Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper 17-017

Beating the Simple Average: Egalitarian LASSO for Combining Economic Forecasts

by

Francis X. Diebold

Minchul Shin

http://ssrn.com/abstract=3032492

Beating the Simple Average: Egalitarian LASSO for Combining Economic Forecasts

Francis X. Diebold
University of Pennsylvania

Minchul Shin
University of Illinois

August 20, 2017

Abstract: Despite the clear success of forecast combination in many economic environments, several important issues remain incompletely resolved. The issues relate to 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 superiority of simple-average combinations, we propose LASSO-based procedures that select and shrink toward equal combining weights. We then provide an empirical assessment of the performance of our "egalitarian LASSO" procedures. The results indicate that simple averages are highly competitive, and that although out-of-sample RMSE improvements on simple averages are possible in principle using our methods, they are hard to achieve in real time, due to the intrinsic difficulty of small-sample real-time cross validation of the LASSO tuning parameter. We therefore propose alternative direct combination procedures, most notably "best average" combination, motivated by the structure of egalitarian LASSO and the lessons learned, which do not require choice of a tuning parameter yet outperform simple averages.

Acknowledgments: For comments we are grateful to Denis Chetverikov, Laura Liu, Dongho Song, Zhentao Shi, and Boyuan Zhang. The usual disclaimer applies.

Key words: Forecast combination, forecast surveys, shrinkage, model selection, LASSO, regularization

JEL codes: C53

Contact: fdiebold@upenn.edu

Contents

1 Introduction

Forecast combination has a long and successful history in economics.[1] Important issues remain incompletely resolved, however, 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 a LASSO-based procedure that addresses all considerations.

We proceed as follows. In section 2, we highlight aspects of the "equal-weights puzzle", that is, the frequently-found superiority of simple-average combinations, as relevant for our eventual concerns. In section 3 we describe our "egalitarian LASSO" procedures, which shrink and select toward simple averages. In section 4 we place our methods in the context of the broader literature. In section 5 we provide ex post empirical assessment of our procedures' performance in comparison to others, and in section 6 we provide ex ante (real-time) assessment. We conclude in section 7.

2 On The Good Performance of Equal Combining Weights

Although it seems natural to average forecasts (i.e., to to use equal-weight combinations), simple averages are generally suboptimal, and potentially highly so. To see the theoretical sub-optimality of equal combining weights, consider K competing unbiased forecasts $f_t^1, ..., f_t^K$ of a series y_t made at time $t-1$, $t = 1, ..., T$. We form a combined forecast as

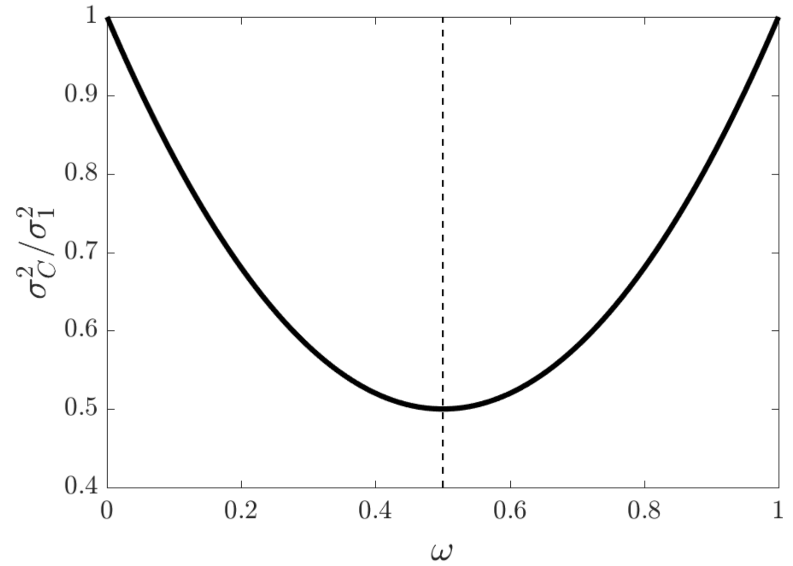
$$C_t = \omega_1 f_t^1 + \omega_2 f_t^2 + ... + \left(1 - \sum_{k=1}^{K-1} \omega_i\right) f_t^K.$$

The corresponding forecast errors (e_C and e_1, ..., e_K, with variances σ_C^2 and $\sigma_1^2, ..., \sigma_1^K$, respectively) satisfy the same equality, from which it follows that the variance of the combined forecast error is minimized using weight vector

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

[1] For a recent overview, see Elliott and Timmermann (2016).

Figure 1: σ_C^2/σ_1^2 vs. ω, for $\omega \in [0,1]$



Notes: We assume $\phi = 1$ and uncorrelated errors. See text for details.

where Σ is the variance-covariance matrix of 1-step-ahead forecast errors and $\mathbf{i}$ is a conformable column vector of ones (Bates and Granger (1969)). Note in particular that *the optimal weights (1) are generally not* $1/K$. As an example, consider two forecasts with uncorrelated errors.[2] Then (1) reduces to

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

where ω^* 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.

