International Journal of Forecasting 35 (2019) 1679–1691

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives

Francis X. Diebold [a], Minchul Shin [b,*]

[a] University of Pennsylvania, United States
[b] University of Illinois, United States

ARTICLE INFO

Keywords:
Forecast combination
Forecast surveys
Shrinkage
Model selection
LASSO
Regularization

ABSTRACT

Despite the clear success of forecast combination in many economic environments, several important issues remain incompletely resolved. The issues relate to the selection of the set of forecasts to combine, and whether some form of additional regularization (e.g., shrinkage) is desirable. Against this background, and also considering the frequently-found good performance of simple-average combinations, we propose a LASSO-based procedure that sets some combining weights to zero and shrinks the survivors toward equality ("partially-egalitarian LASSO"). Ex post analysis reveals that the optimal solution has a very simple form: the vast majority of forecasters should be discarded, and the remainder should be averaged. We therefore propose and explore direct subset-averaging procedures that are motivated by the structure of partially-egalitarian LASSO and the lessons learned, which, unlike LASSO, do not require the choice of a tuning parameter. Intriguingly, in an application to the European Central Bank Survey of Professional Forecasters, our procedures outperform simple average and median forecasts; indeed, they perform approximately as well as the ex post best forecaster.

1. Introduction

Forecast combination has a long and successful history in economics.[1] However, various important issues still have not been resolved completely, related to determining the best set of forecasts to combine ("selection", e.g., via an information criterion), how to combine those selected (e.g., via a linear weighted average), and whether some form of regularization (e.g., via shrinkage) is desirable, given that the historical forecast record is often small relative to the number of candidate forecasters. Against this background, and also considering the frequently-found good performance of simple-average combinations, we propose various LASSO-inspired procedures that address all considerations.

We proceed as follows. Section 2 highlights aspects of the "equal-weights puzzle", that is, the frequently-found good performance of simple-average combinations which motivates our concerns and proposals, and also describes our "partially-egalitarian LASSO" procedures, which shrink and select in desirable ways. Section 3 provides an ex post empirical assessment of our procedure's performance. Section 4 proposes and explores direct ex ante combination procedures motivated by the structure of partially-egalitarian LASSO and the lessons learned. Section 5 places our methods in the context of the broader literature, which

* Corresponding author.
E-mail addresses: fdiebold@upenn.edu (F.X. Diebold), mincshin@illinois.edu (M. Shin).

[1] For overviews, see Diebold and Lopez (1996), Timmermann (2006), and Elliott and Timmermann (2016).

https://doi.org/10.1016/j.ijforecast.2018.09.006
0169-2070/

notably includes the studies by Capistrán and Timmermann (2009), Elliott (2011), Conflitti, De Mol, and Giannone (2015), and Samuels and Sekkel (2017), among many others. Finally, Section 6 concludes.

2. Partially-egalitarian LASSO for forecast combination

This section considers methods for selection and shrinkage in regression-based forecast combination. The key new method is "partially-egalitarian LASSO" (peLASSO), but we build up to it gradually, arriving at it in Section 2.6.

2.1. Aspects of optimal forecast combination

Although it seems natural to average forecasts (i.e., to use equal-weight combinations), simple averages are generally suboptimal. To see the theoretical sub-optimality of equal combining weights, consider K competing unbiased forecasts $f_t^1, \ldots, f_t^K$ of y_t. We form a combined forecast as

$$C_t = \beta_1 f_t^1 + \beta_2 f_t^2 + \cdots + \left(1 - \sum_{k=1}^{K-1} \beta_i\right) f_t^K.$$

The corresponding forecast errors, e_{Ct} and $e_{1t}, \ldots, e_{Kt}$, have variances σ_C^2 and $\sigma_1^2, \ldots, \sigma_K^2$, and satisfy the same equality, from which it follows that the variance of the combined forecast error is minimized using the weight vector

$$\beta^* = \left(\Sigma^{-1}\mathbf{i}\right) / \left(\mathbf{i}'\Sigma^{-1}\mathbf{i}\right), \tag{1}$$

where Σ is the variance–covariance matrix of the forecast errors and $\mathbf{i}$ is a conformable column vector of ones (Bates & Granger, 1969). In particular, equal weights — that is, simple averages — are generally suboptimal.[2]

It is well known (Granger & Ramanathan, 1984) that the population Bates-Granger optimal combining weights in Eq. (1) may be obtained trivially from the population regression (linear projection) $y_t \rightarrow f_t^1, \ldots, f_t^K$, subject to the constraint that the coefficients add to one.[3] Thus, the theoretical optimal linear forecast combination problem is just a population linear regression (projection) problem, and the estimation of finite-sample combining weights involves just a simple linear regression.

Despite the theoretical sub-optimality of equal weights, a large body of literature has found frequent good performances of simple averages under quadratic loss. Indeed, the forecast combination "equal weights puzzle", emphasized long ago by Clemen (1989) and Diebold (1989), refers to the frequently-found good performance of simple averages.[4] By now, though, the equal weights puzzle has been studied thoroughly and is understood better. For example, Aruoba et al. (2012) work in population (i.e., without estimation error) and show that: (1) even if simple averages are not fully optimal, they are likely to be much better than any individual forecast, and (2) even if simple averages are not fully optimal, they are likely to be close to the optimum. In addition, Smith and Wallis (2009) show that finite-sample combining-weight estimation error can degrade empirical attempts at optimal combination seriously, which further increases the relative attractiveness of simple averages, since they do not involve estimation.

Thus far, the discussion strongly suggests that simple averages (equal weights) are a natural shrinkage direction for such combining regressions. With shrinkage, we do not *force* simple averages; rather, we coax things in that direction, blending data (likelihood) information with prior information. This amounts to a Bayesian approach with the prior centered on simple averages.

An important issue remains, however. Particularly when combining large numbers of forecasts, some forecasts may be largely redundant, or not worth including in the combination for a variety of other reasons. Thus, we may potentially want to set *some* combining weights to zero ("select to zero") and shrink the *remaining* weights toward equality ("shrink toward equality"). As we will see, LASSO-based methods almost do the trick, as they both select and shrink; unfortunately, though, they select to zero *and* shrink to zero. In the remainder of this section we begin by discussing the standard LASSO, which we then modify until we arrive at our "partially-egalitarian LASSO", which selects to zero and shrinks to equality. Interestingly, each of the estimators introduced en route proves useful in its implementation.[5]

2.2. Penalized estimation for selection and shrinkage

Consider a penalized forecast combining regression, with "parameter budget" c,

$$\hat{\beta}_{Penalized} = \arg\min_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it}\right)^2$$
$$s.t. \sum_{i=1}^{K} |\beta_i|^q \leq c. \tag{2}$$

Equivalently, in Lagrange-multiplier form we can write

$$\hat{\beta}_{Penalized} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it}\right)^2 + \lambda \sum_{i=1}^{K} |\beta_i|^q\right),$$

[2] As an example, consider two forecasts with uncorrelated errors. Then Eq. (1) reduces to

$$\beta^* = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1}{1+\phi^2},$$

where β^* is the weight placed on forecast 1 and $\phi = \sigma_1/\sigma_2$. Hence, the simple average obtains if and only if $\phi = 1$. This is entirely natural: we want to give more weight to the forecast with lower-variance errors, so we take a simple average only in the equal-variance case.

[3] Moreover, one can allow for biased forecasts by including an intercept, and there is no real need to impose the "sum-to-one" constraint.

