Improving the out-of-sample predictive accuracy of average window stock return forecasts using truncation

Author: Igor Uiterwijk

Student number: *494212*

Bachelor Thesis BSc^2

Supervised by: Dr. Onno Kleen

Second Assessor: Bram van Os

Date: 03/07/2022

Erasmus University Rotterdam

Erasmus School of Economics

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract. Using traditional methods, it is often difficult to outperform the historical average forecast when forecasting stock returns. However, H. Zhang et al. (2020) use an average window estimation method with shrinkage to account for parameter instability and model uncertainty when forecasting stock returns. After applying this method on seven different sophisticated models, they find that all of the models outperform the historical average benchmark model. In this paper, we expand upon their method by setting negative return forecasts to zero, inspired by the truncation method from Campbell and Thompson (2008). Using the data from Welch and Goyal (2008), we show that it in some cases leads to increased forecasting accuracy as well as utility gains for the mean-variance investor. We consider truncation at three separate stages of the method from H. Zhang et al. (2020), and using the Clark and West (2007) test statistic we find that overall, the difference between the benchmark model and sophisticated models becomes more significant after implementing the restrictions. Lastly, we find that during the COVID-19 pandemic, the gains in predictive accuracy from applying truncation are even larger.

Table of Contents

1 Introduction

The worldwide COVID-19 pandemic has been a very turbulent period for investors and those interested in the stock market. All over the world there were unprecedented increases in volatility (Basuony et al. 2021), for some countries resulting in a fall of their stock indices of more than 50% (Ganie et al. 2022). Recently, H. Zhang et al. (2020) devised a method that improves the out of sample predictive accuracy for several popular forecasting models by introducing methods that takes into account parameter instability and model uncertainty, as well as introducing shrinkage, based on previous research from Welch and Goyal (2008) and Pesaran and Timmermann (2007). Their methods improved on the existing methods in the periods of a stock market crash, as well as during times of economic growth. Additionally, Campbell and Thompson (2008) used economic reasoning to improve forecasting accuracy from Welch and Goyal (2008), by imposing restrictions on the forecasted returns to be positive. This leads to the main research question of our study:

> **RQ.1:** How does imposing the restrictions from Campbell and Thompson (2008) affect the methods from H. Zhang et al. (2020)?

Our research uses the ideas from Campbell and Thompson (2008) to truncate negative forecast returns, estimated using the methods and models from H. Zhang et al. (2020). We conclude that often gains in forecasting accuracy can be achieved, especially when only considering recent data. Moreover, both studies note that in periods of extreme economic downturn, their methods tend achieve better results than in less volatile times. As the COVID-19 crisis saw some unprecedented negative returns, we pose the second research question:

> **RQ.2:** How do the results of applying truncation to the methods from H. Zhang et al. (2020) differ during the COVID-19 pandemic compared to pre-COVID19 times?

Using an out-of-sample period of [2020:02-2021:12] to denote the COVID-19 crisis, we find that applying truncation leads to more gains in terms of out-of-sample prediction accuracy and utility during the COVID-19 outbreak compared to the pre-COVID-19 period.

In Section 2, we further elaborate on the research questions. Additionally, we will examine the main findings of the existing research on this topic, highlight the most important results, and outline what this paper will contribute to the literature. Subsequently, in Section 3, we describe how we obtain the data as well as how we construct the variables. In Section 4, we explain which methods and models we use to forecast the returns, as well as how we evaluate the the forecasts. A discussion of the results will be given in Section 5. Lastly, in Section 6, we will discuss our conclusions, the limitations to our research, and suggestions for further research.

2 Literature Review

In this section we will examine the existing literature on the estimation of stock returns and the methods most commonly used. Then, we will outline what our research contributes to the existing literature on this topic. Lastly, we will consider how this research can be of use in practise.

The first important paper which we will consider as a foundation of our research is the paper by Welch and Goyal (2008). They examined and summarized the performance of the predictors that had been deemed the most effective at predicting stock returns, as determined by the historical academic literature. Using simple linear regressions, they conclude that the variables have poor predictive power out-of-sample. They continue by implementing the restrictions proposed by Campbell and Thompson (2008). Two types of restrictions were suggested. Firstly, they considered a sign restriction on the simple regression coefficient. In the case the sign of an estimated coefficient was not consistent with the theoretical relation between the variable and stock returns, they used the historical average to forecast the returns. Additionally, they considered a restriction on the sign of the forecasted returns. Since they defined stock returns as the difference between the index returns and the risk free rate, a negative stock return forecast would imply that the risky asset is expected to have smaller returns than the risk-free asset. In that case, any rational investor would invest in the risk-free asset as opposed to the risky asset, resulting in an excess return of zero. Therefore, they imposed a so-called equity premium restriction by setting all negative returns forecasts to 0. Both of these restrictions combined lead to improvements in forecasting accuracy.

Next, we consider the paper by H. Zhang et al. (2020), who use the variables from Welch and Goyal (2008) in several different sophisticated models, estimated using the average window method devised by Pesaran and Timmermann (2007) to account for parameter instability and model uncertainty. Subsequently, they shrink the model using the historical average forecast, by forecasting with an equal weighted combination of the sophisticated model with the historical average. Campbell and Thompson (2008), Welch and Goyal (2008), and H. Zhang et al. (2020) evaluate the models based on their out of sample R^2, which compares the MSFE or the combination model with that of the benchmark model, which is defined as the historical average. Since the out-of-sample (OS) R^2 is almost always lower compared the in-sample R^2, as stated by Copas (1983), sometimes it is desirable to introduce shrinkage to decrease the out-of-sample Mean Squared Forecast Error (MSFE). This exchange between an increase in bias for a reduction in the variance is known as the bias-variance trade-off.

However, from simply looking at the R^2_{OS} it is unclear how someone might benefit from these improved forecasts. A naive view would be to look at how these affect the expected returns, and therefore utility, using some investment strategy. However, in reality maximizing the expected returns is not the only goal of an investor in general. It is well known that there are factors outside of expected return that influence how an individual perceives utility, which is at the core of behavioural economics. Real phenomena such as risk aversion and loss aversion (Godoi et al. 2005; Morin and Suarez 1983) have to be taken into account when determining utility. Still, the question remains, to what extent do individuals maximize their own utility?

While it does not account for all of the above, Welch and Goyal (2008) and H. Zhang et al. (2020) consider utility to be a function of both the mean and the variance of the realized returns. The measure they use for utility is the annualized difference between the certainty equivalent returns (CER) between the benchmark model and the model that is to be evaluated. This measure takes into account the mean and the variance of the realized returns, as well as a variable risk factor γ. As no two individuals are alike, including this risk factor makes it such that we can account for indivdual differences with regards to risk preferences. Moreover, they use the Sharpe (1966) ratio of the constructed realized returns, which defined as the ratio between the mean and the standard deviation of the excess returns, to further account for the risk associated of the portfolio.

H. Zhang et al. (2020) consider several sophisticated models, starting off with a multiple linear regression, where the dependent variable, excess stock returns, is regressed on the set of independent variables from Welch and Goyal (2008). Additionally, they consider six more elaborate models. Firstly, they consider least absolute shrinkage and selection operator (LASSO), devised by Tibshirani (1996), which is a self regularization method that performs variable selection as to reduce the effect of overfitting. Next, they applied the elastic net method which is an extension of LASSO introduced by Zou and Hastie (2005). The next three models are all model averaging methods, namely Bayesian Model Averaging, Mallows Model Averaging (Hansen 2007; Wan et al. 2010), and Jacknife Model Averaging (Hansen and Racine 2012; X. Zhang et al. 2013). Instead of only considering one specification, these methods use a weighted combination of different models, based on different selection criteria. Lastly, they use a new model averaging method, named weighted-average least squares, first introduced by Magnus, Powell, et al. (2010). These models will be covered in more detail in Section 4.

Even though the out-of-sample R^2 values often do not exceed 2%, H. Zhang et al. (2020) conclude that the proposed methods can lead to utility gains for the rational mean-variance investor. However, as discussed previously, this does not necessarily lead to welfare gains. On the other hand, the relevance of these methods are not limited to the stock market, as they can also be applied in other areas of economics, such as in health economics (Jackson et al. 2009), or when forecasting inflation (Koop and Korobilis 2012). Moreover, some of these methods are used in non-economic settings as well. For instance, model averaging is also used in weather forecasts (Raftery, Gneiting, et al. 2005; Sloughter et al. 2010). Furthermore, the elastic net regressions are extensively used in the studying of genetics (Amini and Hu 2021; Hughey and Butte 2015; Ogutu et al. 2012), industrial data (Yu and Zhao 2019), or when forecasting solar and wind energy production (Nikodinoska et al. 2022). As these models are widely applicable, researching and improving them could be of importance in many different fields.