Despite the theoretical sub-optimality of equal weights, a large literature finds generally good performance of simple averages under quadratic loss. Indeed the forecast combination "equal weights puzzle" (referring to equal weights of $1/K$ in combining $i = 1, ..., K$ forecasts) refers to the frequently-found good performance of simple averages. The equal-weights puzzle

[2] For simplicity and maximum transparency we work with uncorrelated errors throughout this section, but the qualitative results hold even with correlation. See Aruoba et al. (2012).

was noted long ago (e.g., Clemen (1989)), but it has not been fully resolved and continues to be the subject of much research (e.g., Timmermann (2006), Smith and Wallis (2009), Elliott (2011)).[3]

It turns out that the equal weights puzzle is ultimately not so puzzling. Two crucial observations, in particular, favor equal weights. First, *even if simple averages are not fully optimal, they are likely to be much better than any individual forecast.* To build intuition let us stay with the case of two forecasts with uncorrelated errors. It is instructive to compare the error variance of the optimally-combined forecast, σ_C^2, to the error variance of one of the primary forecasts (with no loss of generality, let us compare to σ_1^2), for a range of ω values (including $\omega = \omega^*$, $\omega = 0$, and $\omega = 1$). Simple calculations reveal that:

$$\frac{\sigma_C^2}{\sigma_1^2} = \omega^2 + \frac{(1-\omega)^2}{\phi^2}.$$

In Figure 1 we graph σ_C^2/σ_1^2 for $\omega \in [0,1]$ with $\phi = 1$. Obviously the maximum variance reduction is obtained using $\omega^* = 0.5$, but even for nonoptimal ω we achieve substantial variance reduction relative to using f_1 alone. That is a key result: *for all* ω (except those very close to 0 or 1, of course) we achieve substantial variance reduction.

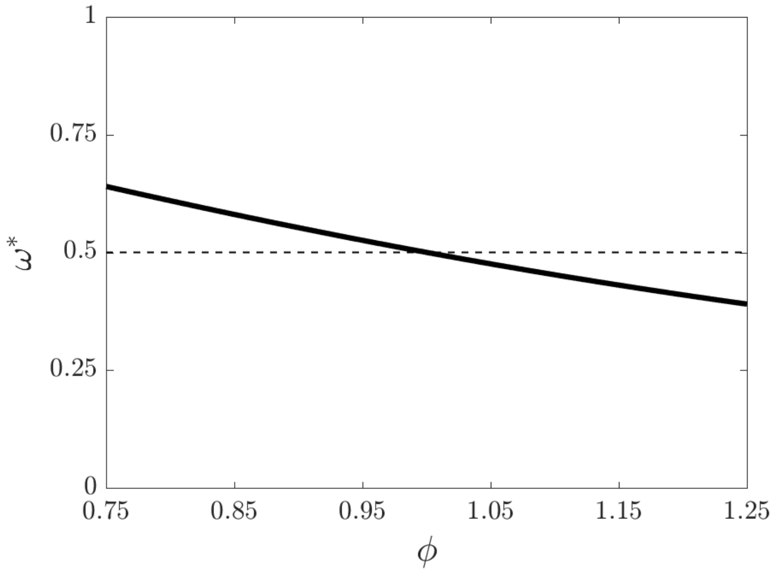
Second, *even if simple averages are not fully optimal, they are likely to be nearly optimal.* In Figure 2 we graph ω^* as a function of ϕ, for $\phi \in [.75, 1.25]$. ω^* is of course decreasing in ϕ, but interestingly, it is only mildly sensitive to ϕ. Indeed, for our range of ϕ values, the optimal combining weight remains close to 0.5. Under a wide range of conditions it is optimal to put significant weight on both f_1 and f_2, with the optimal weights not differing radically from equality.

3 Synthesis: Shrinkage and Selection with Variants of Egalitarian LASSO

It is well known (Granger and Ramanathan (1984)) that the population Bates-Granger optimal combining weights (1) introduced above may be trivially obtained from the population regression (linear projection) $y_t \rightarrow f_t^1, ..., f_t^K$, subject to the constraint that the coefficients

[3] Note well 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. Aruoba et al. (2012), for example, have recently shown that equal weights are optimal under minimax loss.

Figure 2: ω^* vs. ϕ, for $\phi \in [.75, 1.25]$



Notes: We construct ω^* assuming uncorrelated errors, and we include a horizontal line for visual reference at $\omega^* = .5$. See text for details.

add to one.[4] So the theoretical optimal linear forecast combination problem is just a population linear regression (projection) problem, and combining weight estimation involves just a finite-sample linear regression.

The discussion of section 2 strongly suggests, moreover, that simple averages (equal weights) are a natural shrinkage direction for such combining regressions. With shrinkage, we don't *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. So we potentially want to set *some* combining weights to zero ("select to zero") and shrink the *remaining* weights toward equality ("shrink toward equality"). LASSO-based methods almost do the trick – they both select and shrink ((Tibshirani, 1996; Hastie et al., 2009)) – but unfortunately they select to zero *and* shrink

[4] Moreover, there is no real need to impose the "sum-to-one" constraint, and it is now rarely imposed.)

to zero. In the remainder of this section we start with standard LASSO and modify it until we arrive at our "2-step egalitarian LASSO", which selects to zero and shrinks to equality. Interestingly, each of the estimators introduced en route to the 2-step egalitarian LASSO will prove useful in its implementation.

3.1 Penalized Estimation for Selection and Shrinkage

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

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

Equivalently, in Lagrange-multiplier form we can write

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

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

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

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

3.2 Shrinkage Toward Equal Weights: Egalitarian Ridge

Smooth convex penalties in (2) produce pure shrinkage. In particular, $q = 2$ produces ridge regression, which shrinks coefficients toward 0:

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

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

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

Egalitarian ridge is closely related 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 much simpler to implement.

Note that, although egalitarian ridge will feature later in this paper, which is why we introduced it, it is inadequate for our ultimate purpose – it shrinks in the right direction but does not select.