[4] Note that the theoretical suboptimality of simple averages, and hence the equal weights puzzle, refers to combination under quadratic loss. Under other loss functions, equal weights may in fact be optimal. For example, Aruoba, Diebold, Nalewaik, Schorfheide, and Song (2012) show that equal weights are optimal under minimax loss.

[5] For a broad introduction to LASSO and related procedures, see Hastie, Tibshirani, and Friedman (2009).

where λ depends on c. Taking $\lambda = 0$ produces Bates-Granger OLS combining:

$$\hat{\beta}_{BG} = \arg\min_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 .$$

Many estimators that select and/or shrink, both of which are important for our purposes, fit in the penalized estimation framework.[6]

2.3. Shrinkage toward equality: egalitarian ridge

Smooth convex penalties in Eq. (2) produce pure shrinkage. In particular, $q = 2$ produces ridge regression, which shrinks the coefficients toward zero:

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \beta_i{}^2 \right).$$

Taking $q = 2$ and centering the constraint around $1/K$ produces a modified ridge regression that shrinks the coefficients toward equality ("egalitarian ridge", or "eRidge"):

$$\hat{\beta}_{eRidge} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left(\beta_i - \frac{1}{K} \right)^2 \right).$$

eRidge is related closely to the Bayesian shrinkage combining weight estimation of Diebold and Pauly (1990), who take an empirical Bayes approach using the g-prior of Zellner (1986), but it is simpler to implement.

Note that, although eRidge will feature later in this paper (which is why we introduce it), it is inadequate for our ultimate purpose: it shrinks in the right direction but does not select.

2.4. Selection to and shrinkage toward zero: LASSO

As we have noted, $q = 2$ produces pure shrinkage (ridge). Conversely, $q \to 0$ produces pure selection. The intermediate case $q = 1$ produces shrinkage *and* selection, and is known as a LASSO estimator:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i| \right).$$

The seminal reference is the paper by Tibshirani (1996).

There are several variants of LASSO. The most important for our purposes is "adaptive LASSO" (Zou, 2006), which weights the terms in the penalty to encourage small first-round coefficient estimates to be set to zero,

$$\hat{\beta}_{aLASSO} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i |\beta_i| \right),$$

where $w_i = 1/|\hat{\beta}_i|^{\nu}$, $\hat{\beta}_i$ is OLS or ridge, and $\nu > 0$. Others include the "elastic net" (Zou & Hastie, 2005), which uses a convex combination of the LASSO ($q = 1$) and ridge penalties ($q = 2$), namely $\sum_{i=1}^{K} (\alpha|\beta_i| + (1-\alpha)\beta_i^2)$, and "adaptive elastic net", which blends the adaptive LASSO and elastic net penalties as $\sum_{i=1}^{K} (\alpha w_i|\beta_i| + (1-\alpha)\beta_i^2)$.

Under some assumptions, the adaptive versions (adaptive LASSO and adaptive elastic net) have the so-called "oracle property".[7] The elastic net variants have good properties in handling highly-correlated predictors. The adaptive elastic net has both. Unfortunately, though, all LASSO variants, while improving on ridge insofar as they both shrink *and* select, remain inadequate for our purposes: they select in the right direction (to zero) but shrink in the wrong direction (toward zero).

2.5. Selection to and shrinkage toward equality: egalitarian LASSO

All of the standard LASSO variants in Section 2.4 select and shrink combining weights toward zero, but that is not what we want. Instead, as was discussed in Section 2.1, both theory and experience point clearly to shrinkage toward simple averages. We therefore change the LASSO penalized estimation problem to

$$\hat{\beta}_{eLASSO} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left| \beta_i - \frac{1}{K} \right| \right).$$

That is, instead of shrinking the weights toward zero, we shrink the *deviations from equal weights* toward zero. Appendix A shows that eLASSO implementation is straightforward using standard software.

Note that although eLASSO shrinks in the right direction, it is still unappealing, for reasons opposite those of the standard LASSO. Like the standard LASSO, eLASSO shrinks and selects, but whereas LASSO *shrinks* in the wrong direction, eLASSO *selects* in the wrong direction! However, the reason why we introduced Ridge, eRidge, LASSO, and eLASSO was because the procedure to which we now turn, which both shrinks *and* selects in the right directions, is closely related, and because each will feature importantly in our subsequent empirical work.

2.6. Selection to zero and shrinkage toward equality: partially-egalitarian LASSO

eLASSO does not tend to discard forecasters, because it selects and shrinks toward equal weights, not zero weights. In particular, eLASSO implicitly presumes that all forecasters "belong" in the set to be combined. However, one can easily modify the eLASSO such that some forecasters are potentially discarded, and then the survivors are selected

[6] One could also add additional constraints. For example, with unbiased forecasts it may be natural to impose $\beta_i \geq 0$, $\forall i$ and $\sum_{i=1}^{K} \beta_i = 1$, as per Conflitti et al. (2015), but we will not pursue that here.

[7] That is, roughly speaking, they asymptotically select the data-generating process (DGP) almost surely if it is among the models considered, and otherwise select the best predictive approximation to the DGP.

and shrunken toward equality. We call this the "partially-egalitarian LASSO".

2.6.1. One-step conceptualization

The partially-egalitarian LASSO (peLASSO) solves a penalized estimation problem with two penalties,

$$\hat{\beta}_{peLASSO} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda_1 \sum_{i=1}^{K} \left| \beta_i \right| + \lambda_2 \sum_{i=1}^{K} \left| \beta_i - \frac{1}{p(\beta)} \right| \right), \quad (3)$$

where $p(\beta)$ is the number of non-zero elements in β. The first is the standard LASSO penalty, which selects and shrinks to zero, whereas the second selects and shrinks to equality. The optimization of this one-step objective proves difficult, due to the discontinuity of the objective function at $\beta_i = 0$. We therefore reserve it for future work and proceed instead with a two-step approach.

2.6.2. Two-step implementation

The obvious two-step analog of Eq. (3) above is:

Step 1 (Select to zero): Using standard methods, select k forecasts from among the full set of K forecasts.

Step 2 (Shrink towards equality): Using standard methods, shrink the combining weights on the k forecasts that survive step 1 toward $1/k$.

The obvious method for step 1 is the standard LASSO, which requires only one estimation and moreover can handle situations with $K > T$, which are not uncommon in forecast combination. In our subsequent empirical work, for example, such situations are omnipresent, as our combining regressions involve more forecasters than observations.

One obvious method for step 2 is eRidge, which is trivial to implement via a standard ridge regression with a transformed left-hand-side variable, as is discussed in Appendix A. One could go even farther and use eLASSO for step 2, in which case the complete procedure would first select some weights to zero, then select some of the surviving weights to $1/k$ and shrink the rest toward $1/k$.

The empirical work in Sections 3 and 4 emphasizes combining procedures that are motivated by the two-step peLASSO.

3. Ex post optimal peLASSO tuning

This section begins our empirical work, providing a comparative assessment of various forecast combination methods using the European Central Bank's well-known quarterly Survey of Professional Forecasters.[8] Of course, the comparative performances of our methods, using a particular dataset and a particular implementation (choice of sample period, choice of tuning parameters, etc.), cannot establish anything conclusively, but the comparison illustrates our methods in a realistic and important environment, and provides suggestive evidence regarding the methods' performances.

We emphasize that this section examines out-of-sample RMSEs for those procedures that require the selection of a tuning parameter λ, based on the ex post optimal λ, i.e., the λ that optimizes the out-of-sample RMSEs that we would have obtained if we had been using it in real time, which we can determine ex post.[9] Hence, we endow the forecaster with valuable information that is not available ex ante. Section 4 subsequently shows how to address the tuning issue ex ante, the key to which is first to understand the nature of the ex post solution, to which we now turn.