3 Data

As this research builds upon the research done by H. Zhang et al. (2020), we use updated data from Welch and Goyal (2008) over the period [1926:12-2021:12]. From the data set, several economic variables can be constructed, of which we will use 12 as independent variables, similar to the previous research. These variables include log dividend-price ratio (DP) and log dividend yield (DY), which can be constructed by taking the difference between the log of the dividend from the log of prices, or the log of the lag of the prices, respectively. The log earnings-price ratio (EP) is obtained by subtracting the log of prices from the log of earnings. The term spread (TMS) is given by the difference between the long term yield (LTY) on government bonds and the Treasury-bill (TBL). The final variables that need to be constructed are the default yield spread (DFY), which can be constructed from the AAA and BAA rated corporate bond yield, and the default return spread (DFR), which can be obtained by substracting the government long-term bond returns from the corporate long-term bond returns. The remaining independent variables, stock variance (SVAR), book-to-market ratio (BM), net equity expansion (NTIS), long-term return (LTR) and inflation (INFL) are all given in the data set. We do not include LTY as a variable, as it is a linear combination of TBL and TMS. The dependent variable, stock returns, is constructed by taking the log difference between the CRSP value weighted index including dividends and the risk-free rate. The value weighted CRSP index is an index for the US stock market that is covers approximately 26,500 stocks (Center for Research in Security Prices 2022). We use this as our dependent variable to be in line with the previous literature from Campbell and Thompson (2008) and Welch and Goyal (2008). However, H. Zhang et al. (2020) use a different dependent variable, namely an index for the S&P500. Therefore, the results will differ slightly from their paper. Nonetheless, the results should be comparable, as the CRSP value weighted index moves similarly to the S&P500 index, as the latter makes up

Table 1: Summary statistics.

Variable	Mean	St. Dev.	Min	Max
Returns	0.005	0.054	−0.339	0.346
BM	0.554	0.269	0.121	2.028
TBL	0.033	0.031	0.0001	0.163
NTIS	0.016	0.026	−0.056	0.177
INFL	0.002	0.005	−0.021	0.059
LTR	0.005	0.025	−0.112	0.152
SVAR	0.003	0.006	0.0001	0.073
DP	−3.401	0.473	−4.524	−1.873
DY	−3.396	0.470	−4.531	−1.913
EP	−2.758	0.421	−4.836	−1.775
TMS	0.017	0.013	−0.037	0.046
DFY	0.011	0.007	0.003	0.056
DFR	0.0004	0.014	−0.098	0.074

a large portion of the former. Following previous research, we divide the data in an in-sample period [1926:12-1956:12] and an out-of-sample period [1957:01-2016:12]. We also consider a more recent dataset, as the stock market has changed drastically since the digital revolution. For the evaluation of this period, as well as the performance of the models during recessions, as indicated by NBER (2022), we use an in-sample period of [1960:01-1979:12] and an out-of-sample period of [1980:01-2021:12]. Using this data we will also analyse the performance of the models during the COVID-19 crisis, which we define as [2020:02-2021:12] following Azar II (2020). The summary statistics are given in Table 1. We consider a wide range of returns, as implied by the difference between the minimum and maximum. Furthermore, all variables are within their theoretical range, for instance the log dividend and yield ratios are strictly negative, and TBL and BM are strictly greater than 0.

4 Methodology

Firstly, we will give a short summary of the methods from H. Zhang et al. (2020). Afterwards we will explain how we apply the restrictions from Campbell and Thompson (2008) to the aforementioned methods. Finally, we will cover how we will evaluate the performance of the models.

4.1 Sophisticated Models

For our analysis we consider seven different models. The first model we consider is the standard multiple regression model using all 12 explanatory variables with a constant. One of the downsides of this method is that it susceptible to overfitting, as it is unclear whether all variables should be included. A solution to this problem is included in the next two models we consider, namely, least absolute shrinkage and selection operator (LASSO), developed by Tibshirani (1996), and elastic net, an extension of LASSO devised by Zou and Hastie (2005). Where ordinary least squares is designed to minimize the squared residuals, elastic net and LASSO introduce the a penalty term to the squared residuals to be minimized such that,

$$\hat{\beta}_{EN} = \underset{\beta \in \mathbb{R}^N}{\arg\min} \left(\frac{1}{T} \sum_{t=0}^{T-1} (r_{t+1} - X_t \beta)^2 + \lambda \sum_{i=1}^{N} \left(\frac{1-\alpha}{2} \beta_i^2 + \alpha |\beta_i| \right) \right). \tag{1}$$

In order to get estimates for the elastic net method we set $\alpha = 0.5$, whereas for LASSO we use $\alpha = 1$. Lambda is determined by 5-fold cross-validation using the *glmnet* package in R by Friedman et al. (2010).

The following three models we consider are all model averaging methods. Instead of considering only one model specification, model averaging methods consider several models and weight them based on a certain criterion. In general, a model averaging forecast with m candidate models is calculated as follows,

$$\hat{r}_{T+1}^{MA} = \sum_{i=1}^{m} \lambda_i \hat{r}_{T+1}^{i} = X_T \hat{\beta}(\lambda). \tag{2}$$

$\hat{r}_{T+1}^{i}$ in Eq. 2 denote the estimated returns of model specification i for time $T+1$, and is appointed weight λ_i, where $\sum_{i=1}^{m} \lambda_i = 1$. $\hat{\beta}(\lambda)$ denotes an averaging estimator for β weighted by $\lambda = (\lambda_1, ..., \lambda_m)$. The methods we consider differ in the way λ is determined. For Bayesian model averaging (BMA) (Raftery, Madigan, et al. 1997), the weight for model i is calculated using the Bayes Information Criterion, $BIC_i = T \log(\hat{\sigma}_i^2) + N_i \log(T)$ (Priestley 1981). Here, $\hat{\sigma}_i^2$ denotes the estimate of the variance of the errors for model i, and N_i denotes the number of predictors. We assume that the model errors are identically and independently distributed according to a normal distribution and that the derivative of the log likelihood with respect to the true variance of the errors is equal to 0. Buckland et al. (1997) then give an approximation for the model weights,

$$\lambda_i^{BMA} = \frac{\exp(-\frac{1}{2} BIC_i)}{\sum_{j=1}^{m} \exp(-\frac{1}{2} BIC_j)} \tag{3}$$

For Mallow's model averaging (Hansen 2007; Wan et al. 2010) the Mallows Criterion is used, which is defined as:

$$C_p(\lambda) = \left(R - X_T \hat{\beta}(\lambda) \right)' \left(R - X_T \hat{\beta}(\lambda) \right) + 2\sigma^2 \lambda K. \tag{4}$$

Here, $K = (K_1, ..., K_M)$ where K_i denotes the number of explanatory variables is model i. The vector of model weights λ is then determined as follows,

$$\lambda^{MMA} = \underset{\lambda \in \mathbb{H}}{\arg\min} C_p(\lambda), \tag{5}$$

where $\mathbb{H} = \{\lambda \in \mathbb{R}^m : \lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = 1\}$.

The final model averaging method we consider is jackknife model averaging (Hansen and Racine 2012; X. Zhang et al. 2013), where a leave-one-out cross-validation criterion is used to determine the model weights. Let $\tilde{R}_i = \tilde{r}_1^{(i)}, ..., \tilde{r}_T^{(i)}$ denote the estimated returns R,

$$\tilde{r}_t^{(i)} = X_t \hat{\beta}_{-t}, \tag{6}$$

where $\hat{\beta}_{-t}$ denotes the OLS estimator of the model coefficients with the t-th observations deleted. The jackknife returns are then estimated using $\tilde{R}(\lambda) = \sum_{i=1}^{m} \lambda_i \tilde{R}^{(i)}$. The leave-one-out cross-validation criterion is then defined as follows,

$$CV_T(\lambda) = \frac{1}{T} \tilde{e}(\lambda)' \tilde{e}(\lambda), \tag{7}$$

where $\tilde{e}(\lambda) = R - \tilde{R}(\lambda)$ denotes the vector containing the residuals of the jackknife returns. The vector of model weights is determined as follows,

$$\lambda^{JMA} = \underset{\lambda \in \mathbb{H}}{\arg\min} CV_T(\lambda), \tag{8}$$

where $\mathbb{H}$ is the same as defined in Eq. 5.

Lastly, we consider weighted-average least squares (WALS), developed by Magnus, Powell, et al. (2010), which has the advantage over BMA that it is computationally easier to solve, and it uses either Laplace, Subbotin, or in this case Weibull priors, which result in bounded prediction variance, as opposed to the Normal priors in BMA. The intuition behind WALS is that the variables are seperated into two sets X_1 and X_2, where X_1 contains a constant and any variables that must be included in the regression based on prior theoretical knowledge, and X_2 contains the set variables that may or may not be included. The return forecasts are given by the following formula,

$$\hat{r}_{T+1}^{WALS} = X_{1,T} \hat{\beta}_1 + X_{2,T}^* \hat{\beta}_2^*, \tag{9}$$

here, $X_{2,T}^*$ is a transformation of $X_{2,T}$ such that (X_1, X_2^*) becomes a semi-orthogonal matrix. $\hat{\beta}_2^*$ is an estimate for the coefficient of X_2^* that incorporates Weibull priors to determine variable weights. For this paper, we consider X_1 to only consist of the constant. More detailed explanations and derivations of these methods can be found in their respective papers.