3.3 Selection and Shrinkage Toward Zero Weights: Standard LASSO

As we have seen, $q = 2$ 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:[5]

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

There are several variants of LASSO:

> (1) Adaptive LASSO weights the terms in the constraint to encourage setting small first-round coefficient estimates 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 $\nu > 0$.

[5] In addition, $q = 1$ is the smallest q for which the penalized estimation minimization problem is (conveniently) convex.

(2) Elastic net blends the LASSO penalty ($q = 1$) and the ridge penalty ($q = 2$),

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

(3) Adaptive elastic net blends everything,

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

Under conditions, the adaptive LASSO variants have the oracle property. The elastic net variants have good properties in handling highly-correlated predictors. Adaptive elastic net has both. Note that standard LASSO and its variants, although improving on ridge insofar as they shrink *and* select, remain inadequate for our purposes – they select in the right direction but shrink in the wrong direction.

3.4 Selection and Shrinkage Toward Equal Weights: Egalitarian LASSO

All of the standard LASSO variants in section 3.3 select and shrink combining weights toward zero, but that is not what we want. Instead, as discussed in section 2, both theory and experience point clearly to shrinkage toward simple averages. We therefore change the LASSO penalized estimation problem to

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

That is, instead of shrinking the weights toward zero, we shrink the *deviations from equal weights* toward zero.

We can continue to use standard LASSO software to do the modified minimization.[6]

[6] See, for example, the R package glmnet, written by J. Friedman, T. Hastie, N. Simon, and R. Tibshirani, at `https://cran.r-project.org/web/packages/glmnet/index.html`.

Simply note that

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

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

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

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,

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

by simply running the standard LASSO regression,

$$(y_t-\bar{f}_t)\ \rightarrow_{LASSO}\ f_{1t},...,f_{Kt}. \tag{3}$$

Note that although egalitarian LASSO shrinks in the right direction, it is still unappealing, for reasons opposite those of standard LASSO. Like standard LASSO, egalitarian LASSO shrinks and selects, but whereas LASSO *shrinks* in the wrong direction, egalitarian LASSO *selects* in the wrong direction! We nevertheless introduced ridge, LASSO, and egalitarian LASSO because the procedure to which we now turn – which shrinks *and* selects in the right direction – is closely related, and because each will feature importantly in our subsequent empirical work.

3.5 Selection Toward Zero and Shrinkage Toward Equality: 2-step Egalitarian LASSO

Egalitarian LASSO does not tend to discard forecasters, because it selects and shrinks toward equal weights, not toward zero weights. In particular, the selection to equality, rather than zero, may be problematic, as it presumes that all forecasters "belong" in the set to be combined. One can easily modify the egalitarian LASSO, however, such that some forecasters are potentially discarded. Consider, for example, the following 2-step egalitarian LASSO procedure:

2-Step Egalitarian LASSO

> Step 1 (Selection to Zero): Using standard methods, select k forecasts from among the full set of K forecasts.
>
> Step 2 (Shrinkage Toward Equality): Using standard methods, shrink the combining weights on the k forecasts that survive step 1 toward $1/k$.

Obvious methods for step 1 include information criteria like SIC or AIC, potentially in conjunction with forward or backward stepwise methods in high-dimensional cases. Perhaps preferably, one could also use standard LASSO, which requires only one estimation and moreover can handle situations with $K > T$, which are not uncommon in forecast combination.[7]

An obvious method for step 2 is egalitarian ridge.[8] One could go even farther and use egalitarian *LASSO* for step 2, in which case the complete procedure would first select some weights to 0, and then select some of the surviving weights to $1/k$ and shrink the rest toward $1/k$.

In the empirical work of sections 5 and 6 below, we emphasize 2-step egalitarian LASSO and compare it to several competitors. But first, now that we have described our approach, we situate it in the literature.

4 Related Literature

Several authors have considered pure Bayesian shrinkage of combining weights, often toward equal weights.[9] As is well known, under standard conditions the Bayes rule under quadratic loss is

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

[7] In our subsequent empirical work in sections 5 and 6, for example, the combining regressions involve 25 forecasters but only 20 observations.

[8] Note that egalitarian ridge can be trivially implemented via a standard ridge regression with a transformed left-hand-side variable, $(y_t - \bar{f}_t) \rightarrow_{Ridge} f_{1t}, ..., f_{kt}$, in parallel with the with egalitarian LASSO regression (3).

[9] Some authors, including Conflitti et al. (2015) and Fuentes et al. (2014), have introduced selection as well by using standard LASSO regression to estimate combining weights. As we have argued, however, standard LASSO shrinks in an undesirable direction for purposes of forecast combination.

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

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, $y_t \rightarrow_{OLS} c, \bar{f}_t$. Genre et al. (2013) go even farther and segment the forecasters into n groups, projecting onto each of the group averages, $y_t \rightarrow_{OLS} c, \bar{f}_t^1, ..., \bar{f}_t^n$.

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

$$\hat{\beta}_{OSCAR} = \arg\min_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2$$

$$\text{s.t.} \quad (1-\gamma) \sum_{i=1}^{K} |\beta_i| + \gamma \sum_{j<k} \max\left\{|\beta_j|, |\beta_k|\right\} \leq c. \tag{4}$$

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.

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

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

which is the forecast we get by projecting the realized outcome on equal-weight forecasts, as in 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

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

where $G_k = \{i : \ \hat{\beta}_i = \hat{\beta}_k\}$ and $N_k = \#G_k$. Genre et al. (2013)) is obviously in the same spirit.