3.1. Background

Again, we focus on the European Central Bank's well-known quarterly Survey of Professional Forecasters. We consider quarterly 1-year-ahead forecasts of Euro-area real GDP growth (year-on-year percentage change). However, as was noted by Genre, Kenny, Meyler, and Timmermann (2013), forecasts are solicited for one year ahead of the *latest available outcome*. For example, the 2007Q1 survey asked the respondents to forecast the GDP growth over 2006Q3–2007Q3. Hence, our "one-year-ahead" growth forecasts are actually only six to eight months ahead.

We have an unbalanced panel, because forecasters enter and exit in real time, in addition to which those in the panel at any time do not necessarily respond to the survey. Hence, for ease of analysis, we select the 23 forecasters who responded most frequently to the surveys (1999Q1–2016Q2), and impute missing observations using a linear filter as per Genre et al. (2013). We start with the 1999Q1 survey because the survey began then, and we end with the 2016Q2 survey to ensure that all of our growth realizations data are of the final revised form, as we now explain.

Throughout, we calculate forecast errors using "realizations" from the 2018Q1 data vintage (pulled in 2018M5, when the latest revision of this paper was begun, and containing what we will consider to be final revised data for all quarters through 2016Q4). The first release of 2016Q4 GDP was in 2017M2, then it went through several revisions. The statistical agency, Eurostat, makes all "standard" revisions by 100 days after the end of the quarter ("preliminary" 30 days after, "flash" 45 days after, "regular" 60 days after, and "updated" 100 days after), but additional non-standard revisions sometimes occur after more than 100 days, so we wait approximately a year, using "realizations" from the 2018Q1 vintage, to ensure that all realizations are approximately their "final-revised" values, which is desirable because forecasters should be forecasting true GDP growth, the best estimate of which is the final-revised value, not a preliminary release.

We perform the forecast evaluation as follows. Our surveys run over the period 1999Q1–2016Q2, corresponding to growth rate forecasts 1999Q3–2016Q4. We burn in our

[8] See http://www.ecb.europa.eu/stats/prices/indic/forecast/html/index.en.html.

[9] Of course, λ could be a vector, as with peLASSO.

estimation using the first five forecasts 1999Q3–2000Q3, so our actual evaluation period is 2000Q4–2016Q4. We roll through the evaluation sample, estimating combining weights using a 5-year (20-quarter) window and producing 1-year-ahead out-of-sample forecasts. For periods 6–20, we simply estimate the forecasts using all available data from time 1, despite that fact that there are fewer than 20 observations.[10] For periods $t > 20$, we use a full 20-period estimation window.

We focus on combining methods that involve regularization estimators, which is essential in our context as $K > T$. Our main comparison involves combined forecasts based on Ridge, LASSO, eRidge, eLASSO, and three versions of peLASSO (the first step is always LASSO, and the second step is simple average, eRidge, or eLASSO).[11] Throughout, we compare the formally-combined forecasts to simple averages.

Each combining method except simple averages requires the selection of a tuning parameter, λ, which governs the regularization strength. We examine the combined forecast accuracy for many λs, ranging from a very light penalization (small λ; all forecasters included in the combination) to a very heavy penalization (large λ; no forecasters included in the combination). Specifically, we compute forecasts on a grid of 200 λs. We start with an equally-spaced grid on $[-15, 15]$, which we then exponentiate, producing a grid on $(0, 3269017]$, with the grid's coarseness increasing with λ. This grid turns out to be adequate for all LASSO-based combinations that we consider.

3.2. Ex post results

Table 1 presents out-of-sample combined forecast RMSEs. There are many relevant observations that could be made. In no particular order:

1. Granger-Ramanathan OLS combination is infeasible because $K > T$, so we cannot include it in the table.
2. No method performs better than the best individual forecaster. (It can happen that a combined forecast is better than any individual forecast, but it does not happen here.)
3. All methods perform better than the worst individual forecaster.
4. The simple average improves significantly over the worst individual, but is still noticeably worse than the best individual.
5. All procedures that involve selection to zero select a very small number of forecasters on average (approximately three).
6. Ridge and LASSO perform about as well as the simple average, despite the fact that they shrink toward zero weights rather than equal weights.
7. eRidge and eLASSO perform exactly as well as the simple average. This is because the optimal regularization (toward the average) turns out to be very strong, in which case both eRidge and eLASSO produce a simple average.
8. All peLASSO methods perform identically. The reason for this is as follows. They regularize identically in the first step, by construction (all use the standard LASSO in step 1). Then, in the second step, the "LASSO, Average" method averages by construction, and the remaining methods effectively average as well, because heavy step-2 regularization turns out to be optimal.
9. The peLASSO methods reduce the out-of-sample RMSE relative to the simple average by almost ten percent.
10. The peLASSO methods have out-of-sample RMSEs that are as good as that of the best forecaster. This property is reminiscent of procedures that achieve external regret minimization in the "combining expert advice" problem, as was discussed by Arora, Hazon, and Kale (2012), for example.

Table 1
Forecast RMSEs based on ex post optimal λs.

Regularization group	RMSE	λ^*	#	DM	p-val
Ridge	1.51	2.66	23.00	−0.14	0.56
LASSO	1.52	0.38	2.71	−0.10	0.54
eRidge	1.50	max	23.00	−1.14	0.87
eLASSO	1.50	3.60	23.00	0.95	0.17
peLASSO (LASSO, Average)	1.40	0.21	2.95	1.06	0.15
peLASSO (LASSO, eRidge)	1.40	(0.21, max)	2.95	1.06	0.15
peLASSO (LASSO, eLASSO)	1.40	(0.21, 3.10)	2.95	1.07	0.15
Comparisons	**RMSE**	**λ^***	**#**	**DM**	**p-val**
Best	1.40	N/A	1	0.61	0.27
90%	1.44	N/A	1	0.63	0.27
Median	1.53	N/A	1	−0.57	0.72
10%	1.68	N/A	1	−1.61	0.94
Worst	1.74	N/A	1	−1.55	0.94
Average	1.50	N/A	23	N/A	N/A

Notes: λ^* is the ex post optimal penalty parameter(s), # is the average number of forecasters selected, and DM is the one-sided (Diebold & Mariano, 1995) statistic against a simple average, with the p-value denoted by p-val. We compute DM as per Harvey, Leybourne, and Newbold (1999).

The nature of the ex post optimal solution is contained in results 5 and 8: *first discard most forecasters (result 5), then simply average the survivors (result 8)*. The importance of this "trim and average" solution cannot be overemphasized, and we will indeed emphasize and explore it extensively in Section 4.

Appendix B shows that the results are robust to doing the evaluation over only periods $t > 20$, so that we always have an exact 20-period estimation window. Appendix C then shows that the results are also robust (and in fact even better) when using aLASSO rather than LASSO in the two-step peLASSO. The trim-and-average nature of the ex post optimal peLASSO solution remains intact throughout: first discard most forecasters, then average the survivors.

[10] We do this so as to avoid the need to discard the first 20 observations, as degrees of freedom are scarce.

[11] Unlike much of the LASSO literature, we do not standardize our data. Standardization is desirable when the regressors are measured in different units, but such is not the case with forecast combination, so there is no need.