4.2 The Average Window Method

We estimate forecasts for each of the six models using the average window method, from Pesaran and Timmermann (2007). Firstly we construct m estimation windows $\mathbb{W}_i = \{r_{t+1}, X_t\}_{t=T-w_i}^{T-1}$, where $i = 1, ..., m$. w_i is determined as follows,

$$w_i = w_{min} + \frac{i-1}{m-1}(T - w_{min}). \tag{10}$$

Here, w_{min} is a pre-specified minimum window size. Following H. Zhang et al. (2020), we will use $m = 10$ and $w_{min} = 240$. For each of the estimation windows we make a single forecast by applying one of the models using the given estimation window. For a given window, we apply one of the specified models. For illustration, the multiple linear regression would be applied as follows,

$$r_{t+1} = X_t\beta + e_{t+1}, \quad t = T - w_i, ..., T - 1. \tag{11}$$

Here X_t is a vector of explanatory variables with a constant as the first variable, β denotes the parameter vector and e_{t+1} denotes the error term. Given the returns for each window, we can compute $\hat{r}_{T+1}^{AveW}$ using,

$$\hat{r}_{T+1}^{\text{AveW}} = \frac{1}{m}\sum_{i=1}^{m} \hat{r}_{T+1}(\mathbb{W}_i), \tag{12}$$

where $\hat{r}_{T+1}(\mathbb{W}_i)$ denotes the forecast of a specified model using estimation window $\mathbb{W}_i$. Following H. Zhang et al. (2020), using the results from the average window method we construct model C for the returns which we will use for forecasting.

$$\hat{r}_{T+1}^{\text{C}} = (1 - \delta)\hat{r}_{T+1}^{\text{HA}} + \delta\hat{r}_{T+1}^{\text{AveW}}, \tag{13}$$

The model is a linear combination of the shrinkage target, in this case historical average returns, $\hat{r}_{T+1}^{\text{HA}} = \frac{1}{T}\sum_{i=1}^{T} r_i$, and the forecasted returns as per one of the average window method, $\hat{r}_{T+1}^{\text{AveW}}$. Here, δ denotes the shrinkage factor, which we set to 0.5 for equal weights.

4.3 Truncation of Negative Returns

Note that the estimated returns can be either positive or negative. However, as Campbell and Thompson (2008) point out, when the expected excess returns are negative, an investor will only invest in the risk free asset. Therefore, they propose that negative expected returns should be set to zero. In this paper we will implement this truncation approach to the methods average window methods from H. Zhang et al. (2020). Firstly, we apply the truncation to estimates obtained using expanding windows. Then, we consider the average window methods. Note that truncation can be applied at three separate stages in the procedure. Firstly, it can be applied before averaging (BA) the estimates, i.e. before Eq. 12. Next, we consider truncation after averaging (AA), i.e. after Eq. 12, but before applying shrinkage in Eq. 13. Lastly, it can be applied after shrinkage (AS), in Eq. 13. As the historical average forecast is strictly positive in the dataset we consider, each of these truncation methods will result in strictly non negative estimates of the returns, and separate truncation of the historical average forecasts has no effect. As the results are likely to differ depending on when truncation is applied, we will consider all three cases.

4.4 Model Evaluation

The performance of the models will be determined based on their R^2, which is defined as:

$$R^2_{OS} = 1 - \frac{\frac{1}{p}\sum_{t=T+1}^{T+p}(r_t - \hat{r}_t^{\mathrm{C}})}{\frac{1}{p}\sum_{t=T+1}^{T+p}(r_t - \bar{r}_t)} = 1 - \frac{\mathrm{MSFE}^{\mathrm{C}}}{\mathrm{MSFE}^{\mathrm{bmk}}}. \tag{14}$$

Here, $\bar{r}_t$ denote the returns estimated using the benchmark model, which we define as the historical average of the returns. $\mathrm{MSFE}^{\mathrm{C}}$ and $\mathrm{MSFE}^{\mathrm{bmk}}$ denote the mean squared forecast error for the combined model C and the benchmark model, respectively. A negative R^2 indicates that the benchmark model outperforms model C, whereas a positive value indicates that the model C performs better. We expect that applying the truncation methods will result in a higher R^2_{OS} compared to the models without truncation, as a result of a decrease in variance. The significance of the models will be estimated using the Clark and West (2007) test statistic which tests the null-hypothesis, H_0, that $R^2_{OS} < 0$, with the alternative hypothesis, H_1, stating $R^2_{OS} \geq 0$.

Additionally, we evaluate the models based on their utility in an investment setting, as the R^2 measure does not account for the risk element of an investment for the investor. Following Rapach et al. (2010), Campbell and Thompson (2008) and H. Zhang et al. (2020), we assume that a mean-variance investor assigns a portfolio weight of $0\% \leq w_T^{\mathrm{C}} \leq 150\%$ to equities and invests the rest in the risk free asset, such that the realized returns from the portfolio constructed from model C amount to,

$$R_p^{\mathrm{C}} = w_T^{\mathrm{C}} r_{T+1} + r_{T+1}^f, \tag{15}$$

Table 2: R^2 of simple methods with and without truncation.

Methods	R^2_{EW}	$R^2_{EW,TR}$	R^2_{AveW}	$R^2_{AveW,BA}$	$R^2_{AveW,AA}$
Kitchen Sink	-7.73	-2.26	-2.99*	-0.41**	**-0.24*****
MMA	-3.80	-1.35	-0.62*	0.02*	**0.17****
JMA	-1.12	-0.85	-0.25*	0.05*	**0.27***
BMA	-2.26	-1.73	-1.30	-0.54	**-0.07**
LASSO	-0.78	-0.73	0.72**	0.68**	**0.97****
Elastic Net	-0.27	-0.15	**0.91****	0.58**	0.77**
WALS	-4.07	-1.46	0.15**	0.17**	**0.32****

Note: This table denotes the out-of-sample forecasting accuracy, R^2_{OS}, of the simple models without a historical average shrinkage factor, compared to the benchmark model, HA, using an out-of-sample evaluation period of [1957:01-2016:12]. The second and third column denote the results obtained using an expanding window, whereas the last three columns apply the average window method by Pesaran and Timmermann (2007). For $R^2_{EW,TR}$, $R^2_{AveW,BA}$ and $R^2_{AveW,AA}$, negative excess return forecasts are set to 0, similar to the truncation method proposed by Campbell and Thompson (2008). For the average window method we consider both truncation before averaging (BA), and after averaging (AA). For each model, the method that obtains the largest R^2_{OS} is written in boldface, and for each method, the model with the largest R^2_{OS} is underlined. Significance codes: *: 0.1; **: 0.05; ***: 0.01, obtained using the test statistic from Clark and West (2007), which tests $H_0 : R^2_{OS} < 0$, with $H_1 : R^2_{OS} \geq 0$.

where r_{T+1} denote the stock returns at time $T+1$, and r^{f}_{T+1} denote the returns of the risk free asset. The weights are determined as follows,

$$w^{\mathrm{C}}_{T} = \frac{1}{\gamma}\left(\frac{\hat{r}^{\mathrm{C}}_{T+1}}{\hat{\sigma}^2_{T+1}}\right), \tag{16}$$

where γ denotes the risk aversion parameter, which we set to 3. $\hat{\sigma}^2_{T+1}$ denotes the 5-year rolling window estimate of the variance of the stock returns, following H. Zhang et al. (2020). Negative weights are set to 0%, and weights greater than 150% are set to 150%. Using the realized returns, we can calculate utility, in this case the certainty equivalent return (CER), of the portfolio,

$$\mathrm{CER}_{\mathrm{C}} = \hat{\mu}_{\mathrm{C}} - \frac{\gamma}{2}\left(\hat{\sigma}^2_{\mathrm{C}}\right). \tag{17}$$

Here, $\hat{\mu}_{\mathrm{C}}$ and $\hat{\sigma}^2_{\mathrm{C}}$ denote the sample mean and variance for the out of sample realized portfolio returns. As we use monthly data, the annualized utility gains with respect to the benchmark model can be calculated as follows,

$$\mathit{\Delta}(\mathrm{ann}\%) = 100 * 12 * \left(\mathrm{CER}_{\mathrm{C}} - \mathrm{CER}_{\mathrm{bmk}}\right). \tag{18}$$

Lastly, we consider the Sharpe ratio which is defined as the mean of the excess portfolio returns divided by the standard deviation.