5 Combined Forecast Performance with Ex Post Optimal Tuning

In this section we begin our empirical work, providing a comparative assessment of our methods using the European Central Bank's well-known quarterly Survey of Professional Forecasters.[10] For those procedures that require selection of a tuning parameter λ, we examine out-of-sample RMSE based on the ex-post optimal λ (i.e., the λ that optimizes out-of-sample RMSE, which we can determine ex post).

5.1 Background

Again, we focus on the European Central Bank's well-known quarterly Survey of Professional Forecasters. Of course the comparative performance of our methods, whether good or bad, using a particular dataset and a particular implementation (choice of sample period, choice of tuning parameters, etc.) cannot establish anything conclusively, but that is not our intent. Rather, we simply seek to illustrate our methods in a realistic and important environment, and to provide some preliminary yet suggestive evidence regarding their performance.

We consider 1-year-ahead forecasts for the Euro-area real GDP growth rate. We are faced with an unbalanced panel, because forecasters enter and exit the pool in real time, and moreover, those in the pool at any time do not necessarily respond to the survey. Hence for ease of analysis we select the 25 forecasters who responded most frequently during our

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

evaluation sample (2000Q3–2014Q1) and impute missing observations using a linear filter as in Genre et al. (2013).

We roll through the evaluation sample, estimating combining weights using a 5-year (20-quarter) window and producing a 1-year-ahead out-of-sample forecast. 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 standard ridge, standard LASSO, egalitarian ridge, egalitarian LASSO, and three versions of 2-step egalitarian LASSO (the first step is always LASSO, and the second step is either simple average, egalitarian ridge, or egalitarian LASSO).[11] Throughout, we compare the combined forecasts to simple average combinations.

Each combining method except simple averages requires choosing a tuning parameter, λ, which governs regularization strength. We examine combined forecast accuracy for many λ's, ranging from very light penalization (small λ; all forecasters included in the combination) to very heavy penalization (large λ; no forecasters included in the combination). Specifically, we compute forecasts on a grid of 200 λ's. We start with an equally-spaced grid on [-15, 15], which we then exponentiate, producing a grid on $(0, 3269017]$, with grid coarseness increasing with λ. This grid turns out to be adequate for all LASSO-based combinations that we consider. Throughout we calculate forecast errors using "realizations" from the summer 2014 data vintage.

5.2 Basic Results

We present out-of-sample Combined Forecast RMSE's in Table 1. Several observations are in 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 doesn't happen here.)

3. All methods perform better than the worst individual forecaster.

[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 that is not the case in our forecast combination application.

Table 1: RMSE's Based on Ex-Post Optimal λ's

Regularization Group:	RMSE	λ^*	#
Ridge	1.629	3.098	ALL
LASSO	1.596	0.507	2.9
Egalitarian Ridge	1.612	max	ALL
Egalitarian LASSO	1.612	3.602	ALL
2-Step Egal LASSO (LASSO, Average)	1.482	0.239	3.3
2-Step Egal LASSO (LASSO, Egal Ridge)	1.482	(0.239, max)	3.3
2-Step Egal LASSO (LASSO, Egal LASSO)	1.482	(0.239, 3.098)	3.3
Comparison Group:	RMSE	λ^*	#
Best Individual	1.480	N/A	1
Median Individual	1.643	N/A	1
Worst Individual	1.850	N/A	1
Simple Average	1.612	N/A	ALL

Notes: # denotes the average number of forecasters selected.

4. The simple average improves significantly over the worst individual, but it is still noticeably worse than the best individual. Its RMSE is approximately midway between those of the worst and best individuals.

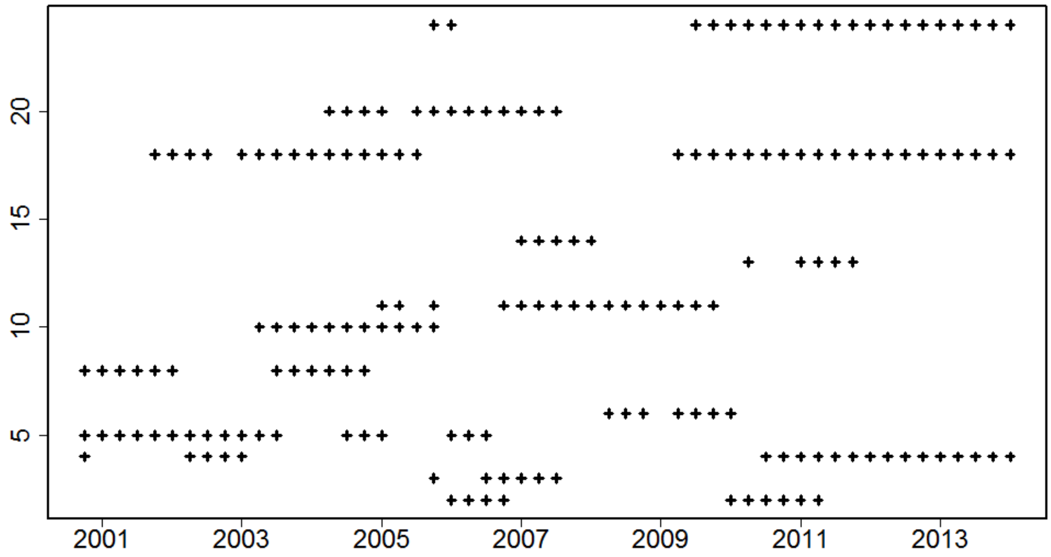
5. All procedures involving selection to zero select a small number of forecasters on average (approximately three).