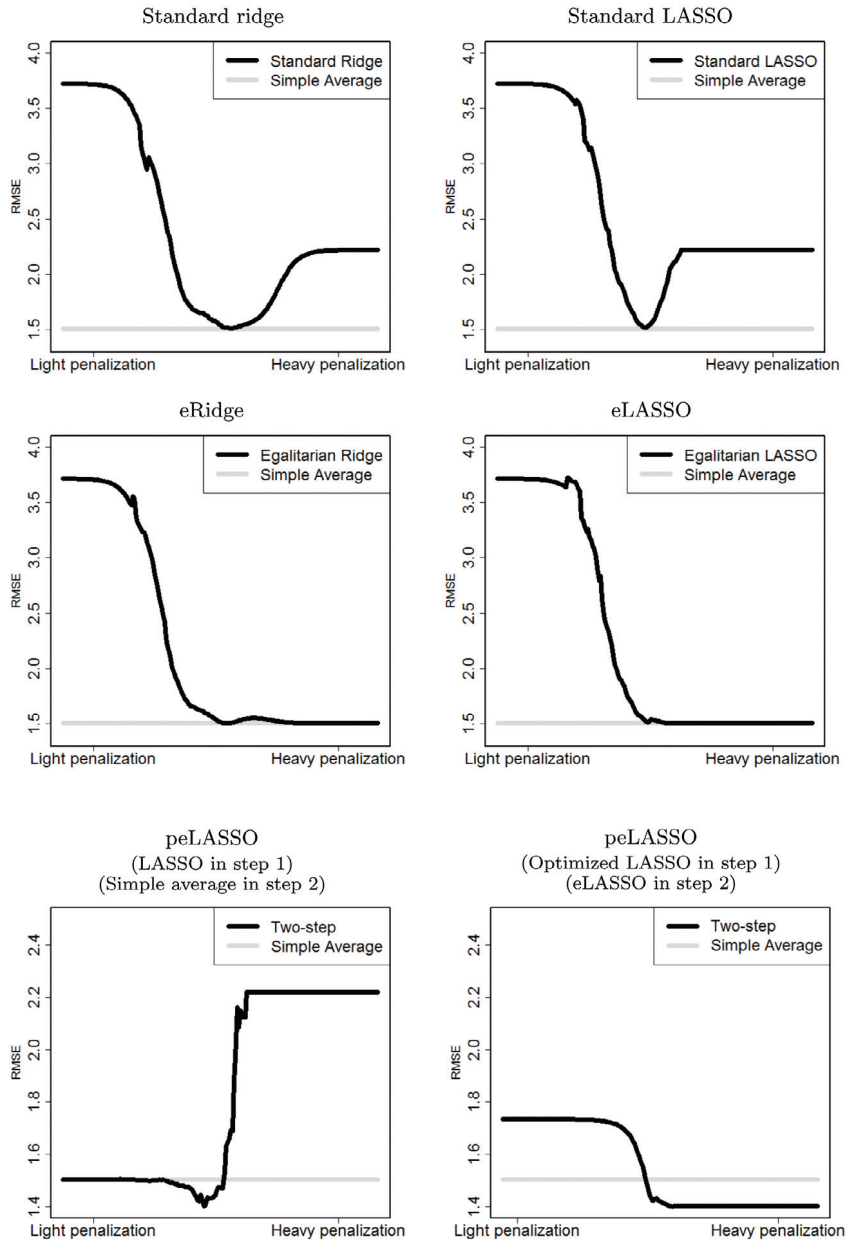

Fig. 1. RMSE as a function of λ for various forecast combination methods. Notes: The lower-right panel implements step 2 using egalitarian LASSO regression on the step-1 selected forecasters, so there is an additional penalty parameter. We show RMSE as a function of the step-2 penalty, with the step-1 penalty being fixed at its optimal value.

3.3. *On the importance of* λ

The results in Table 1 depend on a knowledge of the ex post optimal λ. We get a feel for the sensitivity to λ by showing RMSE as function of λ in Fig. 1. The lighter gray line in each graph is the RMSE for simple averaging. Consider first the top row of Fig. 1, in which we show the standard ridge and standard LASSO. They perform similarly in terms of the optimized value based on the ex post best λ; at that point they are basically indistinguishable both from each other and from a simple average. However, in the limit as the penalization increases, their performances deteriorate as all forecasters are eventually excluded and the "combined forecasts" therefore approach zero. Finally, note that the simple average is never beaten, including at the ex post optimum λs.

Next, consider the second row of Fig. 1, in which we show eRidge and eLASSO. They too perform similarly in terms of the optimized value based on the ex post best λ; at that point they are basically indistinguishable both

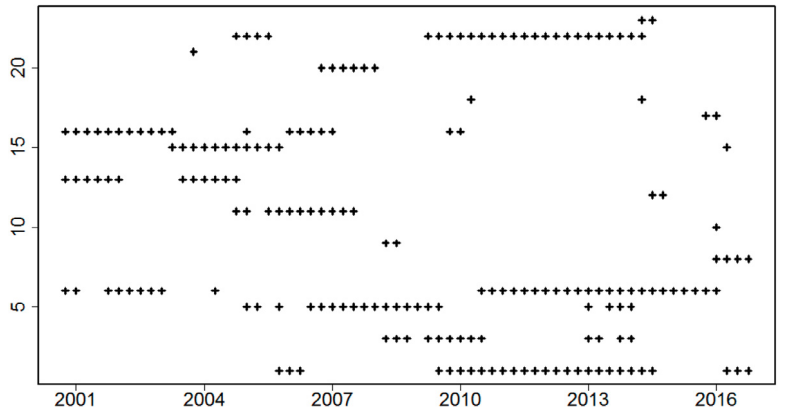

Fig. 2. Selected forecasters. Notes: The x-axis denotes time and the y-axis denotes the forecaster ranking, where a lower y-axis location refers to a forecaster with a smaller overall RMSE. A "+" symbol at location (x, y) indicates that forecaster y was selected at time x.

from each other and from a simple average. However, their penalization limits are very different. In the limit as penalization increases, eRidge, eLASSO, and simple averaging must be (and are) identical. As in the first row of Fig. 1, however, the simple average is never beaten.

Now consider the third row of Fig. 1, in which we show peLASSO, in each case with step 1 performed using the standard LASSO. The left panel implements step 2 by simply averaging the step-1 selected forecasters, so that there is only one penalty parameter to choose. At the ex post optimum penalty, this two-step egalitarian LASSO outperforms other methods, including simple averaging of all forecasters.

The right panel of the third row of Fig. 1 implements step 2 by eLASSO regression on the forecasters selected in step 1, so there is a second penalty parameter to be chosen. Denote the ex post optimal pair by $(\lambda_1^*, \lambda_2^*)$. We show RMSE as a function of λ_2, with λ_1 being fixed at λ_1^*. It turns out that once we select forecasters, it is ex post optimal to shrink those selected strongly toward a simple average; that is, heavy step-2 penalization (large λ_2) is optimal.

The key result is that, unlike other methods (rows 1 and 2 of Fig. 1), peLASSO methods (row 3 of Fig. 1) offer at least the *possibility* of beating the simple average. The remainder of this paper explores various strategies for attaining the ex post theoretical peLASSO gains in ex ante peLASSO practice.

3.4. On the set of selected forecasters

One might wonder about the nature and evolution of the set of forecasters selected by our peLASSO procedures. The selected forecasters are identical across the procedures, period-by-period, because the first step is always the same (LASSO). We show them in Fig. 2, as we roll through the sample. The x-axis denotes time and the y-axis denotes forecaster ranking, where a smaller y-axis location refers to a forecaster with smaller overall RMSE. A "+" symbol at location (x, y) indicates that forecaster y was selected at time x.