5 Results

We obtained the results using the following programmes and packages. Most of the models were implemented using *R* version 4.1.3, in *RStudio* with version 2022.02.0+443. Additionally, the following packages were used; *readxl* (Wickham and Bryan 2022) for importing the data, and *glmnet* (Friedman et al. 2010) for LASSO and elastic net forecasting. Furthermore, the packages *MAMI* (Schomaker and Heumann 2014) and *BMA* (Raftery, Hoeting, et al. 2022) were used for the model averaging methods. To export the results to LaTeX, we use the package *xtable* (Dahl 2013), and *stargazer* (Hlavac 2022), and the Clark and West (2007) statistic were obtained using the *tsm* package (Kotze 2020). Lastly, the WALS forecasts were estimated using MATLAB version 9.12.0.1956245 Update 2, using the code provided by Magnus and De Luca (2013).

Table 3: R^2 of average window method with shrinkage with and without truncation.

Methods	R^2_{AveW}	$R^2_{AveW,BA}$	$R^2_{AveW,AA}$	$R^2_{AveW,AS}$
Kitchen Sink + HA	0.47^{*}	$\underline{0.60}^{**}$	0.72^{**}	$\underline{\mathbf{0.83}}^{***}$
MMA + HA	$\mathbf{0.78}^{**}$	0.40^{**}	0.52^{**}	0.60^{**}
JMA + HA	$\mathbf{0.61}^{*}$	0.28^{*}	0.43^{**}	0.45^{*}
BMA + HA	-0.03	-0.01	**0.26**	0.21
LASSO + HA	$\mathbf{1.01}^{**}$	0.56^{**}	$\underline{0.76}^{***}$	0.82^{**}
Elastic Net + HA	$\mathbf{1.09}^{**}$	0.50^{**}	0.65^{**}	0.76^{**}
WALS + HA	$\underline{\mathbf{1.24}}^{**}$	0.57^{**}	0.69^{**}	0.82^{**}

Note: This table denotes the out-of-sample forecasting accuracy, R^2_{OS}, of a combination of the sophisticated model forecasts, estimated using the average window method by Pesaran and Timmermann (2007), and historical average (HA), with shrinkage factor $\delta = 0.5$, compared to the benchmark model, HA. The out-of-sample evaluation period is [1957:01-2016:12]. For $R^2_{AveW,BA}$, $R^2_{AveW,AA}$, and $R^2_{AveW,AS}$, negative return forecasts are set to 0, similar to the truncation method proposed by Campbell and Thompson (2008), before averaging (BA), after averaging (AA), and after applying shrinkage (AS), respectively. For each model, the method that obtains the largest R^2_{OS} is written in boldface, and for each method, the model with the largest R^2_{OS} is underlined. Significance codes: *: 0.1; **: 0.05; ***: 0.01, obtained using the test statistic from Clark and West (2007), which tests $H_0 : R^2_{OS} < 0$, with $H_1 : R^2_{OS} \geq 0$.

5.1 Replication

In this subsection we will examine the results from H. Zhang et al. (2020) as well as the results obtained after implementing our new truncation methods. We use the same evaluation period, [1926:12-2016:12], such that we can compare the results of our methods those from the previous research.

5.1.1 Simple Methods

The R^2_{OS} values of for the replication of the simple models from H. Zhang et al. (2020) are given in Table 2. Note that the values are slightly different from those in the original paper, this could be due to the different dependent variable, as we use the CRSP value weighted index as opposed to the S&P500 index. Furthermore, as the folds are determined randomly for LASSO and elastic net, for those models results may vary depending on the seed. The forecasts using an expanding window have poor predictive accuracy, as can be seen from their R^2_{OS}. The negative values indicate that all models are outperformed by the historical average forecast.

Table 4: Annualized utility gain and Sharpe ratios of average window methods with truncation using historical data with risk factor $\gamma = 3$

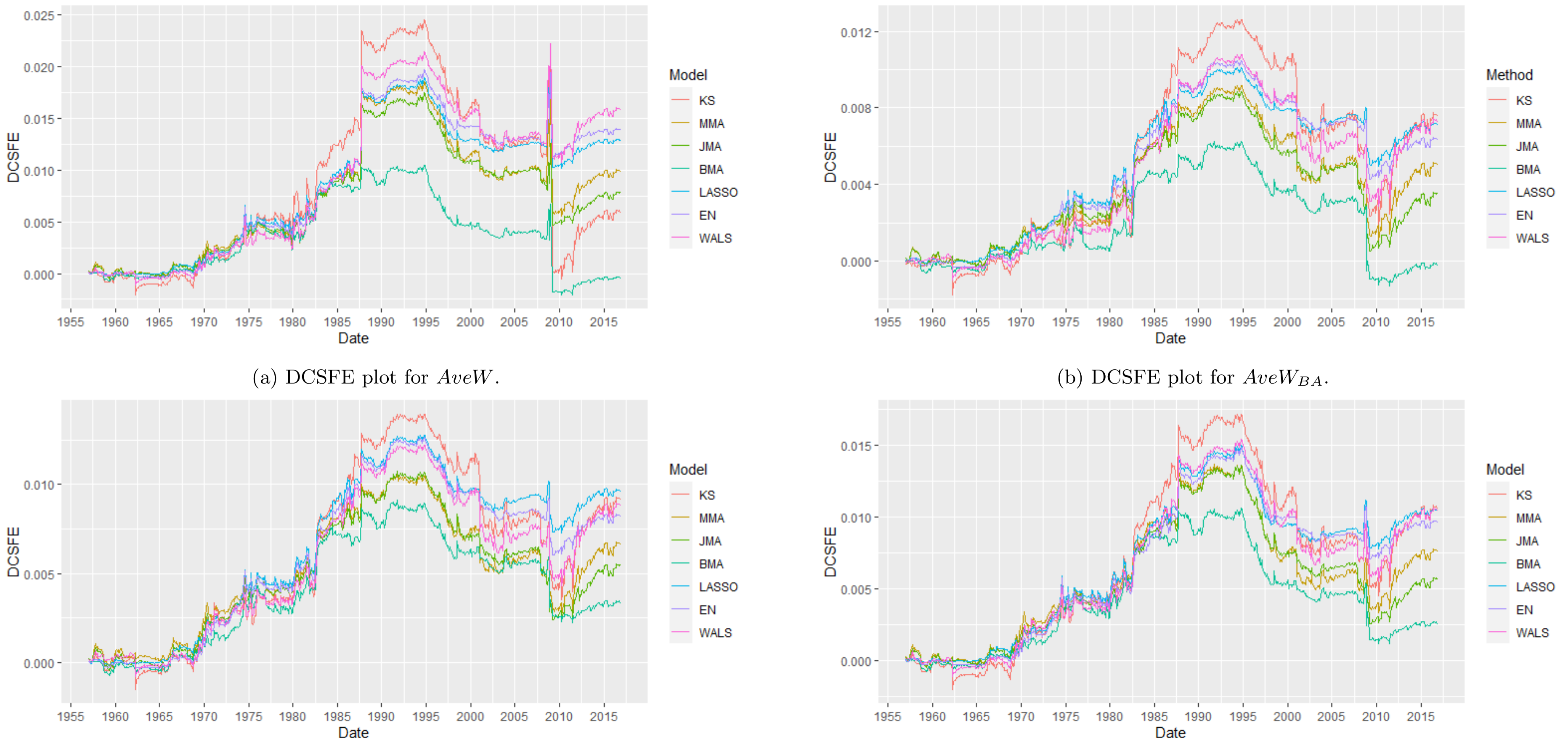
Methods	$AveW_{(AS)}$		$AveW_{BA}$		$AveW_{AA}$	
	$\Delta(ann\%)$	Sharpe	$\Delta(ann\%)$	Sharpe	$\Delta(ann\%)$	Sharpe
Kitchen Sink + HA	**2.39**	0.13	1.54	0.12	2.20	0.13
MMA + HA	**2.17**	0.12	1.38	0.12	1.99	0.12
JMA + HA	1.38	0.11	1.20	0.11	**1.54**	0.11
BMA + HA	0.99	0.10	0.65	0.09	**1.32**	0.11
LASSO + HA	**2.20**	0.13	1.50	0.13	2.11	0.12
Elastic Net + HA	**2.19**	0.12	1.31	0.12	1.91	0.12
WALS + HA	**2.49**	0.13	1.63	0.13	2.13	0.12

Note: This table denotes the annualized utility gains for the combinations of the sophisticated models with the historical average (HA) forecast, with shrinkage factor $\delta = 0.5$. It is expressed as a difference in Certainty Equivalent Return (CER) between the combination and the benchmark model, HA. Furthermore it denotes the Sharpe ratio, which is defined as the mean of the realized returns in excess of the risk-free rate, divided by the standard deviation of the excess returns. Negative excess return forecasts are set to 0, similar to the truncation method proposed by Campbell and Thompson (2008). We implement the truncation at three different stages, before averaging (BA), after averaging (AA), and after shrinkage (AS). By construction, implementing truncation after shrinkage and no truncation give the same results, as negative returns are assigned weight $w = 0$ in Eq. 16. For each model, the method with the largest $\Delta(ann\%)$ is written in boldface, and for each method, the model with the largest $\Delta(ann\%)$ is underlined. The weights are determined using a five year rolling window variance estimate $\hat{\sigma}^2$, and risk factor $\gamma = 3$. The out-of-sample evaluation period is [1957:01-2016:12].