6. Standard ridge and LASSO perform about as well as the simple average, despite their shrinking toward zero weights rather than equal weights.

7. Egalitarian ridge and egalitarian LASSO 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 egalitarian ridge and egalitarian LASSO produce a simple average.

8. All 2-step methods perform identically. The reason is as follows. They regularize identically in the first step, by construction. Then, in the second step, the "step 2 simple average" method averages by construction, and the remaining 2-step methods effectively average as well in the second step, because heavy step-2 regularization turns

Figure 3: Selected Forecasters



Notes: 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.

out to be optimal.

9. The 2-step methods reduce out-of-sample RMSE relative to the simple average by almost ten percent.

10. The 2-step methods have out-of-sample RMSE as good as that of the best forecaster.

5.3 The Nature and Evolution of the Set of Selected Forecasters

One might wonder which forecasters are selected by our 2-step 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 Figure 3, 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:

1. The optimal pool is usually small, with three or four forecasters, as also mentioned earlier in conjunction with Table 1.

2. The optimal pool is not dominated by any one forecaster, or a small set of forecasters. Many different forecasters move in and out of the optimal pool as we roll through the sample.

3. The optimal pool is usually "democratic" in the sense that it is composed of some forecasters from the best performing group, some from the middle group, and some from the worst.

4. The best individual (ID 1) is not always included in the optimal pool. Indeed the best individual is included in the optimal pool only 11 times.

5. Conversely, the worst forecaster (ID 25) is sometimes included in the optimal pool, mostly toward the end of the sample following the Great Recession. Interestingly, the worst individual is actually included in the optimal pool more often than the best.

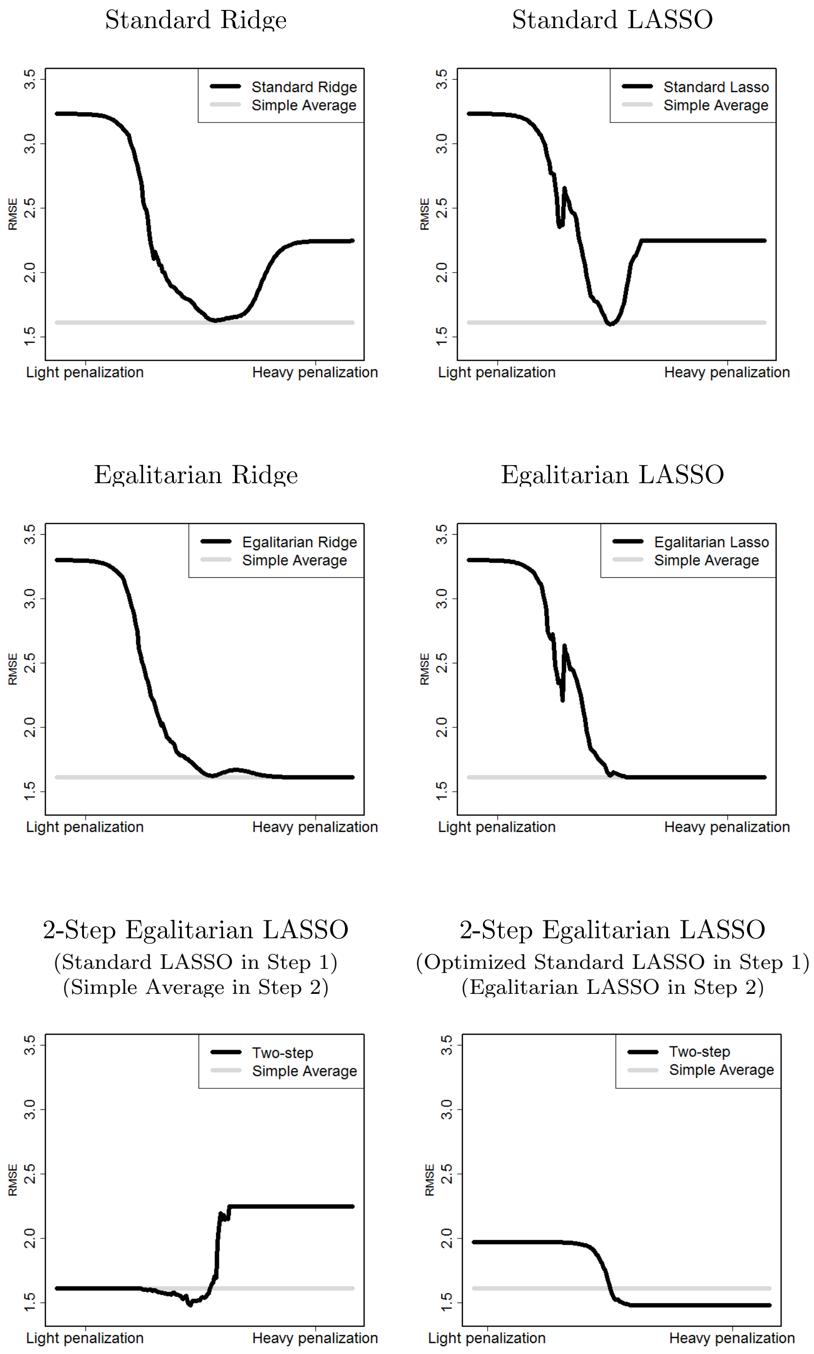
5.4 On the Importance of λ

The results in Table 1 depend on knowledge of the ex post optimal λ. To get a feel for the sensitivity to λ, we show RMSE as function of λ in Figure 4. In each sub-figure, the lighter gray line is the RMSE for simple averaging. Consider first the top row of Figure 4, in which we show 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 from each other and from a simple average. In the limit as penalization increases, however, their performance deteriorates as all forecasters are eventually excluded and the "combined forecasts" therefore approach 0.