A number of results emerge. First, the selected set is usually small, with three or four forecasters (as also mentioned earlier in conjunction with Table 1), yet also usually "democratic" in the sense that it is composed of some ex post top performers, some ex post average performers, and some ex post poor performers. Related, the ex post best forecaster (ID 1) is not always selected, and conversely, the ex post worst forecasters (ID's 22 and 23) are sometimes selected, mostly toward the end of the sample following the Great Recession.

Second, the selected set is not dominated by any one forecaster, or a small set of forecasters. Different forecasters move in and out of the selected set as we roll through the sample.[12] This may be due to different forecasters having different skills, which are relevant at different times. Some may be better in recessions and others in recoveries, some may have more insights into macro-finance interactions, etc.[13]

Finally, although the selected set is evolving, it is not at all independent over time; that is, the forecasters are not exchangeable. If a forecaster is in the selected set at time t, it is highly likely that she will be in the selected set at time $t + 1$. This is evident from the many "horizontal streaks" in Fig. 2.

4. Sophisticated averaging inspired by the ex post optimal peLASSO tuning

Here, motivated by the structure of the ex post (infeasible) peLASSO solution, we propose and explore procedures that implement that structure directly (discard most forecasts and average the survivors), while eliminating the need for penalty parameter selection. Our procedures implicitly perform sophisticated forward-looking cross-validation that is tailored precisely to the forecasting problem at hand, but again, with no need for the selection of penalty parameters.

[12] The forecasters selected most frequently are ID 6 (32 out of 65 quarters), ID 1 (27 quarters), and ID 22 (25 quarters). Five forecasters are never selected: IDs 2, 4, 7, 14, and 19.

[13] Note for example the long streaks of IDs 1, 6, and 22 immediately following the Great Recession.

Table 2
Individual-based average-best forecast combination.

Average-best N	RMSE	#	DM	p-val
$N = 1$	1.46	1	0.33	0.37
$N = 2$	1.42	2	0.77	0.22
$N = 3$	1.41	3	0.84	0.20
$N = 4$	1.41	4	0.95	0.17
$N = 5$	1.42	5	1.09	0.14
$N = 6$	1.43	6	1.11	0.14
Average-Best $\leq N_{max}$	RMSE	#	DM	p-val
$N_{max} = 1$	1.46	1.00	0.33	0.37
$N_{max} = 2$	1.44	1.45	0.55	0.29
$N_{max} = 3$	1.44	1.63	0.54	0.30
$N_{max} = 4$	1.44	1.74	0.57	0.29
$N_{max} = 5$	1.44	1.83	0.57	0.29
$N_{max} = 6$	1.44	1.86	0.57	0.29
Comparisons	RMSE	#	DM	p-val
Best	1.40	1	0.61	0.27
90%	1.44	1	0.63	0.27
Median	1.53	1	−0.57	0.72
10%	1.68	1	−1.61	0.94
Worst	1.74	1	−1.55	0.94
Average	1.50	23	N/A	N/A

Notes: # is the average number of forecasters selected, DM is the one-sided (Diebold & Mariano, 1995) statistic against a simple average, and p-val is the associated p-value. We compute DM as per Harvey et al. (1999).

Table 3
LASSO-based average-best forecast combination.

Average-best N	RMSE	#	DM	p-val
$N = 1$	1.56	1	−1.59	0.94
$N = 2$	1.53	2	−0.55	0.71
$N = 3$	1.45	3	0.87	0.19
$N = 4$	1.45	4	0.92	0.18
$N = 5$	1.46	5	0.86	0.20
$N = 6$	1.47	6	0.89	0.19
Average-best $\leq N_{max}$	RMSE	#	DM	p-val
$N_{max} = 1$	1.56	1	−1.59	0.94
$N_{max} = 2$	1.50	1.82	0.14	0.45
$N_{max} = 3$	1.47	2.35	0.55	0.29
$N_{max} = 4$	1.47	2.51	0.54	0.29
$N_{max} = 5$	1.47	2.57	0.57	0.29
$N_{max} = 6$	1.47	2.57	0.57	0.29
Comparisons	RMSE	#	DM	p-val
Best	1.40	1	0.61	0.27
90%	1.44	1	0.63	0.27
Median	1.53	1	−0.57	0.72
10%	1.68	1	−1.61	0.94
Worst	1.74	1	−1.55	0.94
Average	1.50	23	N/A	N/A

Notes: # is the average number of forecasters selected, DM is the one-sided (Diebold & Mariano, 1995) statistic against a simple average, and p-val is the associated p-value. We compute DM as per Harvey et al. (1999).

4.1. "Average best" combination

Motivated directly by the ex post peLASSO solution, we select a small number N of "best" forecasts and average them. There are two ways of doing the selection: from an individual perspective and from a portfolio perspective. We consider the two in turn.

4.1.1. Individual-based average-best combination

At each time, rolling forward, we determine the best N individual forecasters over the past 20 quarters, then average their 1-year-ahead forecasts. We refer to this as "individual-based average-best N" forecast combination.

Average-best combination requires the selection of N, and the results of course depend on N. As is shown in Table 2, and as expected, there is an internal optimum (minimum) RMSE for small values of N (3 or 4) for the individual-based average-best. Moreover, the optimized RMSE is highly competitive: much better than the ex post worst forecaster, noticeably better than the simple average, and indeed about as good as the ex post best forecaster. However, the DM statistics, are only borderline significant at best, presumably due to the very small forecast evaluation sample size, as was also the case for the infeasible peLASSO.

There is a slight ex post aspect of the good performance of average-best N forecasts, because the optimal N is not known ex ante. Instead of fixing N arbitrarily, we can proceed as follows: examine the historical performance of average-best N for $N = 1, \ldots, N_{max}$ at each time, then pick the best, and use that N and those forecasters to produce the forecast. We refer to this as "individual-based average-best $\leq N_{max}$" forecast combination, and it also appears in Table 2. The RMSEs tend to drop with N, quickly reaching an asymptote around $N = 3$.

4.1.2. Portfolio (LASSO)-based average-best combination

We have already noted the "trim and average" form of the ex post optimal peLASSO solution. However, it is important to note that its trimming is sophisticated, inasmuch as peLASSO does not trim the worst forecasters from an individual perspective. Rather, peLASSO trims the worst forecasters from a *portfolio* perspective; that is, those forecasters with the least to contribute to the combined forecast. The two concepts are very different, and so far we have considered only the individual perspective. The portfolio perspective suggests a related but different portfolio-based average-best N strategy: at each time, rolling forward, use the LASSO to determine the best N forecasters over the relevant window, then average their forecasts. We refer to this as "LASSO-based average-best N" forecast combination. The results appear in Table 3, which also includes results for LASSO-based average-best $\leq N_{max}$ combinations. Surprisingly, the LASSO-based average-best forecasts perform no better than the individual-based average-best forecasts; in fact, they are slightly worse.

4.2. "Best average" combination

In the "average best" approach above, we select some number of best forecasters at each time, rolling forward, and average their forecasts. Here, we move to a "best average" approach, instead selecting directly over averages. At each time, rolling forward, we simply find the historically best-performing average, and use it. Best-average is the more direct approach.

A first strategy is "best N-average": at each time we determine the best-performing N-forecast average over a 20-quarter window and use it. A second strategy is "best

Table 4
Best-average forecast combination.

