Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Seminar Machine Learning

Combining forecasts using peLasso and peElasticNet methods

[redacted], Menno Smit
([redacted], 580595ms)

Supervisor:	Hakan Akyuz
Date final version:	May 26, 2025

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

Abstract

Enhancing prediction accuracy is of major importance for decision-making across many disciplines and is therefore an intensively researched topic. Combining forecasts can be potent for enhancing prediction accuracy when a set of different predictors exists. In past research, it was shown that the average of all forecasts as a forecast combiner can enhance prediction significantly by reducing risk. More recent research has shown that the peLasso method can be used as a more effective method to make forecaster combinations when applied to the European Central Bank Survey of Professional Forecasters to forecast the real GDP of the European Zone. This paper expands upon the area of the peLasso method by implementing a variety of adaptations to the different steps of peLasso, within the same dataset context. This includes the replacement of Lasso with Elastic Net for step 1 (trimming phase) and the introduction of egalitarian Elastic Net, rmseLasso, rmseRidge, and rmseElasticNet as different procedures for step 2 (forecast combination phase). In total, these adaptations give eleven new methods, of which four have simultaneously adapted a change to steps 1 and 2. The results show that all newly introduced methods achieve a higher prediction accuracy than the simple average as a forecast combiner. Compared to the benchmark peLasso however, only the peLasso adaptation involving Elastic Net as a step 2 replacement achieves a forecast accuracy as high as the original peLasso method, which implies that the newly introduced methods do not improve prediction accuracy for the given data set, but can provide as an alternative to the existing methods.

1 Introduction

The quest for making accurate predictions has always been a challenge people strive to overcome. While being able to make a prediction is something everyone can achieve, the ability to make the most precise forecasts using historical data is what researchers are aiming for. The well-known Maya, as early as 3500 years ago, showcased their excellent skills in mathematics and astronomy by predicting optimal harvest times and lunar eclipses. In modern times, human beings try to predict all kinds of things, from the inflation rate of the upcoming year (Stock & Watson, 1999) to the outcome of the presidential election (Abramowitz, 1988). Even though today's society might try to predict different phenomena and have different underlying motives than the Maya, one question has remained central when we make predictions. *How can we leverage past information to make the best possible predictions?*

The answer to this question is important and highly relevant to people in many fields. While the ancients prioritised predictions for optimal harvest timing, modern motivations are diverse, often involving decision-making. This is particularly pertinent when decisions impact numerous stakeholders, which is often the case in institutions such as governments, pension funds, and banks. Achieving high forecast accuracy is then often one of the primary goals since this indicates the forecasts provide valuable information for informed decision-making. Consequently, enhancing the accuracy of predictions is a widely researched topic. In this paper, we aim to add meaningful insights to the existing literature on predictive accuracy.

The original paper written by Diebold & Shin (2019) focuses on using the Lasso technique to combine forecasts by only selecting relevant forecasts and averaging them. In the rest of the paper, the paper by Diebold & Shin (2019) is referred to as the original paper. They first use Lasso in step 1 for selection and step 2 to shrink the coefficients of the remaining survivor forecasts. The RMSE values of eLasso and eRidge both have the same RMSE value as the simple average, as the penalty term is very high. Hence, their approach concludes by shrinking the selected forecasters' coefficients towards each other, promoting a more egalitarian or average-based forecast combination.

In contrast, this paper extends this foundation by evaluating whether implementing Elastic Net can improve the first step of peLasso. This modification aims to address multicollinearity among predictors and maintain a balance between forecast selection and coefficient shrinkage, potentially leading to better predictive performance.

Furthermore, for step 2 this paper introduces the egalitarian Elastic Net model which is an adaptation to the egalitarian Lasso and egalitarian Ridge. Additionally, this paper also introduces the rmseLasso, rmseRidge and rmseElasticNet as shrinkage procedures by using a model-specific rmse-dependent penalty. This approach hypothesises that forecasters with lower historical prediction errors will perform better in the future and, hence, should be assigned more weight.

Our research aims to explore whether modifications to the peLasso method can enhance forecast accuracy. Consequently, our research question is:

Can adaptations to peLasso lead to improved forecast combinations?

We examine this question by first replicating the results from the original paper using standard peLasso, and by subsequently implementing adaptations of peLasso to see whether this leads to a decrease in RMSEs within the same data context. The data we use for this matter is the European Central Bank's quarterly Survey of Professional Forecasters and the real GDP of the Euro Zone, which is the same dataset as in the original paper.

After implementing the procedures from the original paper and the adaptations to the peLasso, we find that all method adaptations perform better than a simple average; however, no adaptations surpass the performance of the standard peLasso method, specifically peLasso (Lasso, AVG), peLasso (Lasso, eRidge), and peLasso (Lasso, eLasso). The optimal results were achieved using the peLasso models, with a minor adaptation, peLasso (Lasso, eEN), which exhibited the lowest RMSE among all implemented adaptations. Additionally, the peElasticNet models underperform compared to the peLasso models. The RMSE-based shrinking methods introduced in section 5.2.1, such as peLasso(Lasso, rmseLasso), underperform compared to shrinking models that shrink towards the average, as proposed in the original paper.

The rest of this paper is organised as follows: First, section 2 reviews existing literature on forecast combination, penalised regression, and the peLasso. Subsequently, section 3 covers the sub-questions of the research and the theoretical motivation for different adaptations of the original peLasso method. Section 4 describes the data that is used for the research. Then, section 5 introduces the methodology that is used throughout section 6, which discusses the findings of the research and implications. Finally, section 7 concludes the research and gives some recommendations for future research.

2 Literature Review

In the contemporary field of forecasting, researchers often use a model to explain the variation in a process, and as a basis for making future inferences. Traditionally, forecasting focuses on selecting a singular optimal model to generate the best possible forecast. However, the practice of combining forecasts from various models, known as *forecast combining*, has gained recognition for its substantial potential in enhancing forecast accuracy (Elliott & Timmermann, 2016). The significance of forecast combining is notably illustrated in the M4 competition, as detailed by Atiya (2020), who provides graphical evidence of its success. At the beginning of forecast combination research, simply using the average of all available forecasts, which is referred to in the literature as the *average weights puzzle*, is a popular method due to its empirical good performing forecasts with low Mean Squared Prediction Error (MSPE) (Clemen, 1989). Despite its empirical success, Aruoba et al. (2012) highlights its theoretical sub-optimality, suggesting inherent limitations. Additionally, Hibon & Evgeniou (2005) shows that while combining forecasts typically reduces overall risk, it seldom surpasses the accuracy of the best individual forecast. Therefore, Kourentzes et al. (2019) studies the idea of selecting a relevant subset or *pool* of forecasts from a larger set of forecasts in combination with combining those forecasts to create a new forecast. They show that this combination is beneficial since it can lead to a significant increase in forecast accuracy.