Next consider the second row of Figure 4, in which we show egalitarian ridge and egalitarian LASSO. They too perform similarly in terms of the optimized value based on the ex post best λ; at that point they are basically indistinguishable from each other and from a simple average. But their penalization limit is very different. In the limit as penalization increases, egalitarian ridge, egalitarian LASSO, and simple averaging must be (and are) identical.

Now consider the third row of Figure 4, in which we show 2-step egalitarian LASSO, in each case with step 1 done by standard lasso. In the left panel we implement step 2 by simply averaging the step-1 selected forecasters, so there is only one penalty parameter to choose. At the ex post optimum penalty, this 2-step egalitarian LASSO outperforms other methods, including simple averaging of all forecasters.

Figure 4: RMSE as a Function of λ,
Various Forecast Combination Methods



Standard Ridge

Standard LASSO

Egalitarian Ridge

Egalitarian LASSO

2-Step Egalitarian LASSO
(Standard LASSO in Step 1)
(Simple Average in Step 2)

2-Step Egalitarian LASSO
(Optimized Standard LASSO in Step 1)
(Egalitarian LASSO in Step 2)

Notes: In the lower-right panel we implement step 2 by 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 fixed at its optimal value.

In the right panel of Figure 4 we implement step 2 by egalitarian LASSO regression on the forecasters selected in step 1, so there is a second penalty parameter to choose. Call the step-1 and step-2 penalty parameters λ_1 and λ_2, respectively, and denote the ex post optimal pair by $(\lambda_1^*, \lambda_2^*)$. We show RMSE as a function of λ_2, with λ_1 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.

6 Ex Ante Combined Forecast Comparison

All told, it is clear that the superior results for the 2-step procedures are crucially dependent on choosing the "right" λ. It remains to be seen whether we can choose λ successfully in real time. Hence we now provide results for real-time λ selection by cross validation, which is natural and appropriate in our forecasting context.

Before proceeding, however, we note an interesting subtlety. If it is true, as mentioned earlier, that the ex-post results endow the 2-step combinations with some knowledge of the future, which potentially artificially enhances their performance, it is also true that the ex-post results enforce constancy of λ, which potentially *degrades* performance relative to alternative procedures that allow for time-variation, as with real-time cross validation. Hence, perhaps surprisingly, the cross-validated real-time results need not be inferior to the ex-post results.

6.1 Ex Ante Optimal Tuning by Overlapping Cross Validation

Standard "S-fold" cross validation for a sample $t = 1, ..., T$ proceeds as follows. Let $\mathcal{T} = \{1, ..., T\}$, and partition $\mathcal{T}$ into S approximately equally-sized blocks, $\mathcal{T} = \mathcal{T}_1 \cup ... \cup \mathcal{T}_S$. Then for any given λ, we approximate mean squared error by

$$\widehat{MSE}(\lambda) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t^c(\lambda))^2$$

where

$$\hat{y}_t^c(\lambda) = \sum_{i=1}^{K} \hat{\delta}_{i,t}(\lambda) f_{i,t} \tag{5}$$

and $\hat{\delta}_t(\lambda)$ is a regularized estimator (recall that we consider several variants) based on observations excluding those in the partition block that includes t. We select the λ that minimizes

Table 2: RMSE's Based on Real-Time Cross-Validated λ's

Regularization Group:	$B = 1$		$B = 2$	
	RMSE	#	RMSE	#
Ridge	1.518	ALL	1.635	ALL
LASSO	2.090	12.5	2.074	11.1
Egalitarian Ridge	1.801	ALL	1.763	ALL
Egalitarian LASSO	2.075	ALL	1.856	ALL
2-Step Egal LASSO (LASSO, Average)	1.562	2.7	1.599	4.8
2-Step Egal LASSO (LASSO, Egal Ridge)	2.065	13.1	2.115	11.4
2-Step Egal LASSO (LASSO, Egal LASSO)	2.011	14.6	2.066	14.5

Comparison Group:	RMSE	#
Best Individual	1.480	1
Median Individual	1.643	1
Worst Individual	1.850	1
Simple Average	1.612	ALL

Notes: # denotes average number of forecasters selected.

$\widehat{MSE}(\lambda)$; that is, $\lambda^* = \arg\min_\lambda \widehat{MSE}(\lambda)$.

Our "overlapping" cross-validation approach is similar except that we use a different $\hat{\delta}_t(\lambda)$ in (5). Rather than sequentially dropping non-overlapping contiguous blocks $\mathcal{T}_1, ..., \mathcal{T}_S$, we follow Stock and Watson (2012) and sequentially drop *overlapping* blocks. In particular, for each t we obtain $\hat{\delta}_t(\lambda)$ from a sample that omits observation t and those from a symmetric neighborhood around it,

$$\big\{(y_s, f_{1,s}, ..., f_{K,s}) : \ s \in \{1, ..., t-B, \quad t+B, ..., T\}\big\}. \tag{6}$$

$B = 1$ corresponds to standard leave-one-out cross validation (non-overlapping "blocks"), and $B \geq 2$ corresponds to overlapping blocks. In our subsequent empirical work we explore $B = 1$ and $B = 2$. We take $\lambda^* = \arg\min_\lambda \widehat{MSE}(\lambda)$.