Best N-average	RMSE	#	DM	p-val
$N = 1$	1.46	1	0.33	0.37
$N = 2$	1.41	2	0.80	0.21
$N = 3$	1.42	3	0.78	0.22
$N = 4$	1.41	4	0.92	0.18
$N = 5$	1.42	5	1.11	0.13
$N = 6$	1.42	6	1.28	0.10
Best $\leq N_{max}$-average	RMSE	#	DM	p-val
$N_{max} = 1$	1.46	1	0.33	0.37
$N_{max} = 2$	1.44	1.52	0.61	0.27
$N_{max} = 3$	1.44	1.72	0.60	0.27
$N_{max} = 4$	1.44	1.80	0.60	0.28
$N_{max} = 5$	1.44	1.83	0.61	0.27
$N_{max} = 6$	1.44	1.83	0.61	0.27
Comparisons	RMSE	#	DM	p-val
Best	1.40	1	0.61	0.27
90%	1.44	1	0.63	0.27
Median	1.53	1	−0.57	0.72
10%	1.68	1	−1.61	0.94
Worst	1.74	1	−1.55	0.94
Average	1.50	23	N/A	N/A

Notes: # is the average number of forecasters selected, DM is the one-sided (Diebold & Mariano, 1995) statistic against a simple average, and p-val is the associated p-value. We compute DM as per Harvey et al. (1999).

Table 5
Forecast combination.

Best $(\leq 6, W)$-average	RMSE	#N	#W	DM	p-val
$W = 1$	1.42	1.14	1	1.14	0.13
$W = 2$	1.36	1.54	2	1.50	0.07
$W = 3$	1.37	1.45	3	1.41	0.08
$W = 4$	1.40	1.29	4	1.10	0.14
$W = 5$	1.42	1.41	5	0.93	0.18
$W = 6$	1.42	1.43	6	0.81	0.21
$W = 7$	1.44	1.43	7	0.65	0.26
$W = 8$	1.46	1.54	8	0.41	0.34
$W = 9$	1.47	1.70	9	0.37	0.36
$W = 10$	1.46	1.70	10	0.43	0.33
$W = 15$	1.44	1.77	15	0.66	0.26
$W = 20$	1.44	1.78	20	0.61	0.27
$W = 25$	1.46	1.57	25	0.40	0.34
$W = 30$	1.48	1.62	30	0.19	0.42
$W = 35$	1.48	1.67	35	0.29	0.39
$W = 40$	1.48	1.74	40	0.22	0.41
Best $(\leq 6, \leq 40)$-average	1.38	1.38	2.02	1.24	0.11
Comparisons	RMSE	#N	#W	DM	p-val
Best	1.40	1	N/A	0.61	0.27
90%	1.44	1	N/A	0.63	0.27
Median	1.53	1	N/A	−0.57	0.72
10%	1.68	1	N/A	−1.61	0.94
Worst	1.74	1	N/A	−1.55	0.94
Average	1.50	23	N/A	N/A	N/A

Notes: #N is the average number of forecasters selected, #W is the average window width selected, DM is the one-sided (Diebold & Mariano, 1995) statistic against a simple average, and p-val is the associated p-value. We compute DM as per Harvey et al. (1999).

$\leq N_{max}$-average": at each time we determine the best-performing $\leq N_{max}$-forecast average by means of a 20-quarter window and use it.

Best-average combining can involve significant computation, depending on K and N or N_{max}. For example, with 23 forecasters, finding the best six-average requires the computation of ${}_{23}C_6$ (= 100,947) simple averages, which must then be sorted to determine the minimum, each period as we roll through time. The per-period computational burden of $\leq N_{max}$-forecast averaging is still larger, because we now consider all subsets. For example, finding the best ≤ 6-average with 23 forecasters requires the computation of ${}_{23}C_6 + {}_{23}C_5 + \cdots + {}_{23}C_1$ (= 145,498) simple averages, which must then be sorted to determine the minimum. Fortunately, the relevant K and N_{max} are quite small in typical economic forecast combinations. In our case, for example, $K = 23$, and $N_{max} \leq 6$ appears more than adequate.

Table 4 shows results for both the best N-average combinations ($N = 1, \ldots, 6$) and the best $\leq N_{max}$-average combinations ($N_{max} = 1, \ldots, 6$). For both variations, the optima are achieved for small values of N or N_{max}. One might expect best-average methods to outperform average-best, because best-average targets the object of interest directly. However, best-average does not outperform, though it does not underperform either: it is at least as good as anything else. The best-average ≤ 6 RMSE is almost as good as that of the best individual, much better than that of the median individual, and, importantly, better than that of the simple average.

4.3. Window width estimation

The essence of the rolling best-average approach is simply to use the particular average that has performed best in the "recent" past. However, there is of course no reason why the appropriate notion of "recent" (that is, the appropriate choice W of the most recent W quarters for evaluation) should be $W = 20$. Using a more complete notation, let us denote our earlier best N-average as the best $(N, 20)$-average, to indicate both an N-forecast average and a 20-period evaluation window. Generically, then, we can speak of the best (N, W)-average or best $(\leq N_{max}, W)$-average combinations.

The first panel of Table 5 shows results for the best $(\leq N_{max}, W)$-average combinations, with $N_{max} = 6$ and W ranging from 1 to 40. The RMSE performance of the best $(\leq 6, W)$-average approach is relatively insensitive to W, but is clearly optimized for very small values of W, around 2 or 3. Interestingly, the average number of forecasters selected around the optimal W is also very small ($N \approx 2$). Thus, the optimal procedure (best $(\leq 6, 2)$-average) is very "localized": each period it basically averages the two forecasts of the two forecasters who have performed the best during the past two quarters. It has an *RMSE* that is better than that of the ex post best forecaster, and much better than that of the average forecaster, with a DM p-value of 0.07.

We can also allow for a time-varying window width W; that is, we can work with best $(\leq N_{max}, \leq W_{max})$-averages, which are completely ex ante. They turn out to work very well: the best $(\leq 6, \leq 40)$-average (in the one-line middle panel of Table 5) has an *RMSE* that is better than that of the best forecaster, and much better than that of the average

forecaster, with a DM p-value of 0.11. All told, allowing for a time-varying window width appears to be highly valuable.

5. Related literature

Now that we have introduced our approach, we can relate it to certain aspects of the broader literature.

5.1. On selection

The structure of the peLASSO solution, which motivates our direct average-best and best-average procedures, clearly involves harsh "trimming", resulting in the elimination of most forecasters. Trimming has been used in forecast combination by many authors, such as Aiolfi and Favero (2005), Aiolfi and Timmermann (2006), Bjørnland, Gerdrup, Jore, Smith, and Thorsrud (2012), Genre et al. (2013) and Stock and Watson (1999). However, as was noted by Granger and Jeon (2004), the attractiveness of trimming may be "more of a pragmatic folk-view than anything based on a clear theory".

One can view our results as showing the clear emergence of the "folk view" in a framework rigorously based on a "clear theory". In particular, although in principle the peLASSO solution need not involve trimming (i.e., it is possible for the peLASSO solution to feature shrinkage but not selection), we have shown that in practice it *does*, and indeed that it involves heavy trimming. Interestingly, Samuels and Sekkel (2017) obtain the same result using a very different approach based on the "model confidence sets" of Hansen, Lunde, and Nason (2011). Note, moreover, that both our trimming procedure (in peLASSO, LASSO-based average-best, and best-average) and that of Samuels and Sekkel (2017) are generally quite sophisticated, inasmuch as they trim from a portfolio perspective rather than a stand-alone perspective. Most impressively, Conflitti et al. (2015) impose sum-to-one and non-negativity constraints, which lead to a sparse solution (that is, some of the combination weights are exactly zero) with combining weights shrunken – indeed forced – to be within [0, 1], all of which is in close touch with the concerns of forecast combination.[14]

5.2. On shrinkage

Several authors have considered Bayesian shrinkage of combining weights. As is well known, under standard conditions the Bayes rule under quadratic loss is

$$\beta_1 = \beta_0 + \delta\left(\hat{\beta}_{OLS} - \beta_0\right),$$

where β_1 is the posterior mean combining weight vector, β_0 is the prior mean vector, and $\delta \in [0, 1]$ is related inversely to the prior precision. Other things being equal, a small value of δ implies a high prior precision, and hence, substantial shrinkage toward β_0. The larger the value of δ, the less shrinkage occurs. Different authors invoke different shrinkage directions (prior means) and different ways of choosing δ. Relevant studies include those by Aiolfi and Timmermann (2006), Chan, Stock, and Watson (1999), Diebold and Pauly (1990), Genre et al. (2013) and Stock and Watson (2004).