Diebold & Shin (2019) introduces a novel method, called the *partially-egalitarian Lasso (peLasso)* approach. This approach utilises penalised regression to choose a pool of forecasts from a bigger set of forecasts and to combine them into a new forecast. The development of peLasso is inspired by several empirical findings: the effective empirical performance of the average weights puzzle, the good performance of the *Lasso regression* in selecting relevant indicators, and the utility of *Ridge regression* in shrinking coefficients towards zero. Both types of regressions fall in the set of *penalized regressions*, with the Ridge regression originating from the 1970s (Hoerl & Kennard, 1970), and the Lasso regression from the 1990s (Tibshirani, 1996), each adding an extended penalty term to the original OLS objective function. An evolution of the Lasso, the *Adaptive Lasso* (aLasso), is proposed by Zou (2006). This model deals with both the occasional inconsistency of Lasso in variable selection and its tendency to uniformly shrink all parameters. A further advanced development in Lasso technology is the *Bayesian Adaptive Lasso* model, which incorporates the adaptive Lasso strategy within a Bayesian framework that allows for both adaptive regularization and facilitates the incorporation of prior knowledge and uncertainty about the model parameters (Leng et al., 2014). It is important to recognise that both Ridge and Lasso are primarily used to address issues of high dimensionality and to produce forecasts with a lower RMSE than those obtained via traditional OLS regression. The introduction of penalized regression methods in the field of forecast combining, in combination with the result that the peLasso is found to outperform the traditional well-performing average weights puzzle, makes Diebold & Shin (2019) a good contribution to literature and a promising direction for future research.

3 Background

We divide our main research question, as discussed in section 1, into multiple sub-questions. The sub-questions are later answered in Section 6.

Sub-question 1: Can we improve the first step of peLasso?

The first step of the original peLasso method relies on using a Lasso regression to choose a set of forecasters from a bigger forecasters pool. It is insightful to explore alternative selection methods for the first step, while maintaining the second step unchanged (ceteris paribus), to determine if this can lead to improved forecast combination accuracy. We hypothesise that we can enhance the trimming by using the *Elastic Net* procedure, which is a combination of both the Ridge and Lasso procedure (Zou & Hastie, 2005). This approach can be advantageous since traditional Lasso struggles with multicollinearity, often allowing certain predictors to disproportionately dominate the model. The Elastic Net has the advantage that it balances out the effect of Lasso, which selects certain forecasts, with the effect of Ridge which only shrinks all forecast coefficients.

Sub-question 2: Can we improve the second step of peLasso?

The second step of the original peLasso procedure utilises the eLasso or eRidge method to form forecast combinations by shrinking the OLS estimated coefficients of the selected predictors towards equality. It is insightful to explore alternative shrinkage methods in the second step, while maintaining the first step unchanged (ceteris paribus), to ascertain if such modifications can enhance forecast combination accuracy. Drawing on the rationale discussed earlier, we hypothesise that *eElasticNet* can prove beneficial in refining the second-step shrinkage by leveraging the combined advantages of eLasso and eRidge. Furthermore, we hypothesise that shrinkage to a different value than the average has a valuable impact on forecast accuracy. Hence, this paper introduces a family of shrinking methods, namely the rmseLasso, rmseRidge, and rmseElasticNet, which shrinks the coefficient of the forecaster with a penalty term that is related to the forecasters corresponding RMSE value. This approach hypothesises that forecasters with lower historical prediction errors should also perform better in the future, and hence should be assigned a higher weight. This approach should thus potentially outperform the simple average method used in traditional peLasso.

Sub-question 3: Can we improve accuracy by changing both step 1 and 2 simultaneously?

It is also insightful to investigate unnested models that incorporate simultaneous changes to both steps, to determine if this approach enhances forecast combination accuracy. The adaptations that are suggested as adaptations to steps one and two can also be implemented simultaneously to form a variety of new methods. Given the significant departure from the original peLasso framework, the outcomes of such an approach are challenging to predict. However, if the adaptations to both steps are individually effective, they potentially lead to improved forecast accuracy. In particular, synergistic effects may arise if better forecast trimming in the first step further increases the effectiveness of better shrinkage in the second step.

4 Data

For the replication part of our study, we utilise the identical dataset as employed in the original study, kindly provided to us by the authors. Specifically, the authors analyse data from the European Central Bank's quarterly Survey of Professional Forecasters, focusing on real GDP projections on a quarterly 1-year ahead percentage growth basis. The sample period spans from 1999Q1 to 2016Q2, during which the authors select the 23 most consistently responding forecasters, resulting in 70 observations. Missing observations are addressed using a linear filter technique as introduced by Genre et al. (2013). In this paper, no modifications or additional cleaning techniques are applied to the dataset beyond those already implemented in the original paper. Table 1 shows the resulting data. The first column enumerates the final selection of 23 forecasters, chosen based on their maximal frequency of responses. The second and third columns reveal a relative homogeneity among the selected forecasters in terms of means and standard deviations. In particular, the means range from 1.31 to 1.70, indicating a modest variation across the dataset. Likewise, the standard deviations range from 0.84 to 1.36, which underlines the consistency of the data. This suggests the 23 selected forecasters show a remarkable degree of agreement in their predictions.

In terms of the actual real GDP to compare the forecasts with, the original research uses a dataset provided by *Eurostat*, where all 'standard' revisions have been taken into account 100 days after the end of the quarter. We use this same dataset as in the original paper, covering also 70 observations for the same years for which the forecasts were gathered. The summary statistics of this dataset are shown in Table 2.

Table 1: Summary statistics of quarterly survey of professional forecasters from the European Central Bank.