Using the average window method as opposed to an expanding window improves results, as LASSO, elastic net and WALS now achieve a positive R^2_{OS}. After truncating the negative return forecasts, most of the results improve even further, as can be seen from the last two columns. When comparing the methods, it would seem as if applying truncation after averaging achieves better results compared to the case where the returns are truncated before averaging. However, note that for elastic net results slightly worsen after truncation. As the R^2 of LASSO does improve after truncation, it even overtakes elastic net as the best performing method. Notably, of all the model averaging methods, BMA consistently has the lowest R^2_{OS}. As BMA relies on the assumption of equal model priors it could be that this assumption causes the forecasts to be less accurate compared to JMA and MMA, which solely rely on the data to determine the model weights.

5.1.2 Applying Truncation to the Average Window Models with Shrinkage

Next, we consider the results obtained from an equal weighted combination of the average window method with the historical average forecast. The R^2_{OS} values are given in Table 3. Note that the linear combinations without truncation strictly outperform the corresponding models given in Table 2. Notably, even the kitchen sink model outperforms the historical average forecast after applying shrinkage. Applying truncation once again gives mixed results. Imposing the return restrictions before averaging over the windows gives the worst results. While for the kitchen sink model and for BMA it does lead to a slight improvement, the gains are larger when truncating at a different stage, and for the other models forecasting accuracy worsens compared no truncation. Imposing the restrictions after averaging gives results in the highest R^2_{OS} for BMA. While the other models do achieve a higher R^2_{OS} compared to the results from the third column, they do not outperform the models without truncation or $AveW_{AS}$, which applies truncation after combining the models with the historical average forecast. In this case, the kitchen sink model achieves its best results, outperforming all other models. However, as MMA, JMA, LASSO, elastic net, and WALS get better results without truncation, it is unclear whether truncation improves results in general. However, in general, truncation does seem to improve the significance of the Clark and West (2007) test statistic, even for JMA and LASSO. Notably, once more BMA has the worst performance of all the model averaging models. We assume this is the case due to similar reasons as discussed in Section 5.1.1. Additionally, LASSO and elastic net likely benefit from the built-in additional shrinkage as a result of the self-regularization that is present in those models, which is not present in any of the model averaging methods.



(a) DCSFE plot for $AveW$.

(b) DCSFE plot for $AveW_{BA}$.

(c) DCSFE plot for $AveW_{AA}$.

(d) DCSFE plot for $AveW_{AS}$.

Fig. 1: Time series plots of the forecasting accuracy of the different models, forecasted using the different AveW methods, with and without truncation. The accuracy is given in terms of DCSFE, which denotes the difference between the cumulative squared forecast error of the benchmark model and the sophisticated model. The greater the difference, the better the performance of the model. For each model we consider a linear combination of a sophisticated method with the historical average, estimated using the AveW method from H. Zhang et al. (2020) using shrinkage factor $\delta = 0.5$. We set the number of estimation windows $m = 10$. In panel (b), (c), and (d) we truncate negative return forecasts before averaging (BA), after averaging (AA), and after shrinkage (AS), respectively. The sophisticated models we consider are the kitchen sink model (KS), which is a standard multivariate regression, Mallows model averaging (MMA), by Hansen (2007), jackknife model averaging (JMA), by Hansen and Racine (2012), Bayesian model averaging with diffuse priors, by Raftery, Madigan, et al. (1997), LASSO (Tibshirani 1996) and elastic net (EN) (Zou and Hastie 2005) with fivevold cross-validation, and weighted-average least squares by Magnus, Powell, et al. (2010). The out-of-sample evaluation period we consider is [1957:01-2016:12].

We continue our examination by investigating the annualized utility gains, based on the difference in CER compared to the benchmark model, as well as the Sharpe ratio's. These results are given in Table 4. Upon first glance, each estimation method leads to utility gains compared to the benchmark model, the historical average forecast. Upon further investigation, it becomes clear that there are differences in utility gains depending on when the returns are truncated. For JMA and BMA truncation after averaging leads to the greatest utility gain, whereas for the other models truncation after shrinkage leads to the greatest utility.

Lastly we will consider the difference in cumulative squared forecast errors (DCSFE) between the benchmark model and the sophisticated models, estimated using the different *AveW* methods. These results are given in Fig. 1. For all models, the highest peak values are attained using the average window method with shrinkage without truncation, as can be seen from Fig. 1a. Note that all types of truncation reduces the difference in DCSFE between the different methods. This is likely a result of greater correlation between the estimates of different methods as a result of the truncation of negative return forecasts. The shape of the plots in Fig. 1b and Fig. 1c are similar, whereas Fig. 1a resembles Fig. 1d more closely.

Table 5: R^2 of average window method with shrinkage methods with and without truncation using recent data.

Methods	R^2_{AveW}	$R^2_{AveW,BA}$	$R^2_{AveW,AA}$	$R^2_{AveW,AS}$
Kitchen Sink + HA	-0.09^{*}	1.65^{***}	$\underline{\mathbf{1.80}}^{***}$	$\underline{1.75}^{***}$
MMA + HA	0.73^{*}	1.23^{***}	$\mathbf{1.32}^{***}$	1.24^{***}
JMA + HA	0.33	0.79^{**}	$\mathbf{0.87}^{**}$	0.70^{**}
BMA + HA	-1.19	0.60^{*}	$\mathbf{0.71}^{**}$	0.68^{**}
LASSO + HA	0.07	0.96^{**}	$\mathbf{0.98}^{**}$	0.90^{**}
Elastic Net + HA	0.27	1.03^{**}	$\mathbf{1.12}^{***}$	0.94^{**}
WALS + HA	$\underline{1.43}^{**}$	$\underline{1.70}^{***}$	$\mathbf{1.76}^{***}$	$\underline{1.75}^{***}$

Note: This table denotes the out-of-sample forecasting accuracy, R^2_{OS}, of a combination of the sophisticated model forecasts, estimated using the average window method by Pesaran and Timmermann (2007), and historical average (HA), with shrinkage factor $\delta = 0.5$, compared to the benchmark model, HA. The out-of-sample evaluation period is [1980:01-2021:12]. For $R^2_{AveW,BA}$, $R^2_{AveW,AA}$, and $R^2_{AveW,AS}$, negative return forecasts are set to 0, similar to the truncation method proposed by Campbell and Thompson (2008), before averaging (BA), after averaging (AA), and after applying shrinkage (AS), respectively. For each model, the method that obtains the largest R^2_{OS} is written in boldface, and for each method, the model with the largest R^2_{OS} is underlined. Significance codes: *: 0.1; **: 0.05; ***: 0.01, obtained using the test statistic from Clark and West (2007), which tests $H_0 : R^2_{OS} < 0$, with $H_1 : R^2_{OS} \geq 0$.

For all methods, the kitchen sink achieves the highest peak DCSFE, but for Fig. 1a the relative performance drops after the financial crisis 2007, whereas for the truncation methods the drop in performance is less severe. The sharp drops in relative performance seem to be attenuated by truncation. Additionally, BMA consistently has one of the lowest DCSFE for all methods, which is in line with the results from Table 3 and Table 4.