6.2 Combined Forecast Performance with Ex Ante Optimal Tuning

As in the earlier ex-post exercise, we roll through the evaluation sample, sequentially estimating (using a 20-quarter rolling window) and forecasting (four quarters ahead). The difference is that we now cross validate λ in real time as part of the rolling estimation, instead of endowing the forecaster with knowledge of the ex post optimal λ.

It is interesting to note the differences between this ex ante rolling cross validation analysis and the ex-post analysis of section 5. The ex ante rolling cross validation analysis both (1) constrains to real-time the information set on which forecasts are based, and (2) allows for adaptive (time-varying) λ^*. In particular, it is possible in principle for the ex-ante approach to perform better than the ex-post approach, despite (1), because of (2).

In Table 2 we collect RMSEs and the number of included forecasters (#) under $B = 1$ and $B = 2$ cross validation.

1. We lose the predictive gains that we obtained when using ex post optimal λ.

2. There are only three cases that outperform the simple average (ridge for $B = 1$, and 2-Step egalitarian LASSO (second step average) for $B = 1$ and $B = 2$).

3. Ridge with $B = 1$ cross-validation actually performs better than that based on the ex post optimal λ. The gain comes from the fact that under cross validation we allow for time-varying λ.

4. All told, the results suggest that it is hard to beat the simple average.

Perhaps, however, there is remaining cause for some optimism, to which we now turn.

6.3 Alternative Approaches to Ex Ante LASSO Tuning

6.3.1 "Average Best"

Here we select a subset of forecasts and average those selected.

"Average Best (Individual-Based) N"

Note that the RMSE performances of the 2-step egalitarian methods in the third row of Figure 4 are almost identical at the optimum, but that there is a point optimum in the bottom left (LASSO in step 1, simple averaging in step 2) and a broad plateau optimum

Table 3: N-Forecast RMSE's
"Average Best" and "Best Average"

	RMSE	avg λ^*	avg N
Average Best N			
Average of (Individual-Based) Best 1	1.537	N/A	1
Average of (Individual-Based) Best 2	**1.496**	N/A	2
Average of (Individual-Based) Best 3	1.500	N/A	3
Average of (Individual-Based) Best 4	1.509	N/A	4
Average of (Individual-Based) Best 5	1.517	N/A	5
Average of (Individual-Based) Best 6	1.530	N/A	6
Average Best N			
Average of (LASSO-Based) Best 1	1.787	5.13	1
Average of (LASSO-Based) Best 2	1.730	2.33	2
Average of (LASSO-Based) Best 3	1.562	1.25	3
Average of (LASSO-Based) Best 4	**1.542**	0.50	4
Average of (LASSO-Based) Best 5	1.575	0.21	5
Average of (LASSO-Based) Best 6	1.582	0.11	6
Best Average N			
Best Average 1	1.537	N/A	1
Best Average 2	1.510	N/A	2
Best Average 3	**1.501**	N/A	3
Best Average 4	1.510	N/A	4
Best Average 5	1.150	N/A	5
Best Average 6	1.524	N/A	6
Best Average $\leq N$			
Best Average $\leq$1	1.537	N/A	1
Best Average $\leq$2	1.514	N/A	1.52
Best Average $\leq$3	**1.509**	N/A	1.87
Best Average $\leq$4	1.510	N/A	2.00
Best Average $\leq$5	1.510	N/A	2.00
Best Average $\leq$6	1.510	N/A	2.00
Comparison Group:	RMSE	λ^*	N
Best Individual	1.480	N/A	1
Median Individual	1.643	N/A	1
Worst Individual	1.850	N/A	1
Simple Average	1.612	N/A	N

in the bottom right (optimized LASSO in step 1, egalitarian LASSO in step 2). At the optimum, the bottom-right procedure effectively selects a few forecasts and averages them. This suggests a strategy of abandoning cross validation and instead simply averaging the best N individual forecasts in each 20-quarter window. We refer to this as "averaging the best (individual-based) N".

"Average Best (LASSO-Based) N"

There is a subtlety, however, when averaging the best individual performers. In general the best forecast combination does *not* simply use the best individual forecasts, as we stressed earlier in section 5.3 when discussing the members and evolution of the set of selected forecasts. This suggests a related but different strategy: In each 20-quarter window, find the LASSO λ that keeps N forecasts, and average *those* forecasts. We refer to this as "averaging the best (LASSO-based) N".

6.3.2 "Best Average"

In the "average best" approach above, at each time we select some best forecasts and average them, moving through 20-quarter windows. Here we flip to a more brute-force "best average" approach, instead selecting directly over averages.

Best Average N

At each time we use a 20-quarter window and determine the best-performing N-forecast average and use it. Note that this method rapidly becomes costly as the number of forecasters increases. With 25 forecasters, for example, finding the best 5 requires computing ${}_{25}C_5$ (= 53,130) simple averages and then sorting them to determine the minimum.

Best Average $\leq N$

At each time we use a 20-quarter window and determine the best-performing $\leq N$-forecast average and use it. The approach is similar to the "subset averaging" of Elliott (2011), but Elliott averages subset averages, whereas we obtain and use the best subset average.

The computational burden of $\leq N$-forecast averaging grows even more quickly with N than it does with N-forecast averaging, because we now consider all subsets. With 25 forecasters, finding the best ≤ 5 requires computing ${}_{25}C_5 +_{25} C_4 + ... +_{25} C_1$ (= 68405) simple averages and then sorting them to determine the minimum. Fortunately, the empirical results to which we now turn indicate that small N is generally optimal for both the average-best and best-average methods.