Forecasters	Mean	St.deviation	Min	Max	Observations	RMSE
ID_1	1.52	0.99	-2.30	3.70	70	1.55
ID_2	1.70	1.02	-2.50	3.38	70	1.63
ID_4	1.62	1.09	-1.50	3.90	70	1.50
ID_5	1.52	1.02	-2.50	3.30	70	1.60
ID_7	1.37	1.31	-3.90	3.90	70	1.44
ID_15	1.53	0.96	-2.60	3.00	70	1.68
ID_16	1.44	1.07	-3.00	3.40	70	1.54
ID_20	1.51	1.08	-2.80	3.60	70	1.51
ID_24	1.45	1.02	-2.50	3.60	70	1.56
ID_26	1.31	1.36	-4.80	3.70	70	1.46
ID_29	1.58	0.84	-0.30	3.26	70	1.74
ID_31	1.55	0.85	-1.82	3.40	70	1.61
ID_36	1.64	0.86	-1.66	3.00	70	1.65
ID_37	1.48	1.09	-2.70	3.40	70	1.53
ID_39	1.47	1.09	-2.00	3.60	70	1.51
ID_42	1.41	1.04	-2.10	3.42	70	1.45
ID_48	1.55	1.03	-1.52	3.70	70	1.72
ID_52	1.40	1.12	-3.00	3.40	70	1.44
ID_54	1.51	1.06	-1.89	3.80	70	1.52
ID_85	1.46	0.97	-2.00	3.60	70	1.50
ID_89	1.55	0.96	-1.55	3.70	70	1.60
ID_94	1.33	1.26	-3.10	3.40	70	1.40
ID_95	1.52	1.03	-2.30	3.40	70	1.50

Table 2: Summary statistics of the actual GDP growth.

	Mean	St.deviation	Min	Max	Observations
Real GDP growth	1.34	1.99	-5.50	4.50	70

5 Methodology

This Section describes all methodologies that we used throughout the results. It starts by describing the methodology behind forecast combining and the standard peLasso method as the original paper. Subsequently, it discusses our extension which involves adaptations to both the first step, the trimming phase, and the second step, the shrinkage phase. The trimming phase literature is enriched using Elastic Net as a trimming method resulting in the peElastic-Net method, while the shrinkage phase literature is enriched with egalitarian Elastic Net and shrinkage methods based on the RMSE of individual forecasts which will be discussed in section 5.2.

5.1 Forecast combinations and the original peLasso model

Here, we discuss the existing theory of forecaster and forecast combinations, and subsequently the peLasso method for forming these combinations.

5.1.1 Forecast Combinations

The idea of forecast combinations is to combine information from different forecasts to construct a more accurate forecast than all forecasts individually. What different forecasts are based on is often unobserved to the forecast user, this is why the idea of combining forecasts, to bundle this *private information* comes into existence (Timmermann, 2006). We now introduce the notation that we use in this research paper. Let $\{f_1, f_2, ...f_k\}$ be the set of k different available *forecasters* from different models, all trying to forecast the same variable y. Then, a *forecaster combiner* can be seen as a function $g(.)$, which maps the input forecasters to a new *forecaster combination* $g(f_1, f_2, ..., f_k)$. Different forecaster combiners can be used, and have been examined in the past. In particular, Aiolfi et al. (2010) shows that an average weights puzzle, which falls in the set of *linear forecaster combiners*, is often very accurate empirically and outperforms many more complex combiners. This paper focuses on linear forecaster combiners only. Linear *forecast combinations* for a *forecast* at time t can be written as

$$g(f_{1,t}, f_{2,t}, \dots, f_{k,t}) = \sum_{i=1}^{k} \omega_i f_{i,t} \quad = \omega_1 f_{1,t} + \omega_2 f_{2,t} + \cdots + \omega_k f_{k,t},$$

where $f_{i,t}$ is the forecast made by forecaster f_i at time t, and ω_i is the scalar weight for forecaster i.

An often-used restriction for forecaster combiners is

$$\sum_{i=1}^{k} \omega_i = 1,$$

which enforces that all forecaster weights sum up to one.

Having introduced this notation, the equal weight puzzle that is mentioned earlier can be seen as a linear forecast combiner where $\omega_i = \frac{1}{k}$, for all $i \in \{1, 2, ..., k\}$.

Forecaster combiners can also be used for a subset of all possible forecasters. In this case, the number of selected forecasters is strictly less than the total number of forecasters, and only the selected forecasters have a weight ω_i unequal to zero.

One other forecaster combiner covered in Franses (2024) is combining forecasts using model information criteria. We prefer this since it gives better-fitting forecasters more weight in determining forecasts. As information criteria, we prefer the Akaike Information Criterion (AIC) over the Bayesian information criterion (BIC) since it is good at evaluating and predicting the accuracy of a forecaster when it is unclear whether the true underlying model is within the set of candidate models, which is often true when predicting Vrieze (2012). Kolassa (2011) describes one procedure of using the AIC for this particular matter. The weight for each forecast is then selected based on

$$\omega_i = \frac{e^{-\frac{1}{2}AIC_i}}{e^{-\frac{1}{2}AIC_1} + e^{-\frac{1}{2}AIC_2} + ... + e^{-\frac{1}{2}AIC_k}}.$$

The weight of this forecaster also satisfies the restriction of all forecasters' weights summing up to one.

5.1.2 The original peLasso model

In the original paper, the authors use the peLasso procedure for both choosing a subset of forecasters and subsequently finding a well-performing forecaster combiner in terms of accuracy. To do so, the method minimises

$$\hat{\omega}_{\text{peLasso}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k} \omega_i f_{i,t} \right)^2 + \lambda_1 \sum_{i=1}^{k} |\omega_i| + \lambda_2 \sum_{i=1}^{k} \left| \omega_i - \frac{1}{p(\omega)} \right| \right), \tag{1}$$

where ω is the *weighting vector* of the forecasts, and $t = \{1, 2, ..., T\}$ is the subsample that is considered for determining the optimal weights for the forecaster combiner at time $T + 1$, in which y_t is the actual observed value at time $t = \{1, 2, ..., T\}$, $p(\omega)$ is the number of selected forecasters (and thus $k = p(\omega)$ if all forecasters are selected), and λ_1 and λ_2 are constants which respectively penalise large weights on individual forecasters and deviations from the average weights puzzle solution. The objective function can be interpreted as a composition of different objective functions expressed in literature. Firstly, the first term of the equation is from a traditional OLS regression. Combining this part with the second term is the typical objective function of a Lasso regression. Secondly, the first term in combination with the last term can be interpreted as a Lasso regression, where the weight shrinks towards the average rather than zero.

$$\hat{\omega}_{\text{eLasso}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k} \omega_i f_{i,t} \right)^2 + \lambda_2 \sum_{i=1}^{k} \left| \omega_i - \frac{1}{p(\omega)} \right| \right), \tag{2}$$

Everything together the objective function both chooses which forecasters are selected, meaning with a weight unequal to zero, and shrinks the selected forecasters to the average.