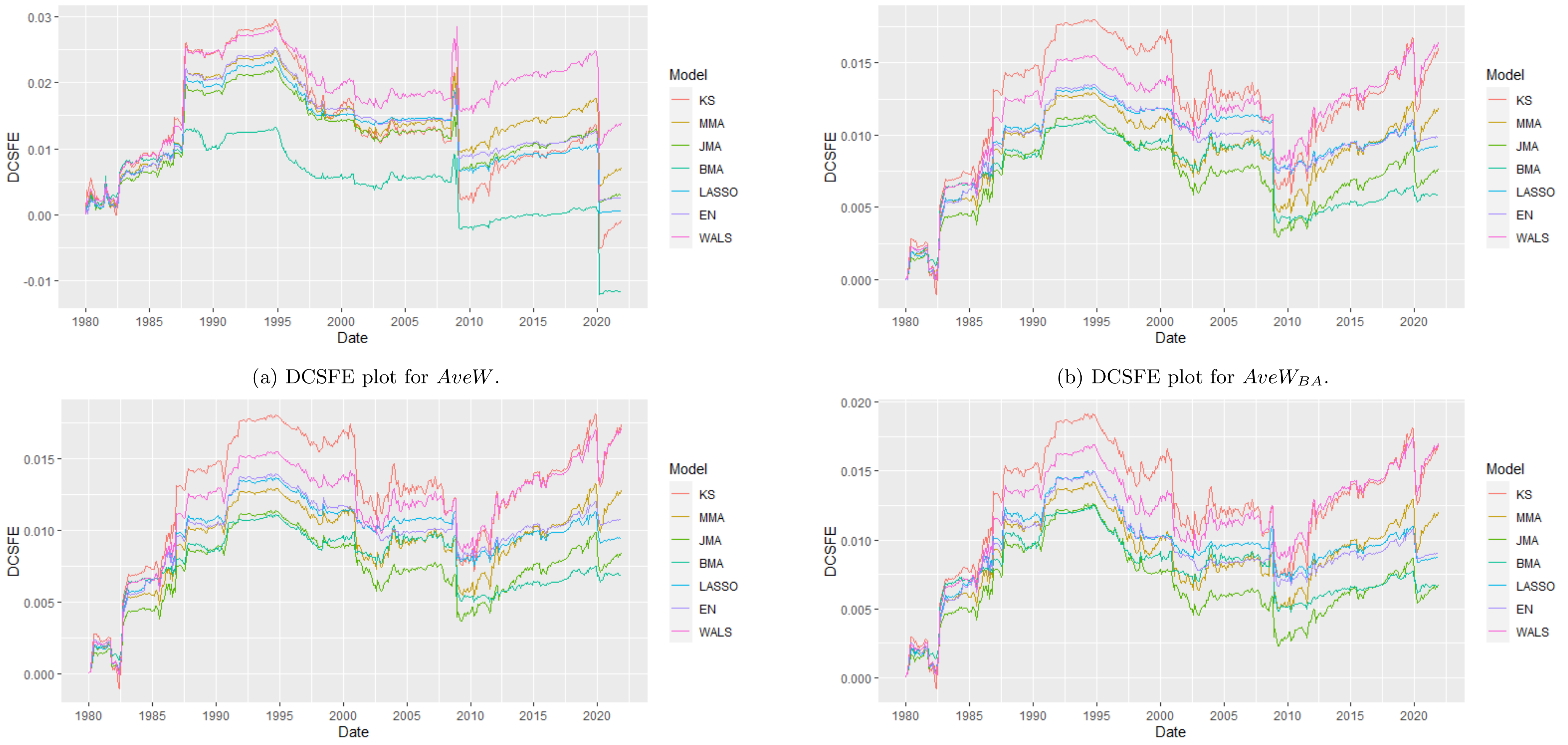
5.2 Analysis Recent Data and COVID-19 Crisis

As a result of the digital revolution and other rapid technological advancements, there have been large changes in the composition of the American stock market. New types of companies have taken over the market, think of Amazon, Apple and Google. Therefore, in this subsection we will analyse the performance of the models using data from [1960:01-1979:12] as our in-sample, and using considering [1980:01-2021:12] for our out-of-sample analysis. Additionally, using this data we will analyse the performance of the methods during the COVID-19 crisis, as it has been an extremely volatile period.

5.2.1 Method and Model Performances Using Recent Data

The results of the analysis using only the data from 1960 and onward are given in table 5. In contrast with the results in Table 3, for all of the sophisticated models, the largest R^2_{OS} values are obtained using a truncation method, more specifically, applying truncation after averaging. The Clark and West (2007) test statistic further confirms this, as for $AveW_{AA}$ the test rejects the null hypothesis of a negative R^2_{OS} much more strongly compared to no truncation for every model similar to Table 3. Additionally, this improvement in forecasting accuracy using the method $AveW_{AA}$ is reflected in the annualized utility gains, given in Table A.2 in the Appendix, as the utility gains are much higher compared to truncation after shrinkage.

Next, we examine the forecasting accuracy of the different models and methods using DCSFE, given in Fig. 2. Similar to 1, the DCSFE peaks attained by the average window method without shrinkage in Fig. 2a are much larger than those attained by any of the truncation methods, for all models except BMA. However, once more large drops in DCSFE seem to be attenuated by truncation. This is evident from the fact that for the models estimated using truncation, the drops after the 2007 financial crisis and in the beginning of the COVID-19 crisis are much smaller. Moreover, this is reflected in the R^2_{OS} values of the different methods during NBER dated recessions, given in Table 8 in the Appendix, as they see large increases after applying any of the truncation methods.



(a) DCSFE plot for $AveW$.

(b) DCSFE plot for $AveW_{BA}$.

(c) DCSFE plot for $AveW_{AA}$.

(d) DCSFE plot for $AveW_{AS}$.

Fig. 2: Time series plots of the forecasting accuracy of the different models, forecasted using the different AveW methods, with and without truncation. The accuracy is given in terms of DCSFE, which denotes the difference between the cumulative squared forecast error of the benchmark model and the sophisticated model. The greater the difference, the better the performance of the model. For each model we consider a linear combination of a sophisticated method with the historical average, estimated using the AveW method from H. Zhang et al. (2020) using shrinkage factor $\delta = 0.5$. We set the number of estimation windows $m = 10$. In panel (b), (c), and (d) we truncate negative return forecasts before averaging (BA), after averaging (AA), and after shrinkage (AS), respectively. The sophisticated models we consider are the kitchen sink model (KS), which is a standard multivariate regression, Mallows model averaging (MMA), by Hansen (2007), jackknife model averaging (JMA), by Hansen and Racine (2012), Bayesian model averaging with diffuse priors, by Raftery, Madigan, et al. (1997), LASSO (Tibshirani 1996) and elastic net (EN) (Zou and Hastie 2005) with fivevold cross-validation, and weighted-average least squares by Magnus, Powell, et al. (2010). The out-of-sample evaluation period we consider is [1980:01-2021:12].

Table 6: R^2_{OS} of average window method with shrinkage methods with and without truncation during the COVID-19 crisis.

Methods	R^2_{AveW}	$R^2_{AveW,BA}$	$R^2_{AveW,AA}$	$R^2_{AveW,AS}$
Kitchen Sink	-18.43	**1.61**	1.57	0.86
MMA + HA	-13.51	1.19	**1.20**	0.50
JMA + HA	-12.72	**-0.61**	**-0.61**	-1.32
BMA + HA	-17.64	0.10	**0.11**	-0.60
LASSO + HA	-9.08	**-0.09**	-0.46	-1.16
Elastic Net + HA	-13.31	**-0.19**	-0.31	-1.01
WALS + HA	-13.78	2.16	**2.17**	1.46

Note: This table details the out-of-sample forecasting accuracy, R^2_{OS}, of a combination of the sophisticated model forecasts, estimated using the average window method by Pesaran and Timmermann (2007), and historical average (HA), with shrinkage factor $\delta = 0.5$, compared to the benchmark model, HA. The out-of-sample evaluation period is [2021:02-2022:12]. For $R^2_{AveW,BA}$, $R^2_{AveW,AA}$, and $R^2_{AveW,AS}$, negative return forecasts are set to 0, similar to the truncation method proposed by Campbell and Thompson (2008), before averaging (BA), after averaging (AA), and after applying shrinkage (AS), respectively. For each model, the method with the largest R^2_{OS} is written in boldface, and for each method, the model with the largest R^2_{OS} is underlined. Significance codes: *: 0.1; **: 0.05; ***: 0.01, obtained using the test statistic from Clark and West (2007), which tests $H_0 : R^2_{OS} < 0$, with $H_1 : R^2_{OS} \geq 0$.

Additionally, after truncation, Bayesian model averaging estimates see a relative increase in DCSFE compared to the other model. Where in Fig. 1a it is clearly the worst performing model, in Fig. 1b, 1c and 1d it achieves a similar difference in cumulative squared forecast errors as jackknife model averaging. Furthermore, after truncation the kitchen sink model seem to get a relative boost in performance, compared to the weighted-average least squares. These relative increases in performance of the BMA and KS models are also reflected in Table 5, as the R^2_{OS} of both methods see a much larger increase after truncation compared to the other models.

5.2.2 Method and Model Performances During the COVID-19 Crisis

Lastly, we will evaluate the impact of the return restrictions based on its effect on the performance of the models during the COVID-19 crisis. From the second column it is clear that the average window method without truncation performs very poorly in this time-period. However, it seems as if the overall the gains made by applying truncation are much larger compared to the results using pre-COVID data. This supported by Fig. 2, as is the drop in DCSFE in the beginning of 2020 is much larger in Fig. 2a than in Fig. 2b, 2c and 2d. Per Table 6, either truncation before averaging or after averaging seems to lead to the largest gain in forecasting accuracy. While none of the R^2_{OS} are significantly larger than 0, as per the Clark and West (2007) test statistic, it does seem as if truncation leads to large gains in forecasting accuracy.

The corresponding utility gains expressed in terms of changes in CER difference are given in Table 7. As opposed to the utility changes in Tables 4 and A.2, some models are outperformed by the benchmark model in terms of CER. On the other hand, the Sharpe ratios are larger than those from the previous results. Another difference between the COVID-19 crisis and the other out-of-sample evaluation period, is that applying truncation, before or after averaging, leads to an increase in annualized utility gains for all models. Previously it lead to decreased utility gains. Using risk factors $\gamma = 2, 4$, similar utility gains are found after truncation. These results are given in Tables 9 and 10 in the Appendix.