6.3.3 Empirical Performance

We show the performance of all average-best and best-average methods in Table 3 for $N = 1, ..., 6$. Several remarks are in order:

1. For all variations, the optima are achieved for small N (2, 3, or 4).
2. One might expect best-average methods to perform best, because they directly target the object of interest. Although they do not out-perform, they also do not underperform: they are at least as good as anything else (not worrying about the third decimal place).
3. Best average N requires choosing N, and the results depend on N. In contrast, best average $\leq N$ is fully automatic, requiring only choice of a maximal number of forecasters, which table 3 as well as our earlier results indicates can be safely set to 6, say.
4. Best average ≤ 6 RMSE (1.51) is almost as good as that of the best individual (1.48), much better than that of the median individual (1.64), and importantly, better than that of the simple average (1.61).

7 Conclusions and Directions for Future Research

Against a background of frequently-found superiority of simple-average forecast combinations, we have proposed "egalitarian LASSO" procedures that discard some forecasts and then select and shrink – without forcing – those remaining toward equal weights. We found that simple averages remain highly competitive. That is, although out-of-sample RMSE improvements on simple averages are possible in principle using egalitarian LASSO, they are hard to achieve in small samples in real time, due to difficulty of successfully cross validating the LASSO tuning parameter. We therefore proposed alternative direct combination procedures, most notably "best average" combinations, which do not require choice of a tuning parameter, and we showed that they outperform simple averages in forecasting Eurozone GDP growth.

A key insight in understanding the above progression is that the structure of the best-average procedure is entirely motivated by the lessons learned from egalitarian LASSO. Without the intermediate egalitarian LASSO step, we would likely never have arrived at best-average. For example, we learn from egalitarian LASSO analysis that:

1. The selection penalty should be quite harsh – only a few forecasts need be combined.
2. The forecasts selected for combination should be allowed to change over time.
3. The forecasts selected for combination should be regularized via shrinkage.
4. The shrinkage direction should be toward a simple average, not toward zero, or anything else.
5. The shrinkage should be extreme; that is, the selected forecasts should simply be averaged.
6. Small-sample real-time selection of tuning parameters by cross-validation (or any other method) is very difficult in the context of forecast combination and should be avoided.

Best average ≤ 6 achieves all of this, and many extensions are possible. It would be of interest, for example, to explore the performance of "best average" versions of OSCAR LASSO and the related HORSES procedure of Jang et al. (2013).

References

Aiolfi, M. and A. Timmermann (2006), "Persistence in Forecasting Performance and Conditional Combination Strategies," *Journal of Econometrics*, 135, 31–53.

Aruoba, S.B., F.X. Diebold, J. Nalewaik, F Schorfheide, and D. Song (2012), "Improving GDP Measurement: A Forecast Combination Perspective," In X. Chen and N. Swanson (eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honour of Halbert L. White Jr.*, Springer, 1-26.

Bates, J.M. and C.W.J Granger (1969), "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451–468.

Bondell, H. D. and B. J. Reich (2008), "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR," *Biometrics*, 64, 115–123.

Capistrán, C. and A. Timmermann (2009), "Forecast Combination with Entry and Exit of Experts," *Journal of Business and Economic Statistics*, 27, 428–440.

Chan, Y. L., J. H. Stock, and M. W. Watson (1999), "A Dynamic Factor Model Framework for Forecast Combination," *Spanish Economic Review*, 1, 91–121.

Clemen, R. T. (1989), "Combining Forecasts: A Review and Annotated Bibliography (With Discussion)," *International Journal of Forecasting*, 5, 559–583.

Conflitti, C., C. De Mol, and D. Giannone (2015), "Optimal Combination of Survey Forecasts," *International Journal of Forecasting*, 31, 1096–1103.

Diebold, F. X. and P. Pauly (1990), "The Use of Prior Information in Forecast Combination," *International Journal of Forecasting*, 6, 503–508.

Elliott, G. (2011), "Averaging and the Optimal Combination of Forecasts," Manuscript, Department of Economics, UCSD.

Elliott, G. and A. Timmermann (2016), *Economic Forecasting*, Princeton University Press.

Fuentes, J., P. Poncela, and J. Rodríguez (2014), "Selecting and Combining Experts from Survey Forecasts," *Working Paper 14-09, Statistics and Econometrics Series 05, Universidad Carlos III de Madrid*.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013), "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting*, 29, 108–121.

Granger, C.W.J. and R. Ramanathan (1984), "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3, 197–204.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, Springer.

Jang, W., J. Lim, N.A. Lazar, J.M. Loh, and D. Yu (2013), "Regression Shrinkage and Grouping of Highly Correlated Predictors with HORSES," Working paper, `https://arxiv.org/abs/1302.0256`.

Smith, J. and K. F. Wallis (2009), "A Simple Explanation of the Forecast Combination Puzzle," *Oxford Bulletin of Economics and Statistics*, 71, 331–355.

Stock, J. H. and M. W. Watson (2004), "Combination Forecasts of Output Growth in a Seven-Country Data Set," *Journal of Forecasting*, 23, 405–430.

Stock, J. H. and M. W. Watson (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30, 481–493.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 267–288.

Timmermann, A. (2006), "Forecast Combinations," *Handbook of Economic Forecasting*, 135–196.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions," in P.K. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 233–243.