As the authors explain, the objective function as is noted in equation 2 is difficult to optimise because the function is non-differentiable. Therefore, a two-step procedure can be used to simulate the procedure, which can be summarised by the following:

$$\textit{Step 1:} \quad \hat{\omega}_{\text{Lasso}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k} \omega_i f_{i,t} \right)^2 + \lambda_1 \sum_{i=1}^{k} |\omega_i| \right).$$

$$\textit{Step 2:} \quad \hat{\omega}_{\text{eLasso}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k'} \omega_i f_{i,t} \right)^2 + \lambda_2 \sum_{i=1}^{k'} \left| \omega_i - \frac{1}{p(\omega)} \right| \right).$$

Step 1 effectively selects the relevant forecasters via a Lasso procedure, after which $k' = p(\omega)$ forecasters remain, while step 2 chooses the optimal linear combination of forecasts, given the surviving forecasts using an *egalitarian-Lasso* (eLasso) procedure. This means that the weight of an unselected forecaster is set to zero in step 1 and therefore has no effect on step 2 (namely their weight stays zero). For the optimal forecaster combiner, the empirical good performance of the average forecast combiner is used to shrink the weights of the forecaster combiner via the penalty term.

In addition to the original peLasso method, the original paper also uses slight variations of the peLasso. They suggest it is possible to change step 2 to shrink via eRidge, which is:

$$\textit{Step 2: } \hat{\omega}_{\text{eRidge}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k'} \omega_i f_{i,t} \right)^2 + \lambda_2 \sum_{i=1}^{k'} \left(\omega_i - \frac{1}{p(\omega)} \right)^2 \right).$$

In the paper, they conclude that both variations lead to similar results, which are optimal for choosing a high λ_2, implying that complete shrinkage to the average gives forecast combinations a minimum forecast accuracy for both methods. The authors afterward decide to name both methods using either eLasso or eRidge for the second step 'peLasso' methods, emphasising that the first optimisation step is most important for the outcome of the method, and therefore the determinant of the method name. In this paper, the naming of newly introduced methods is done based on this idea.

5.2 Extension of peLasso

Given the methodology of the original peLasso procedures introduced in the original paper, now we discuss the extensions of the method. The methodology of the variations we propose is straightforward given the methods of the original peLasso and will now be discussed in detail. First, we discuss the methodology of adjustments to step 1 of peLasso, and then the methodology of adjustments to step 2. Extensions in step 1 selection phase As suggested earlier an adjustment to step 1, which is the trimming phase, is the use of the Elastic Net procedure as a replacement for the Lasso. This results in a method that uses a revised step 1 and standard step 2 and is called *peElasticNet*. The revised step 1 then looks as follows:

$$\textit{Step 1: } \hat{\omega}_{\text{ElasticNet}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k} \omega_i f_{i,t} \right)^2 + \lambda_1 \sum_{i=1}^{k} |\omega_i| + (1-\lambda_1) \sum_{i=1}^{k} \omega_i^2 \right).$$

All coefficients corresponding to the forecasters with non-zero weights are chosen for step 2. As motivated earlier, it is insightful to see whether this can enhance the trimming phase.

5.2.1 Extensions in the shrinkage phase

As described earlier, an approach for step 2, which is the shrinkage phase, is to replace the eLasso procedure with an eElasticNet procedure. The eElasticNet shrinks the weights to equality using a combination of the eLasso and eRidge procedures. Step 2 then becomes:

$$\textit{Step 2: } \hat{\beta}_{\text{eElasticNet}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k'} \omega_i f_{i,t} \right)^2 + \lambda_2 \sum_{i=1}^{k'} \left| \omega_i - \frac{1}{p(\omega)} \right| \right.$$
$$\left. + (1-\lambda_2) \sum_{i=1}^{k'} \left(\omega_i - \frac{1}{p(\omega)} \right)^2 \right),$$

where $k' = p(\omega)$ equals the number of leftover forecasts from step 1.

A second idea we motivate is to replace the average that is used as the shrinkage factor with a different value. The average is chosen as a good shrinkage factor by the original paper since it performs well in general forecast combinations. As is explained in section 5.1, we can use different weights to combine forecasts. It is insightful to use a weight determined by information criteria, which is discussed at the end of section 5.1, to shrink towards. There is however a slight complication since the dataset that we use does not directly provide the information criteria of the forecasts themselves. Our idea is therefore to use a different measure to measure the fit of the underlying models that produce the forecast in the dataset. It is highly likely that a model that performs well based on RMSE also has a low AIC. Therefore, we substitute the AIC with the RMSE and scale the formula appropriately, such that a weighting combination similar to the original idea is obtained. The second step of peLasso is changed to what we call the *rmseLasso*. In formula:

$$\textit{Step 2: } \hat{\omega}_{\text{rmseLasso}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k'} \omega_i f_{i,t} \right)^2 + \lambda_2 \sum_{i=1}^{k'} |\omega_i - \delta_i| \right),$$

in which

$$\delta_i(c) = \frac{e^{\left(\frac{c}{RMSE_i}\right)}}{e^{\left(\frac{c}{RMSE_1}\right)} + e^{\left(\frac{c}{RMSE_2}\right)} + \cdots + e^{\left(\frac{c}{RMSE_{k'}}\right)}}.$$

The new shrinkage factor in the formula is δ_i, and $c \geq 0$ is a constant to scale the formula flexibly. Furthermore, $RMSE_i$ is the Root Mean Prediction Error of forecast i, to which ω_i belongs as a weight. When c increases, $\delta_i(c)$ is influenced more by the difference in RMSE of the forecast i compared to the RMSE of other forecasters, which leads to shrinkage to a value further away from the average and more based on the RMSE of the forecaster.