In an interesting development, Capistrán and Timmermann (2009) take a reverse approach. Whereas Bayesian shrinkage adjusts least-squares combining weights toward a simple average, Capistrán and Timmermann (2009) start with a simple average and adjust *away* from it via a Mincer–Zarnowitz regression, $y_t \rightarrow c, \bar{f}_t$.

5.3. Relatives of peLASSO

The reverse approach of Capistrán and Timmermann (2009) has an interesting connection to the so-called "OSCAR LASSO" proposed by Bondell and Reich (2008), which is also related closely to our methods.

First let us introduce OSCAR. It is defined by the penalized regression:

$$\hat{\beta}_{OSCAR} = \arg\min_{\beta} \sum_{t=1}^{T}\left(y_t - \sum_{i=1}^{K}\beta_i x_{it}\right)^2$$
$$\text{s.t. } (1-\gamma)\sum_{i=1}^{K}|\beta_i| + \gamma\sum_{j<k}\max\left\{|\beta_j|, |\beta_k|\right\} \leq c. \qquad (4)$$

The first part of the constraint involves the L_1 norm; it is just the standard LASSO constraint, producing selection and shrinkage toward zero. The second part of the constraint involves the pairwise L_∞ norm, which selects and shrinks toward equal coefficients. Overall, then, OSCAR regression encourages parsimony not only in standard LASSO fashion, but also by encouraging a small number of unique nonzero coefficients on surviving covariates.[15]

Now let us link to Capistrán and Timmermann (2009). Suppose that the OSCAR solution is "all coefficients are the same". This can occur because of the second part of the OSCAR constraint. Then the combined forecast is

$$\hat{C}_t = \hat{\beta}\sum_{i}^{K} f_{i,t}$$
$$= \hat{\alpha}\left(\frac{1}{K}\sum_{i=1}^{K} f_{i,t}\right),$$

which is the forecast we get by projecting the realized outcome onto equal-weight forecasts, as per Capistrán and Timmermann (2009). The OSCAR solution may also have more than one unique coefficient. In particular, it may have multiple groups, as for example with

$$\hat{C}_t = \hat{\beta}_1\sum_{i\in G_1} f_{i,t} + \hat{\beta}_2\sum_{i\in G_2} f_{i,t}$$
$$= \hat{\alpha}_1\left(\frac{1}{N_1}\sum_{i\in G_1} f_{i,t}\right) + \hat{\alpha}_2\left(\frac{1}{N_2}\sum_{i\in G_2} f_{i,t}\right),$$

[14] Their estimator can be shown to be a special case of LASSO.

[15] Note however that, although OSCAR shrinks toward "equal weights", the equal weights need not correspond to simple averages (e.g., each of three selected forecasters might get a weight of 1/2). This is potentially very important in our context.

where $G_k = \{i : \hat{\beta}_i = \hat{\beta}_k\}$ and N_k is the size of group G_k. The approaches of Aiolfi and Timmermann (2006) and Genre et al. (2013), which allow for grouping, are in the same spirit, as are the "homogeneity pursuit" procedure of Ke, Fan, and Wu (2015) and the "HORSES" procedure of Jang, Lim, Lazar, Loh, and Yu (2015).

5.4. Relatives of average-best and best-average

The work of Burgi and Sinclair (2017) is related to our average-best approach, as it essentially amounts to a refinement of our "individual" average-best. They proceed as follows: (1) for each forecaster, calculate a variable that takes a value of one in a given period if that forecaster has a lower squared error in that period than the simple average and zero otherwise[16]; (2) if a forecaster beats the simple average more often than a given percentage threshold p, include that forecaster in the selected subset for the next forecasting period; and finally (3) average over the selected forecasters.

However, the work that is related most closely to ours is the seminal (and, to the best of our knowledge, relatively unknown) work of Elliott (2011), who examines the gains from optimal combination relative to simple averaging, provides conditions under which the two are equivalent, and explores aspects of what he calls "best subset averaging". Effectively, we provide a foundation for Elliott's subset-averaging procedures, which initially appear ad hoc in theory, even if highly effective in practice. That is, we show that Elliott's procedures are *not* ad hoc in theory.

6. Concluding remarks

Against a background of frequently-found superiority of simple-average forecast combinations, we have proposed "partially egalitarian LASSO" (peLASSO) procedures that discard some forecasts and then select and shrink — without forcing — the remaining forecasts toward equal weights. We found that the peLASSO solution involves discarding most forecasters and simply averaging the survivors. We have therefore proposed alternative direct combination procedures, most notably "best average" combinations, and showed that they seem highly competitive for out-of-sample forecasting. In particular, they often dominate simple averages in forecasting Eurozone GDP growth.

A key insight is that the structure and success of our averaging procedures are entirely motivated by and consistent with the lessons learned from peLASSO. Among other things, we learn from peLASSO that (1) the selection penalty should be quite harsh, as only a few forecasts need be combined; (2) the forecasts selected for combination should be regularized via shrinkage; (3) the shrinkage direction should be toward a simple average, not toward zero or anything else; and (4) the shrinkage should be extreme, so that the selected forecasts should simply be averaged. All of this is embedded in our best $\leq (N_{max}, W_{max})$-average procedure, for small values of N_{max} and W_{max}.

[16] The time-average of this variable is the historical percentage share of times that the forecaster has beaten the average.

Acknowledgments

This is a revised and extended version of our previously-circulated manuscript, "Beating the Simple Average: Egalitarian LASSO for Combining Economic Forecasts". We are grateful for comments from the editors (Domenico Giannone, George Kapetanios, and Mike McCracken), two anonymous referees, and Umut Akovali, Xu Cheng, Denis Chetverikov, Edgar Dobriban, Ed George, Mike Kearns, Laura Liu, Ken McAlinn, Rob McCulloch, Hashem Pesaran, Veronika Rockova, Zhentao Shi, Tara Sinclair, Stephen Stigler, Dongho Song, Allan Timmermann, Weijie Su, Jonathan Wright, and Boyuan Zhang. The paper also benefitted from presentations at FRB St. Louis, Brown, Chicago, TU Vienna, York, the ECB, and the NBER. The usual disclaimer applies.

Appendix A. Egalitarian LASSO and egalitarian ridge implementation

The egalitarian LASSO can be implemented via a straightforward adaptation of standard LASSO software, such as the R package glmnet, written by J. Friedman, T. Hastie, N. Simon, and R. Tibshirani and found at https://cran.r-project.org/web/packages/glmnet/index.html. Simply note that

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

where

$$\delta_i = \beta_i - \frac{1}{K} \quad \text{and} \quad \bar{f}_t = \frac{1}{K}\sum_{i=1}^{K} f_{it}.$$

Hence, we obtain the egalitarian LASSO regression

$$y_t \rightarrow_{EgalLASSO} f_{1t}, \ldots, f_{Kt},$$

by simply running the standard LASSO regression

$$(y_t - \bar{f}_t) \rightarrow_{LASSO} f_{1t}, \ldots, f_{Kt}. \tag{A.1}$$

Similarly, the egalitarian ridge can be implemented trivially by $(y_t - \bar{f}_t) \rightarrow_{Ridge} f_{1t}, \ldots, f_{kt}$, in precise parallel with egalitarian LASSO implementation.