Table 7: Annualized utility gain and Sharpe ratios of average window methods with truncation during the COVID-19 crisis with riskfactor $\gamma = 3$

Methods	$AveW_{(AS)}$		$AveW_{BA}$		$AveW_{AA}$	
	$\varDelta(ann\%)$	Sharpe	$\varDelta(ann\%)$	Sharpe	$\varDelta(ann\%)$	Sharpe
Kitchen Sink + HA	3.36	0.34	**6.39**	0.37	6.18	0.37
MMA + HA	1.87	0.32	4.63	0.36	**4.67**	0.36
JMA + HA	-2.29	0.28	**0.48**	0.31	**0.48**	0.31
BMA + HA	-1.41	0.33	1.30	0.38	**1.34**	0.38
LASSO + HA	-5.57	0.24	**-1.61**	0.30	-2.84	0.28
Elastic Net + HA	-4.72	0.25	**-1.58**	0.30	-1.99	0.30
WALS + HA	3.03	0.34	**5.85**	0.37	**5.85**	0.37

Note: This table denotes the annualized utility gains for the combinations of the sophisticated models with the historical average (HA) forecast, with shrinkage factor $\delta = 0.5$. It is expressed as a difference in Certainty Equivalent Return (CER) between the combination and the benchmark model, HA. Furthermore it denotes the Sharpe ratio, which is defined as the mean of the realized returns in excess of the risk-free rate, divided by the standard deviation of the excess returns. We implement the truncation method by Pesaran and Timmermann (2007) at three different stages, before averaging (BA), after averaging (AA), and after shrinkage (AS). By construction, implementing truncation after shrinkage and no truncation give the same results, as negative returns are assigned weight $w = 0$ in Eq. 16. The weights are determined using a five year rolling window variance esitmate $\hat{\sigma}^2$, and risk factor $\gamma = 3$. For each model, the method with the largest $\varDelta(ann\%)$ is written in boldface, and for each method, the model with the largest $\varDelta(ann\%)$ is underlined. The out-of-sample evaluation period is [1957:01-2016:12].

6 Conclusions

The main goal of this paper was to answer the following research question:

> **RQ.1** How does imposing the restrictions from Campbell and Thompson (2008) affect the methods from H. Zhang et al. (2020)?

In order to answer this research question, we consider several methods to estimate 7 sophisticated models. Firstly, for an expanding window approach or the average window method without shrinkage, we find that setting negative excess forecast returns to 0 leads to an increase in R^2_{OS} for all models under consideration. For the average window method with shrinkage, we find similar results. We impose restrictions on negative return forecasts at three separate stages of the average window method, namely before averaging, after averaging and after shrinkage. When considering a large sample, we find that only for some models truncation leads to an improvement of the R^2_{OS}. However, the Clark and West (2007) test statistic indicates that the difference between the benchmark model and sophisticated models becomes more significant after implementing the restrictions. Moreover, using a more recent dataset we do find that the return restrictions improve forecasting accuracy as well as utility gains across the board, relative to the benchmark model. This could be due to truncation being relatively more effective when considering a smaller in-sample period. Additionally, it seems that during NBER dated recessions restricting negative returns forecasts has a larger effect. In order to further solidify this suspicion, we perform a case study on the performance of the methods during the COVID-19 crisis, using the following research question:

> **RQ.2** How do the results of applying truncation to the methods from H. Zhang et al. (2020) differ during the COVID-19 pandemic compared to pre-COVID19 times?

Applying the same methods, we find that during the COVID-19 crisis truncation improves the out-of-sample prediction accuracy as well as our utility measure across the board. Compared to the results from the other samples, the gains in forecasting accuracy and utility are estimated to be much larger. However, using the Clark and West (2007) test statistic, none of the estimated values for the out-of-sample R^2 were found to be significantly larger than 0. This could be a result of the out-of-sample size being too small.

For further research, one could examine how using a different measure for utility affects the results, as the CER only takes into account the mean and variance of the returns. For instance, one could add transaction costs such that it resembles a real life investment setting more closely. Furthermore, Pettenuzzo et al. (2014) devised a new approach with respect to introducing economic constraints to the forecasting of stock returns, by using constraints on equity premia and Sharpe ratio's to modify the posterior distribution of the parameters. Perhaps, combining this approach with the methods from H. Zhang et al. (2020) could improve on the methods from this paper. Additionally, in this paper we consider only 7 sophisticated model, however, there are other and newer models to consider, such as the SLOPE model by Kremer et al. (2020), to see how they compare to the models used in this paper. One could also consider using a different dataset, different explanatory variables or a different stock index as a robustness check. Lastly, perhaps the combination of the average window method with truncation could be applied in other fields where similar theoretical restrictions can be made for the dependent variable.

References

Amini, F., & Hu, G. (2021). A two-layer feature selection method using genetic algorithm and elastic net. *Expert Systems with Applications*, *166*.

Azar II, A. M. (2020). *Determination that a public health emergency exists*. Retrieved May 10, 2022, from https://www.phe.gov/emergency/news/healthactions/phe/Pages/2019-nCoV.aspx

Basuony, M. A., Bouaddi, M., Ali, H., & Emadeldeen, R. (2021). The effect of covid-19 pandemic on global stock markets: Return, volatility, and bad state probability dynamics. *Journal of Public Affairs*, e2761.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 603–618.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Center for Research in Security Prices, L. (2022). *Data*. Retrieved April 10, 2022, from https://www.crsp.org/resources/data

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, *138*(1), 291–311.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B*, *45*(3), 311–335.

Dahl, D. B. (2013). *xtable: Export tables to LaTeX or HTML* [http://xtable.r-forge.r-project.org/].

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Ganie, I. R., Wani, T. A., & Yadav, M. P. (2022). Impact of covid-19 outbreak on the stock market: An evidence from select economies. *Business Perspectives and Research.*

Godoi, C. K., Marcon, R., & Barbosa daSilva, A. (2005). Loss aversion: A qualitative study in behavioural finance. *Managerial Finance.*

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189.

Hansen, B. E., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, *167*(1), 38–46.

Hlavac, M. (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3.* Retrieved May 10, 2022, from https://CRAN.R-project.org/package=stargazer

Hughey, J. J., & Butte, A. J. (2015). Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Research*, *43*(12), e79–e79.

Jackson, C. H., Thompson, S. G., & Sharples, L. D. (2009). Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society*, *172*(2), 383–404.

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

Kotze, K. (2020). *KevinKotze/tsm: Time Series Modelling.* Retrieved May 10, 2022, from https://rdrr.io/github/KevinKotze/tsm/

Kremer, P., Brzyski, D., Bogdan, M., & Paterlini, S. (2020). Sparse index clones via the sorted l1-norm. *Available at SSRN 3412061.*

Magnus, J. R., & De Luca, J. (2013). *Weighted-average least squares.* Retrieved May 10, 2022, from https://www.janmagnus.nl/items/WALS.pdf

Magnus, J. R., Powell, O., & Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of econometrics*, *154*(2), 139–153.

Morin, R.-A., & Suarez, A. F. (1983). Risk aversion revisited. *The journal of finance*, *38*(4), 1201–1216.

NBER. (2022). *Business cycle dating.* Retrieved May 10, 2022, from https://www-nber-org.eur.idm.oclc.org/research/business-cycle-dating

Nikodinoska, D., Käso, M., & Müsgens, F. (2022). Solar and wind power generation forecasts using elastic net in time-varying forecast combinations. *Applied Energy*, *306*, 117983.

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC proceedings*, *6*(2), 1–6.

Pesaran, M. H., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, *137*(1), 134–161.

Pettenuzzo, D., Timmermann, A., & Valkanov, R. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics*, *114*(3), 517–553.

Priestley, M. B. (1981). *Spectral analysis and time series: Univariate series* (Vol. 1). Academic press.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174.

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. Y. (2022). *Bayesian model averaging.* https://cran.r-project.org/web/packages/BMA/BMA.pdf

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191.

Rapach, D., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, *23*(2), 821–862.

Schomaker, M., & Heumann, C. (2014). Model selection and model averaging after multiple imputation. *Computational Statistics and Data Analysis*, *71*, 758–770.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of business*, *39*(1), 119–138.

Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and bayesian model averaging. *Journal of the American Statistical Association*, *105*(489), 25–35.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288.

Wan, A. T., Zhang, X., & Zou, G. (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics*, *156*(2), 277–283.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

Wickham, H., & Bryan, J. (2022). *readxl: Read Excel Files.* Retrieved May 10, 2022, from https://readxl.tidyverse.org,%20https://github.com/tidyverse/readxl

Yu, W., & Zhao, C. (2019). Robust monitoring and fault isolation of nonlinear industrial processes using denoising autoencoder and elastic net. *IEEE Transactions on Control Systems Technology*, *28*(3), 1083–1091.

Zhang, H., He, Q., Jacobsen, B., & Jiang, F. (2020). Forecasting stock returns with model uncertainty and parameter instability. *Journal of Applied Econometrics*, *35*(5), 629–644.