As $c \to \infty$, the shrinkage factor for the model with the lowest RMSE ($RMSE_{min}$) approaches 1, while the shrinkage factor for all other models approaches 0. In mathematical terms:

$$\lim_{c \to \infty} \delta_i(c) = \begin{cases} 1 & \text{if } i = \arg\min_j \{RMSE_1, ..., RMSE_j\} \text{ with } i, j \in \{1, 2, ..., k'\} \\ 0 & \text{otherwise.} \end{cases}$$

This is because the exponential function $e^{-c \cdot RMSE_i}$ decreases towards 0 much more rapidly for models with higher RMSEs compared to the model with the lowest RMSE. This results in the lowest RMSE being selected exclusively for the forecast combination, and ignoring the rest of the forecasters.

Conversely, as $c \to 0$, the exponential term $e^{-c \cdot RMSE_i}$ for each model approaches 1, regardless of the RMSE values. Thus, all models receive equal weights in the forecast combination:

$$\lim_{c \to 0} \delta_i(c) = \frac{1}{k'}.$$

This scenario reflects a simple average of all models, where each model contributes equally to the combined forecast, irrespective of its individual accuracy.

The RMSE is made such that it is based on a sample short before the forecast at time $T+1$, since this offers the most valuable information, as is found in the original paper. When the RMSE of a particular forecaster is low compared to the rest, the $\delta_i(c)$ is set higher, which results in weight shrinkage to a higher value compared to the rest. By using this type of shrinkage, we can use the past performance very close to the forecast horizon to steer the weight more accurately.

A few different variations combining previously introduced ideas are the *rmseRidge* and *rmseElasticNet*. Their formulas are as follows:

$$\textit{Step 2: } \hat{\omega}_{\text{rmseRidge}} = \arg\min_{\omega} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k'} \omega_i f_{i,t} \right)^2 + \lambda_2 \sum_{i=1}^{k'} (\omega_i - \delta_i)^2 \right),$$

$$\begin{aligned} \textit{Step 2: } \hat{\omega}_{\text{rmseElasticNet}} = \arg\min_{\omega} & \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k'} \omega_i f_{i,t} \right)^2 + \lambda_2 \sum_{i=1}^{k'} |\omega_i - \delta_i| \right. \\ & \left. + (1-\lambda_2) \sum_{i=1}^{k'} (\omega_i - \delta_i)^2 \right). \end{aligned}$$

It is insightful to see whether this unique shrinkage via δ_i in combination with Ridge or Elastic Net gives better shrinkage for step 2.

5.3 Evaluating different models

5.3.1 RMSE

Now we have introduced all methods, the next step is evaluating the forecast accuracy of different forecast combiners that are created by different models. We do so by using the RMSE as is done by the original paper, which is:

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (y_t - g(f_{1,t}, f_{2,t}, ..., f_{k,t}))^2}.$$

In the formula, T is the time horizon that is used to predict for time $T+1$, in which y_t is the actual value that is being estimated as time t, and $g(f_{1,t}, f_{2,t}, ..., f_{k,t})$ is the forecast combination made by the forecast combiner at time t.

5.3.2 Diebold-Mariano Test

The Diebold-Mariano (DM) test statistic (Diebold & Mariano, 2002) is used to compare the predictive accuracy of two forecasting models. It is defined as:

$$DM = \frac{\overline{d}}{\sqrt{\hat{f}(0) \cdot \frac{2\pi}{T}}}$$

where:

- $\overline{d} = \frac{1}{T}\sum_{t=1}^{T}(g(e_{1t}) - g(e_{2t}))$ is the mean of the loss differentials, where e_{1t} and e_{2t} are the forecast errors from the two forecasters at time t,
- $g(\cdot)$ is the loss function, typically squared error,
- $\hat{f}(0)$ is an estimate of the spectral density of the loss differentials at frequency zero,
- T is the number of observations.

The hypotheses for the DM test compare the forecast accuracy of the two models:

$$\begin{aligned} H_0 : &\quad E[d_t] = 0, \\ H_a : &\quad E[d_t] \neq 0, \end{aligned}$$

where $d_t = g(e_{1t}) - g(e_{2t})$. The null hypothesis states that there is no difference in forecasting accuracy between the two models.

6 Results

Table 3: Results of different peLasso and peElasticNet adaptations.

Category	RMSE	λ_1 or λ^*	λ_2	NoF	DM	P-value
peLasso (Lasso, AVG)	1.39	0.31	N/A	3	-1.47	0.14
peLasso (Lasso, eRidge)	1.39	0.31	4.75×10^{17}	3	-1.08	0.29
peLasso (Lasso, eLasso)	1.39	0.31	0.03	3	-1.46	0.15
peLasso (Lasso, eEN)	1.39	0.31	0.031	3	-1.33	0.18
peLasso (Lasso, rmseLasso)	1.42	0.31	0.03	3	-1.34	0.18
peLasso (Lasso, rmseRidge)	1.43	0.31	0.03	3	-1.08	0.29
peLasso (Lasso, rmseEN)	1.43	0.31	0.03	3	-1.08	0.29
peElasticNet (EN, AVG)	1.40	0.31	N/A	6	-1.60	0.12
peElasticNet (EN, eRidge)	1.41	0.31	4.75×10^{17}	6	-0.51	0.61
peElasticNet (EN, eLasso)	1.41	0.31	3.20	6	-0.81	0.42
peElasticNet (EN, eEN)	1.41	0.31	3.20	6	-0.62	0.54
peElasticNet (EN, rmseLasso)	1.47	0.31	0.31	6	-0.51	0.61
peElasticNet (EN, rmseRidge)	1.47	0.31	0.31	6	-0.51	0.61
peElasticNet (EN, rmseEN)	1.47	0.31	0.31	6	-0.51	0.61
Individual forecasters						
Best	1.40	N/A	N/A	1	0.61	0.27
90%	1.44	N/A	N/A	1	0.63	0.27
Median	1.53	N/A	N/A	1	-0.57	0.72
10%	1.68	N/A	N/A	1	-1.61	0.94
Worst	1.74	N/A	N/A	1	-1.55	0.94
Simple Average	1.50	N/A	N/A	23	N/A	N/A

Table 3 presents performance results of combinations of forecasting methodologies presented previously in the methodology section. For step 1, two methods are utilised, which are the Lasso and Elastic Net which are used for trimming. For step 2, seven methods are used for trimming, which is Averaging, eLasso, eRidge, eElasticNet, rmseLasso, rmseRidge and rmseElasticNet. Hence, in total, 14 different combinations can be used as methods for predicting. Additionally, the bottom part of Table 3 contains the performance of the Best, 90%, Median, 10%, and Worst individual forecasters in terms of RMSE to compare. The columns starting from the second column of Table 3 display the RMSE and the corresponding λ values of the two-step method, and the three final columns indicate the number of forecasters remaining after the trimming process, the DM test statistics and its corresponding P-value respectively.