Appendix B. Equal-length, 20-quarter evaluation windows

Here, we start from $t = 21$ rather than $t = 6$ when evaluating forecasts, and hence, the estimation samples are always of length 20. The results, reported in Table B.1, are identical qualitatively to those reported in the main text, for which the estimation sample sizes grow to $t = 21$, after which they are always of length 20.

Table B.1
Forecast RMSEs based on ex post optimal λs, evaluation starts at $t = 21$.

Regularization group	RMSE	λ*	#	DM	p-val
Ridge	1.60	2.29	23.00	0.65	0.26
LASSO	1.61	0.38	2.78	0.22	0.42
eRidge	1.58	1.97	23.00	0.96	0.17
eLASSO	1.60	0.51	23.00	0.82	0.21
peLASSO (LASSO, Average)	1.51	0.21	3.12	1.14	0.13
peLASSO (LASSO, eRidge)	1.50	(0.21, 3.10)	3.12	1.06	0.15
peLASSO (LASSO, eLASSO)	1.50	(0.21, 0.51)	3.12	1.03	0.16
Comparisons	RMSE	λ*	#	DM	p-val
Best	1.49	N/A	1	0.76	0.23
90%	1.54	N/A	1	0.93	0.18
Median	1.65	N/A	1	−0.38	0.65
10%	1.82	N/A	1	−1.37	0.91
Worst	1.90	N/A	1	−1.46	0.92
Average	1.64	N/A	23	N/A	N/A

Notes: λ* is the ex post optimal penalty parameter(s), # is the average number of forecasters selected, and DM is the one-sided (Diebold & Mariano, 1995) statistic against a simple average, with p-value denoted p-val. We compute DM as per Harvey et al. (1999).

Appendix C. Adaptive partially-egalitarian LASSO

We change the LASSO penalty from $\lambda \sum_{k=1}^{K} |\beta_k|$ to $\lambda \sum_{k=1}^{K} \frac{1}{|\hat{\beta}|^{1/3}} |\beta_k|$, where $\hat{\beta}$ is a preliminary consistent estimator, which we set to the Ridge regression estimate. The use of aLASSO improves the ex post performance of two-step LASSO procedures, as can be seen from Table C.1.

Table C.1
RMSEs based on ex post optimal λs, using the adaptive lasso.

Regularization group	RMSE	λ*	#	DM	p-val
Ridge	1.51	2.66	23.00	−0.02	0.51
LASSO	1.46	0.80	2.09	0.22	0.41
eRidge	1.49	1.97	23.00	0.15	0.44
eLASSO	1.50	max	23.00	0.55	0.29
peLASSO(aLASSO, Average)	1.33	1.08	1.69	0.95	0.17
peLASSO (aLASSO, eRidge)	1.33	(1.08, max)	1.69	0.95	0.17
peLASSO (aLASSO, eLASSO)	1.33	(1.08, max)	1.69	0.95	0.17
Comparisons	RMSE	λ*	#	DM	p-val
Best	1.40	N/A	1	0.61	0.27
90%	1.44	N/A	1	0.63	0.27
Median	1.53	N/A	1	−0.57	0.72
10%	1.68	N/A	1	−1.61	0.94
Worst	1.74	N/A	1	−1.55	0.94
Average	1.50	N/A	23	N/A	N/A

Notes: λ* is the ex post optimal penalty parameter(s), # is the average number of forecasters selected, and DM is the one-sided (Diebold & Mariano, 1995) statistic against a simple average, with p-value denoted p-val. We compute DM as per Harvey et al. (1999).

References

Aiolfi, M., & Favero, C. (2005). Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting*, *24*(4), 233–254.

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1), 31–53.

Arora, S., Hazon, E., & Kale, S. (2012). The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, *8*, 121–164.

Aruoba, S., Diebold, F., Nalewaik, J., Schorfheide, F., & Song, D. (2012). Improving GDP measurement: A forecast combination perspective. In X. Chen, & N. Swanson (Eds.), *Recent advances and future directions in causality, prediction, and specification analysis: Essays in honour of Halbert L. White Jr.* (pp. 1–26). Springer.

Bates, J., & Granger, C. (1969). The combination of forecasts. *Operations Research Quarterly*, *20*, 451–468.

Bjørnland, H., Gerdrup, K., Jore, A., Smith, C., & Thorsrud, L. (2012). Does forecast combination improve norges bank inflation forecasts? *Oxford Bulletin of Economics and Statistics*, *74*, 163–179.

Bondell, H., & Reich, B. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, *64*(1), 115–123.

Burgi, C., & Sinclair, T. (2017). A nonparametric approach to identifying a subset of forecasters that outperforms the simple average. *Empirical Economics*, *53*, 101–115.

Capistrán, C., & Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics*, *27*(4), 428–440.

Chan, Y., Stock, J., & Watson, M. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review*, *1*(2), 91–121.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography (with discussion). *International Journal of Forecasting*, *5*(4), 559–583.

Conflitti, C., De Mol, C., & Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, *31*, 1096–1103.

Diebold, F. (1989). Forecast combination and encompassing: Reconciling two divergent literatures. *International Journal of Forecasting*, *5*, 589–592.

Diebold, F., & Lopez, J. (1996). Forecast evaluation and combination. In G. Maddala, & C. Rao (Eds.), *Handbook of statistics* (pp. 241–268). North-Holland.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 253–365.

Diebold, F., & Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, *6*(4), 503–508.

Elliott, G. (2011). *Averaging and the optimal combination of forecasts, manuscript*. Department of Economics, UCSD.

Elliott, G., & Timmermann, A. (2016). *Economic forecasting*. Princeton 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.

Granger, C., & Jeon, Y. (2004). Thick modeling. *Empirical Economics*, *21*, 323–343.

Granger, C., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3*, 197–204.

Hansen, P., Lunde, A., & Nason, J. (2011). The model confidence set. *Econometrica*, *79*, 453–497.

Harvey, D., Leybourne, S., & Newbold, P. (1999). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*, 281–291.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer.

Jang, W., Lim, J., Lazar, N., Loh, J., & Yu, D. (2015). Some properties of generalized fused lasso and its applications to high dimensional data. *Journal of the Korean Statistical Society*, *44*(3), 352–365.

Ke, Z., Fan, J., & Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association*, *110*, 175–194.

Samuels, J., & Sekkel, R. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, *33*(1), 48–60.

Smith, J., & Wallis, K. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355.

Stock, J., & Watson, M. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R. Engle, & H. White (Eds.), *Cointegration, causality, and forecasting*. Oxford University Press.

Stock, J., & Watson, M. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430.

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

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1) (pp. 135–196). Elsevier.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel, & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of bruno de finetti* (pp. 233–243). North Holland, Amsterdam.

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

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*, 302–320.

Francis X. Diebold is Paul F. and Warren S. Miller Professor of Social Sciences and Professor of Economics, Finance, and Statistics at the University of Pennsylvania. He is also a Faculty Research Associate at the National Bureau of Economic Research, an elected Fellow of the Econometric Society, the American Statistical Association, and the International Institute of Forecasters.

Minchul Shin is an Assistant Professor of Economics at the University of Illinois at Urbana-Champaign.