Zhang, X., Wan, A. T., & Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, *174*(2), 82–94.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, *67*(2), 301–320.

A Appendix

A.1 Description of R and Matlab code

In this section we will give a short explanation on the MATLAB and R code given in the folder 'code_thesis.zip'. Firstly, navigate to the main folder 'Code Thesis' and run the script 'Variables.R'.

Then, to obtain the results from Section 5.1.1 and 5.1.2, firstly run the script 'WalsRep.m' in the 'MATLAB' subfolder, and afterwards run the script Replication.r in the main folder.

Next, to obtain the results from Section 5.2.1, firstly run the script 'WalsUpd.m' in the 'MATLAB' subfolder, and afterwards run the script Updated.r in the main folder.

Finally, to obtain the results from Section 5.2.2, firstly run the script 'WalsCov.m' in the 'MATLAB' subfolder, and afterwards run the script Covid.r in the main folder.

Using an HP ENVY 13-ba0750nd laptop, we found the total running time to be approximately 3 hours.

A.2 Annualized Utility Gains and Sharpe Ratio's Using Recent Data

Methods	$AveW_{(AS)}$		$AveW_{BA}$		$AveW_{AA}$	
	$\Delta(ann\%)$	Sharpe	$\Delta(ann\%)$	Sharpe	$\Delta(ann\%)$	Sharpe
Kitchen Sink + HA	3.48	0.18	3.20	0.17	**3.77**	0.18
MMA + HA	2.86	0.17	2.70	0.16	**3.25**	0.17
JMA + HA	1.70	0.14	1.94	0.15	**2.50**	0.16
BMA + HA	2.39	0.16	2.35	0.16	**2.58**	0.16
LASSO + HA	1.82	0.15	1.93	0.15	**2.26**	0.16
Elastic Net + HA	1.54	0.14	1.69	0.14	**2.23**	0.16
WALS + HA	3.37	0.18	3.16	0.17	**3.57**	0.18

Note: This table denotes the annualized utility gains for the combinations of the sophisticated models with the historical average (HA) forecast, with shrinkage factor $\delta = 0.5$. It is expressed as a difference in Certainty Equivalent Return (CER) between the combination and the benchmark model, HA. Furthermore it denotes the Sharpe ratio, which is defined as the mean of the realized returns in excess of the risk-free rate, divided by the standard deviation of the excess returns. We implement the truncation method by Pesaran and Timmermann (2007) at three different stages, before averaging (BA), after averaging (AA), and after shrinkage (AS). By construction, implementing truncation after shrinkage and no truncation give the same results, as negative returns are assigned weight $w = 0$ in Eq. 16. The weights are determined using a five year rolling window variance esitmate $\hat{\sigma}^2$, and risk factor $\gamma = 3$. For each method, the model with the largest $\Delta(ann\%)$ is written in boldface. The out-of-sample evaluation period is [1980:01-2021:12].

A.3 Out-of-Sample R^2 Values During NBER dated Recessions

Table 8: R^2_{OS} of average window method with shrinkage methods with and without truncation during NBER dated recessions.

Methods	R^2_{AveW}	$R^2_{AveW,BA}$	$R^2_{AveW,AA}$	$R^2_{AveW,AS}$
Kitchen Sink + HA	-9.37	-2.29	-1.81	**-1.75**
MMA + HA	-4.86	-1.23	-0.94	**-0.84**
JMA + HA	-4.71	-1.20	**-0.81**	-0.96
BMA + HA	-5.53	-0.16	0.25	**0.50**
LASSO + HA	-4.95	-0.28	0.04	**0.16**
Elastic Net + HA	$\underline{\text{-3.96}}$	$\underline{\text{0.28}}$	$\underline{\textbf{0.72}}$	$\underline{\text{0.68}}$
WALS + HA	-4.42	-0.95	-0.72	**-0.58**

Note: This table denotes the out-of-sample forecasting accuracy, R^2_{OS}, of a combination of the sophisticated model forecasts, estimated using the average window method by Pesaran and Timmermann (2007), and historical average (HA), with shrinkage factor $\delta = 0.5$, compared to the benchmark model, HA, during NBER dated recessions. The out-of-sample evaluation period is [1980:01-2021:12]. For $R^2_{AveW,BA}$, $R^2_{AveW,AA}$, and $R^2_{AveW,AS}$, negative return forecasts are set to 0, similar to the truncation method proposed by Campbell and Thompson (2008), before averaging (BA), after averaging (AA), and after applying shrinkage (AS), respectively. For each model, the method that obtains the largest R^2_{OS} is written in boldface, and for each method, the model with the largest R^2_{OS} is underlined. Significance codes: *: 0.1; **: 0.05; ***: 0.01, obtained using the test statistic from Clark and West (2007), which tests $H_0 : R^2_{OS} < 0$, with $H_1 : R^2_{OS} \geq 0$.

A.4 Utility Gains and Sharpe Ratio's Using Other Risk Factors

Table 9: Annualized utility gain and Sharpe ratios of average window methods with truncation during the COVID-19 crisis with riskfactor $\gamma = 2$

Methods	$AveW_{(AS)}$		$AveW_{BA}$		$AveW_{AA}$	
	$\mathit{\Delta}(ann\%)$	Sharpe	$\mathit{\Delta}(ann\%)$	Sharpe	$\mathit{\Delta}(ann\%)$	Sharpe
Kitchen Sink + HA	-0.77	0.36	2.91	0.40	**3.40**	0.41
MMA + HA	-1.44	0.35	**2.82**	0.40	2.72	0.40
JMA + HA	-1.34	0.35	**2.82**	0.40	**2.82**	0.40
BMA + HA	-5.86	0.32	-1.79	0.37	**-1.75**	0.37
LASSO + HA	-4.59	0.33	1.41	**0.39**	-0.46	0.38
Elastic Net + HA	-4.32	0.33	**0.42**	0.38	-0.19	0.38
WALS + HA	-1.01	0.36	2.85	0.40	**3.15**	0.40

Note: This table denotes the annualized utility gains for the combinations of the sophisticated models with the historical average (HA) forecast, with shrinkage factor $\delta = 0.5$. It is expressed as a difference in Certainty Equivalent Return (CER) between the combination and the benchmark model, HA. Furthermore it denotes the Sharpe ratio, which is defined as the mean of the realized returns in excess of the risk-free rate, divided by the standard deviation of the excess returns. Negative excess return forecasts are set to 0, similar to the truncation method proposed by Campbell and Thompson (2008). We implement the truncation at three different stages, before averaging (BA), after averaging (AA), and after shrinkage (AS). By construction, implementing truncation after shrinkage and no truncation give the same results, as negative returns are assigned weight $w = 0$ in Eq. 16. The weights are determined using a five year rolling window variance esitmate $\hat{\sigma}^2$, and risk factor $\gamma = 2$. For each method, the model with the largest $\mathit{\Delta}(ann\%)$ is written in boldface. The out-of-sample evaluation period is [1957:01-2016:12].

Table 10: Annualized utility gain and Sharpe ratios of average window methods with truncation during the COVID-19 crisis with riskfactor $\gamma = 4$

Methods	$AveW_{(AS)}$		$AveW_{BA}$		$AveW_{AA}$	
	$\mathit{\Delta}(ann\%)$	Sharpe	$\mathit{\Delta}(ann\%)$	Sharpe	$\mathit{\Delta}(ann\%)$	Sharpe
Kitchen Sink	1.40	0.30	3.69	0.33	3.53	0.33
MMA	-0.83	0.27	1.25	0.30	1.28	0.30
JMA	-6.46	0.19	-4.40	0.22	-4.40	0.22
BMA	-1.05	0.33	0.98	0.38	1.01	0.38
LASSO	-4.17	0.24	-1.20	0.30	-2.13	0.28
Elastic Net	-3.54	0.25	-1.18	0.30	-1.49	0.30
WALS	1.26	0.30	3.38	0.33	3.38	0.33

Note: This table denotes the annualized utility gains for the combinations of the sophisticated models with the historical average (HA) forecast, with shrinkage factor $\delta = 0.5$. It is expressed as a difference in Certainty Equivalent Return (CER) between the combination and the benchmark model, HA. Furthermore it denotes the Sharpe ratio, which is defined as the mean of the realized returns in excess of the risk-free rate, divided by the standard deviation of the excess returns. We implement the truncation method by Pesaran and Timmermann (2007) at three different stages, before averaging (BA), after averaging (AA), and after shrinkage (AS). By construction, implementing truncation after shrinkage and no truncation give the same results, as negative returns are assigned weight $w = 0$ in Eq. 16. The weights are determined using a five year rolling window variance esitmate $\hat{\sigma}^2$, and risk factor $\gamma = 4$. For each method, the model with the largest $\mathit{\Delta}(ann\%)$ is written in boldface. The out-of-sample evaluation period is [1957:01-2016:12].