In the original study, the three different partially-egalitarian methods employed by the authors are collectively termed peLasso, as all these methods utilise Lasso in the first step. Consistent with this naming convention, we choose to name our methods based on the first step utilised. As illustrated in Table 3, all methods that implement Elastic Net as the initial step are named "peElasticNet". This systematic approach to naming ensures clarity and consistency in the differentiation of methodologies within our study. The first column lists all combinations of partially egalitarian (PE) models for comparison, with the initial seven models being peLasso models and the subsequent entries peElasticNet models. This indicates that the first step (trimming) of our two-step method employs either Lasso or Elastic Net, respectively. Notably, the first three peLasso models are used in the original study, and we refer to these three models in the later part as the benchmark. The remaining models represent new implementations introduced in this study.

6.1 Implementation

For the implementation of all models, we initially employ the Lasso and Elastic Net models using the 'sklearn.linear-model' package in Python 3.12. For subsequent models, including eRidge, eLasso, eElasticNet, and the rmseLasso model, we first reformulate the equations into their fundamental mathematical expressions. Following this, we apply variable transformations. Ultimately, we replicate the procedures established for Lasso and Elastic Net. All the details regarding how we perform the variable transformations are documented in Appendix B.

In the original study, grid search is used to tune the λ parameters; however, in our research, we opt for K-fold (4-fold) cross-validation. The decision to use K-fold cross-validation over grid search is motivated by concerns about the potential for over-fitting associated with grid search, while K-fold cross-validation is preferred for its more robust estimation of model performance. This is because K-fold splits the dataset into multiple folds, evaluates the performance of each fold, and then rotates the folds until it has the performance of every fold combination. This process ensures that every possible fold combination is assessed, allowing for a more comprehensive evaluation. Furthermore, whereas the original authors utilise R packages to conduct their analysis, we opt for Python packages. As a result, the λ values and RMSE's obtained in our study are slightly lower than those reported in the original paper. For the peLasso (Lasso, AVG), (Lasso, eRidge), (Lasso, eLasso) the original paper report an RMSE of 1.40 as we show in Appendix 4, while we achieve an RMSE of 1.39. Additionally, the λ_1 reported in the original study is 0.21, whereas we obtain a value of 0.31 for the first three peLasso models in Table 3.

6.2 Findings

1. All the method's extensions introduced in the paper outperform the simple average way of combining forecasts in terms of forecast accuracy. The RMSE of all method's extensions are lower than the RMSE of the Simple Average of 1.50.

2. The best performing methods are the family of peLasso which shrinks to average, peLasso (Lasso, AVG), peLasso (Lasso, eRidge), peLasso (Lasso, eRidge), and peLasso (Lasso, EN), the mentioned methods all have RMSE of 1.39. The worst-performing methods are the family of peEN which shrinks the weight inversely proportional to the corresponding individual forecasters' RMSE. These methods are namely peElasticNet (EN, rmseLasso), peElasticNet (EN, rmseRidge), and peElasticNet (EN, rmseEN), and they all have a RMSE of 1.47.

3. The peLasso performs better than peElasticNet. In terms of the RMSEs for peElasticNet compared to its counterpart, peLasso. For example, comparing peElasticNet (EN, AVG) with peLasso(lasso, AVG), peElasticNet (EN, eRidge) with peLasso (lasso, Ridge), and vice versa. The obtained peElasticNet RMSEs are uniformly higher than their corresponding peLasso RMSEs. Consequently, replacing the step 1 Lasso with Elastic Net does

not enhance forecast accuracy; this contradicts our initial hypothesis, indicating that the substitution of Elastic Net for Lasso in the trimming process does not improve prediction performance.

4. The forecast accuracy does not improve when eElasticNet is utilised as a replacement for AVG, eLasso, or eRidge in the second step. This observation is evidenced by the RMSEs for peLasso (Lasso, eEN) which is 1.39 and is the same as the RMSE for peLasso (Lasso, AVG), peLasso (Lasso, Ridge), and peLasso (Lasso, eLasso). The same phenomenon also occurs with peElasticNet (EN, eEN) with RMSE of 1.41 which is equivalent to those of the peElasticNet (EN, AVG), peElasticNet (EN, eRidge) and, peElasticNet (EN, eLasso), with corresponding RMSE being 1.40, 1.41 and 1.41, respectively.

   This result shows that given the same trimming method applied in step 1, the step 2 shrinking procedures, namely eRidge, eLasso and eEN which shrink to average, all perform equally.

   This can be attributed to the phenomenon where Averaging forecasters perform very well. There is a favour for high penalty in penalized regression, which produces simple average results. This can be implicated from the result that the RMSE of the models which for step 2 use eLasso, eRidge or eElasticNet have the same RMSE as simple averaging in step 2.

   Furthermore, when we use eRidge for the second step, then λ_2 for peLasso (Lasso, eRidge) and peElasticNet (EN, eRidge) are both extremely high. We interpret this as indicative of a preference for simple averaging in the second step, where a higher λ_2 implies a stronger penalty, leading to an approach closer to a simple average. Given the structure of our dataset, this observation seems adequate since all forecasters have similar performance, which means that the reduction in risk of averaging seems more important than heterogeneously weighing forecasters based on their past performance. This observation aligns with the findings of the original study, which also note high λ_2 values when using eRidge in the second step of the peLasso model. Again, this underlines the previously noted characteristic of the dataset, that risk reduction is more important than heterogeneously weighing forecasts based on past performance.

