-index models to predict accident severity with telematics data. Working paper WP E-IR18/29, Universitat de Barcelona. Facultat d'Economia i Empresa. https://diposit.ub.edu/dspace/bitstream/2445/126954/1/IR18-029_Bolance%2bCao%2bGuillen.pdf
Bolancé C, Cao R, Guillén M (2022) Single-index model for motor insurance claim severity: kernel estimated conditional likelihood based inference. Mendeley Data V1. https://data.mendeley.com/datasets/py3kb2hn2b/1
Bolancé C, Cao R, Guillén M (2023) Conditional likelihood based inference on single index-models for motor insurance claim severity. SORT Stat Oper Res Trans (to appear)
Brodeur A, Cook N, Heyes A (2020) Methods matter: p-hacking and publication bias in causal analysis in economics. Am Econ Rev 110:3634–60. https://doi.org/10.1257/aer.20190687
Brunhuemer A, Larcher L, Seidl P, Desmettre S, Kofler J, Larcher G (2022) Supervised machine learning classification for short straddles on the SP500. Risks 10:235
Campbell JY, Thompson SB (2008) Predicting excess stock returns out of sample: Can anything beat the historical average? Rev Financ Stud 21:1509–1531
Clemen R (1989) Combining forecasts: a review and annotated bibliography. J Forecast 5:559–583
Fan Q, Wu R, Yang Y, Zhong W (2022) Time-varying minimum variance portfolio. J Econom 239:105339
Filzmoser P, Nordhausen K (2021) Robust linear regression for high-dimensional data: an overview. WIREs Comput Stat 13:e1524
Frattini A, Bianchini I, Garzonio A, Mercuri L (2022) Financial technical indicator and algorithmic trading strategy based on machine learning and alternative data. Risks 10:225
Gong X, Zhang W, Xu W, Li Z (2022) Uncertainty index and stock volatility prediction: evidence from international markets. Finant Innov 8:57
Hastie T, Tibshirani R, Friedman J (2017) The elements of statistical learning. Springer, New York
Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67
James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York
Kelly BT, Malamud S, Zhou K (2022) The virtue of complexity in return prediction. Working paper 30217, National Bureau of Economic Research. http://www.nber.org/papers/w30217
Kou G, Xu Y, Peng Y, Shen F, Chen Y, Chang K, Kou S (2021) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140:113429
Kou G, Dinçer H, Yüksel S (2024a) Pattern recognition of financial innovation life cycle for renewable energy investments with integer code series and multiple technology S-curves based on Q-ROF DEMATEL. Financ Innov 10:53
Kou G, Dinçer H, Yüksel S, Alotaibi FS (2024b) Imputed expert decision recommendation system for QFD-based omnichannel strategy selection for financial services. Int J Inf Technol Decis Mak 23:141–170. https://doi.org/10.1142/S0219622023300033
Kyriakou I, Mousavi P, Nielsen JP, Scholz M (2020) Longer-term forecasting of excess stock returns-the five-year case. Mathematics 8:1–20
Kyriakou I, Mousavi P, Nielsen JP, Scholz M (2021a) Forecasting benchmarks of long-term stock returns via machine learning. Ann Oper Res 287:221–240
Kyriakou I, Mousavi P, Nielsen JP, Scholz M (2021b) Short-term exuberance and long-term stability: a simultaneous optimization of stock return predictions for short and long horizons. Mathematics 9:1–19
Leeb H, Steinberger L (2021) Statistical inference with F-statistics when fitting simple models to high-dimensional data. Econom Theory 39:1–24
Mammen E, Nielsen JP (2003) Generalised structured models. Biometrika 90:551–566. https://doi.org/10.1093/biomet/90.3.551
Mammen E, Nielsen JP, Scholz M, Sperlich S (2019) Conditional variance forecasts for long-term stock returns. Risks 7:1–22
McGibney G, Smith MR (1993) An unbiased signal-to-noise ratio measure for magnetic resonance images. Med Phys 20:1077–1078. https://doi.org/10.1118/1.597004
Nielsen JP, Sperlich S (2003) Prediction of stock returns: a new way to look at it. ASTIN Bull 33:399–417
Rapach DE, Zhou G (2020) Time-series and cross-sectional stock return forecasting: new machine learning methods. In: Jurczenko E (ed) Machine learning for asset management: new developments and financial applications. Wiley, Hoboken, pp 1–33
Scholz M (2022) Forecast combinations for benchmarks of long-term stock returns using machine learning methods. Ann Oper Res. https://doi.org/10.1007/s10479-022-04880-4
Shiller RJ (1989) Market volatility. MIT Press, Cambridge
Spiliotis E, Abolghasemi M, Hyndman RJ, Petropoulos F, Assimakopoulos V (2021) Hierarchical forecast reconciliation with machine learning. Appl Soft Comput 112:107756
Steinberger L, Leeb H (2019) Prediction when fitting simple models to high-dimensional data. Ann Stat 47:1408–1442
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 58:267–288
Wand M, Jones M (1994) Kernel smoothing. Monographs on statistics and applied probability. Chapman and Hall/CRC, London
Wu K, Karmakar S (2023) A model-free approach to do long-term volatility forecasting and its variants. Finant Innov 9:59
Xu Y, Kou G, Peng Y, Ding K, Ergu D, Alotaibi FS (2024) Profit- and risk-driven credit scoring under parameter uncertainty: a multiobjective approach. Omega 125:103004

Yae J, Luo Y (2023) Robust monitoring machine: a machine learning solution for out-of-sample R2-hacking in return predictability monitoring. Financ Innov 9:1–28
Zahid M, Iqbal F, Koutmos D (2022) Forecasting bitcoin volatility using hybrid Garch models with machine learning. Risks 10:237
Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429

Publisher's Note