5. Given that in step 1 the trimming method used is the same, during the step 2 shrinking phase, methods that shrink to average perform better than the methods shrinking to the weight that is used in rmseLasso, rmseRidge and rmseElasticNet methods. This result holds regardless of the choice of the tuning parameter c as can be seen in Table 5, which shows that any strict positive choice of c results in the same forecast accuracy. In the case of peLasso, the RMSE of peLasso (Lasso, rmseLasso), peLasso (Lasso, rmseRidge) and, peLasso (Lasso, rmseEN) are 1.42, 1.43, and 1.43, which is higher than methods which shrink to average peLasso (Lasso, AVG), peLasso (Lasso, eRidge) and, peLasso (Lasso, eLasso) which are 1.39. Same phenomenon happens with peElasticNet where in step 2 shrinking to average performs better than shrinking to the RMSE-based factor.

6. Additionally, the DM tests and their corresponding P-values are reported. The different methods are compared to the average forecast, with the Null Hypothesis being whether

the corresponding forecasting combination methodology and the forecasting combination by simple average have the same prediction accuracy. All the methodologies introduced in the original paper and the adaptations to the methodologies introduced in this paper have forecasting accuracy which is statistically indifferent to the simple average with 5% significance. This is because the P-value of the DM tests are all bigger than 0.05, hence we can not reject the Null hypothesis that the forecast combinations, namely the family of peLasso and peElasticNet significantly differ in forecasting accuracy from the simple average. From the Number of Forecasters (NoF) presented in Table 3, it is not surprising that fewer forecasters remain after using Lasso compared to Elastic Net in the first step of the two-step method. This observation is referable to the inherent characteristics of Lasso, which employs L1 regularization. L1 regularization is known for zeroing out the coefficients of less significant features, effectively reducing the number of variables. On the other hand, Elastic Net incorporates both L1 and L2 (Ridge) penalties. This combination not only retains the variable selection properties of Lasso but also allows for simple shrinkage of variables. In essence, the integration of L1 and L2 penalty terms mitigates the propensity of aggressively setting coefficients to zero, as seen with Lasso alone. Consequently, Elastic Net tends to retain more forecasters than Lasso, shown in Table 3.

7 Conclusions

In this paper, the challenge of improving prediction accuracy was addressed by constructing various models based on the two-step partially-egalitarian Lasso (peLasso) framework described by Diebold & Shin (2019). Originally, the peLasso model employed the Lasso method in its first step, and either the average, eLasso or eRidge methods as the second step. The objective was to evaluate whether modifying the first step to incorporate an Elastic Net model or modifying the second step to implement eElasticNet, rmseLasso, rmseRidge or rmseElasticNet, could enhance prediction performance (measured by RMSE). Additionally, the potential benefits of changing both steps simultaneously was explored.

Our findings indicate that extensions to the peLasso all perform better than the simple average. Yet, the extensions we introduce in this extension perform worse than the peLasso from the original paper. The best-performing models are the peLasso from the original paper and peLasso (Lasso, eEN) which perform equally well. Additionally, the peElasticNet models underperform compared to the peLasso models. The RMSE-based shrinking methods introduced in section 5.2.1, such as peLasso(Lasso, rmseLasso), underperform compared to shrinking models that shrink towards the average, as proposed by the original paper.

Consequently, we did not identify a superior method that improves upon the original forecast accuracy documented in the original paper. The previous leads us to answer the research question

Can adaptations to peLasso lead to improved forecast combinations?

with a no for the adaptations that we suggested.

For future research, we propose several avenues of investigation. Firstly, it would be instructive to apply the methodologies from the original study to a distinctly different dataset. Given that the current dataset features forecasters with relatively homogeneous statistics (mean, standard deviation, minimum, and maximum), it would be enlightening to assess whether these methods maintain their efficiency in datasets characterised by greater variability among forecasters. We hypothesise that in particular in those datasets, the trimming phase increases in importance, resulting in an optimal trimming that differs from the simple average. Additionally, we recommend employing the adaptive Lasso (aLasso) or adaptive Elastic Net (aEN) for the initial trimming phase, followed by shrinking using egalitarian adaptive Lasso or egalitarian adaptive Elastic Net. Adaptive Elastic Net could possess advantage such as the *oracle property*, which is obtained when using aLasso. The oracle property means that the model performs as well asymptotically as if the true underlying model were given in advance (Zou, 2006). This approach is promising, considering the favourable outcomes achieved with aLasso in the original study. Moreover, exploring alternative trimming techniques could yield significant insights, as our results suggest that the first step of the two-step procedure largely determines the outcomes within the data set context. Particularly, techniques like Random Forest may prove particularly beneficial for identifying crucial regressors due to their ability to capture nonlinear effects. Finally, alternative cross-validation methods for tuning the λ should be considered, as the k-fold cross-validation employed in this study may be computationally expensive with larger datasets.

References

Abramowitz, A. I. (1988). An improved model for predicting presidential election outcomes. *PS: Political Science & Politics*, *21*(4), 843–847.

Aiolfi, M., Capistrán, C., & Timmermann, A. (2010). Forecast combinations. *CREATES research paper*(2010-21).

Aruoba, S. B., Diebold, F. X., Nalewaik, J., Schorfheide, F., & Song, D. (2012). Improving us gdp measurement: A forecast combination perspective. In *Recent advances and future directions in causality, prediction, and specification analysis: Essays in honor of halbert l. white jr* (pp. 1–25). Springer.

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, *36*(1), 197–200.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559-583. Retrieved from `https://www.sciencedirect.com/science/article/pii/0169207089900125` doi: https://doi.org/10.1016/0169-2070(89)90012-5

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, *35*(4), 1679–1691.

Elliott, G., & Timmermann, A. (2016). Forecasting in economics and finance. *Annual Review of Economics*, *8*, 81–110.

Franses, P. H. (2024). *Ethics in econometrics: A guide to research practice.* Cambridge University Press.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International journal of forecasting*, *21*(1), 15–24.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Kolassa, S. (2011). Combining exponential smoothing forecasts using akaike weights. *International Journal of Forecasting*, *27*(2), 238-251. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0169207010001032` doi: https://doi.org/10.1016/j.ijforecast.2010.04.006

Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, *209*, 226-235. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0925527318302196` (The Proceedings of the 19th International Symposium on Inventories) doi: https://doi.org/10.1016/j.ijpe.2018.05.019

Leng, C., Tran, M.-N., & Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, *66*, 221–244.

Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of monetary economics*, *44*(2), 293–335.

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

Timmermann, A. (2006). Chapter 4 forecast combinations. In G. Elliott, C. Granger, & A. Timmermann (Eds.), (Vol. 1, p. 135-196). Elsevier. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1574070605010049` doi: https://doi.org/10.1016/S1574-0706(05)01004-9

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, *17*(2), 228.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005, 03). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320. Retrieved from `https://doi.org/10.1111/j.1467-9868.2005.00503.x` doi: 10.1111/j.1467-9868.2005.00503.x

A peLasso results from the original paper

Table 4: Forecast RMSEs based on ex post optimal λs from the original paper

Category	**RMSE**	$\boldsymbol{\lambda}$	**DM**	**P-value**
peLasso (Lasso, Average)	1.40	0.21	1.06	0.15
peLasso (Lasso, eRidge)	1.40	(0.21, max)	1.06	0.15
peLasso (Lasso, eLasso)	1.40	(0.21, 3.1)	1.07	0.15

Table 5: The table shows the RMSE for different choices of the tuning parameter c for the RMSE-based shrinkage methods.

	RMSE				
Category	**c = 0.001**	**c = 0.01**	**c = 1**	**c = 10**	**c = 100**
peLASSO (Lasso, rmseLasso)	1.42	1.42	1.42	1.42	1.42
peLASSO (Lasso, rmseLasso)	1.43	1.43	1.43	1.43	1.43
peLASSO (Lasso, rmseLasso)	1.43	1.43	1.43	1.43	1.43
peELASTICNET (EN, rmseLasso)	1.47	1.47	1.47	1.47	1.47
peELASTICNET (EN, rmseLasso)	1.47	1.47	1.47	1.47	1.47
peELASTICNET (EN, rmseLasso)	1.47	1.47	1.47	1.47	1.47

B Derivations

B.1 Rewriting eElasticNet as ElasticNet

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}\beta_i f_{it}\right)^2+\lambda_1\sum_{i=1}^{K}\left|\beta_i-\frac{1}{K}\right|+\lambda_2\sum_{i=1}^{K}\left(\beta_i-\frac{1}{K}\right)^2\\
&=\sum_{t=1}^{T}\left(y_t-\overline{f_t}+\overline{f_t}-\sum_{i=1}^{K}\beta_i f_{it}\right)^2+\lambda_1\sum_{i=1}^{K}\left|\beta_i-\frac{1}{K}\right|+\lambda_2\sum_{i=1}^{K}\left(\beta_i-\frac{1}{K}\right)^2\\
&=\sum_{t=1}^{T}\left(y_t-\overline{f_t}+\sum_{i=1}^{K}\left(\frac{1}{K}-\beta_i\right)f_{it}\right)^2+\lambda_1\sum_{i=1}^{K}\left|\beta_i-\frac{1}{K}\right|+\lambda_2\sum_{i=1}^{K}\left(\beta_i-\frac{1}{K}\right)^2\\
&=\sum_{t=1}^{T}\left((y_t-\overline{f_t})-\sum_{i=1}^{K}\left(\beta_i-\frac{1}{K}\right)f_{it}\right)^2+\lambda_1\sum_{i=1}^{K}\left|\beta_i-\frac{1}{K}\right|+\lambda_2\sum_{i=1}^{K}\left(\beta_i-\frac{1}{K}\right)^2\\
&=\sum_{t=1}^{T}\left(y^*-\sum_{i=1}^{K}\delta_i f_{it}\right)^2+\lambda_1\sum_{i=1}^{K}|\delta_i|+\lambda_2\sum_{i=1}^{K}\delta_i^2,
\end{aligned}
$$

where

$$y^* = y_t - \overline{f_t} \quad \text{and} \quad \delta_i = \beta_i - \frac{1}{K}.$$

B.2 Rewriting rmseLasso as Lasso

$$\begin{aligned}
&\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}\beta_i f_{it}\right)^2+\lambda\sum_{i=1}^{K}|\beta_i-w_i| \\
&=\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}w_i f_{it}+\sum_{i=1}^{K}w_i f_{it}-\sum_{i=1}^{K}\beta_i f_{it}\right)^2+\lambda\sum_{i=1}^{K}|\beta_i-w_i| \\
&=\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}w_i f_{it}+\sum_{i=1}^{K}(w_i f_{it}-\beta_i f_{it})\right)^2+\lambda\sum_{i=1}^{K}|\beta_i-w_i| \\
&=\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}w_i f_{it}+\sum_{i=1}^{K}(w_i-\beta_i) f_{it}\right)^2+\lambda\sum_{i=1}^{K}|\beta_i-w_i| \\
&=\sum_{t=1}^{T}\left(\left(y_t-\sum_{i=1}^{K}w_i f_{it}\right)-\sum_{i=1}^{K}(\beta_i-w_i) f_{it}\right)^2+\lambda\sum_{i=1}^{K}|\beta_i-w_i| \\
&=\sum_{t=1}^{T}\left(y_t^*-\sum_{i=1}^{K}\delta_i f_{it}\right)^2+\lambda\sum_{i=1}^{K}|\delta_i|,
\end{aligned}$$

where

$$y_t^*=y_t-\sum_{i=1}^{K}w_i f_{it}\quad\text{and}\quad\delta_i=\beta_i-w_i.$$

B.3 Rewriting aLasso as Lasso

$$\begin{aligned}
&\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}\beta_i f_{it}\right)^2+\lambda\sum_{i=1}^{K}\frac{1}{\left|\hat\beta_i\right|}|\beta_i| \\
&=\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}\beta_i f_{it}\right)^2+\lambda\sum_{i=1}^{K}\left|\frac{\beta_i}{\hat\beta_i}\right| \\
&=\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}\beta_i f_{it}(\frac{\hat\beta_i}{\hat\beta_i})\right)^2+\lambda\sum_{i=1}^{K}\left|\frac{\beta_i}{\hat\beta_i}\right| \\
&=\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}(\frac{\beta_i}{\hat\beta_i})(f_{it}\hat\beta_i)\right)^2+\lambda\sum_{i=1}^{K}\left|\frac{\beta_i}{\hat\beta_i}\right| \\
&=\sum_{t=1}^{T}\left(y_t-\sum_{i=1}^{K}f_{it}^*\delta_i\right)^2+\lambda\sum_{i=1}^{K}|\delta_i|,
\end{aligned}$$

where

$$f_{it}^*=f_{it}\hat\beta_i,\quad\text{and}\quad\delta_i=\frac{\beta_i}{\hat\beta_i}.$$