Partially-Egalitarian Lasso for Shrinkage and Selection in Forecast Combination

Francis X. Diebold, Penn
Minchul Shin, Illinois
Umut Akovali, Koç and Penn

June 14, 2018

Background

Lots of Interesting Predictive Modeling Issues and Ideas

– Economic surveys in Europe and U.S.

– Forecast combination, "ensemble averaging"

– Machine-learning methods
(Selection, shrinkage, regularization, ...)

– Big-data methods for small-data problems

Forecast Combination

$$C_t = \lambda f_{1t} + (1 - \lambda) f_{2t}$$

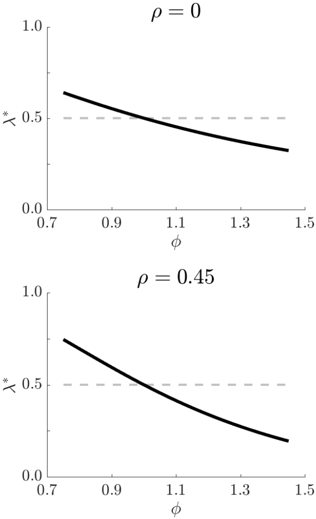
$$\lambda^* = \frac{1 - \phi\rho}{1 + \phi - 2\phi\rho}$$

$$\phi = \frac{var(e_1)}{var(e_2)}$$

$$\rho = corr(e_1, e_2)$$

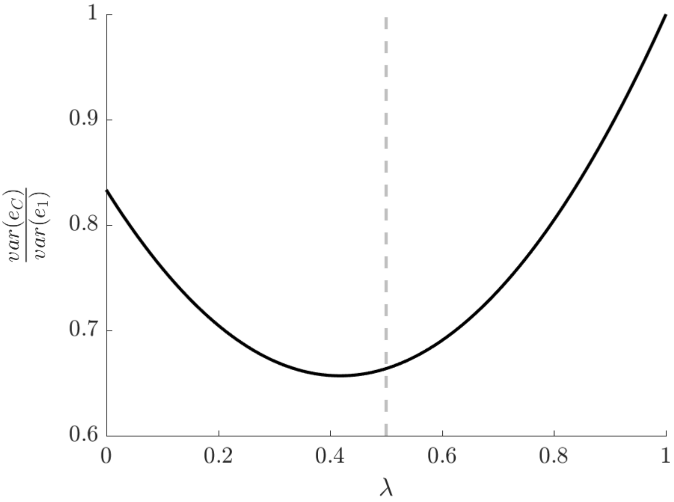
Optimal weights are not equal.

Optimal Combining Weights are Far From 0 and 1 (And Near 1/2)



λ^* vs. ϕ for Various ρ Values.

Gains From Combining Are Huge



$\frac{var(e_C)}{var(e_1)}$ for $\lambda \in [0, 1]$. We set $\phi = 1.20$ and $\rho = 0.45$.

Summary and More

– Large gains from combining

– Optimal combining weights are not equal

– But they're likely not too far from equality

– Estimation issues make equal weights even more attractive

"Can anything beat the simple average?"
(Genre, Kenny, Meyler and Timmermann, 2013)

So we may want to shrink, if not force, weights toward equality...

But We May Want to Trim Before Shrinking

"Trim and average" procedures
have been percolating for many years

(e.g., Stock and Watson, 1999)

But as noted by Granger and Jeon (2004):

"... more of a pragmatic folk-view
than anything based on a clear theory"

We will provide a formal framework and empirical evidence

(and our trimming is sophisticated...)

So:

– First select some weights to 0

"Select to 0"

– Then shrink the survivors' weights toward equality

"Shrink to $1/k$"

Literature

Ancient:

Bayes (1764), ...

Middle Ages:

Bates-Granger (1969), Granger-Ramanathan (1984), ...

Renaissance / Modern:

Diebold-Pauly (1990), Stock-Watson (2004), ...

Post-Modern:

Capistran and Timmermann (2009), Czado, Gneiting, Held (2009), Conflitti, De Mol, and Giannone (2015), Amisano and Geweke (2017), Elliott (2011), ...

Methods:
Penalized Estimation

Penalized Estimation

$$\hat{w} = argmin_w \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} w_i f_{it} \right)^2 \quad s.t. \quad \sum_{i=1}^{K} |w_i|^q \leq c$$

$$\hat{w} = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} w_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} |w_i|^q \right)$$

It's all about q:

Concave penalty function non-differentiable at the origin, encourages selection to 0 (e.g., $q = 1/2$)

Smooth convex penalty, encourages shrinkage toward 0 (e.g., $q = 2$)

$q = 1$ is both concave and convex, encourages both selection to 0 and shrinkage to 0

Penn
University of Pennsylvania

LASSO

$$\hat{w} = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} w_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} |w_i| \right)$$

$$(q = 1)$$

– Selects to 0, shrinks toward 0
No shrinkage ($\lambda \to 0$): Bates-Granger-Ramanathan
Full shrinkage ($\lambda \to \infty$): 0 weights

"Selects in the right direction, shrinks in the wrong direction"

– Can handle situations with $K > T$

Generalized Penalized Estimation and Egailtarian LASSO

Generalized penalized estimation:

$$\hat{w} = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} w_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} |w_i - w_i^0|^q \right)$$

Egalitarian LASSO:

$$\hat{w} = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} w_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left| w_i - \frac{1}{K} \right| \right)$$

$$\left(q = 1, w_i^0 = \tfrac{1}{K} \ \forall i \right)$$

– Selects to $1/K$, shrinks toward $1/K$

No shrinkage ($\lambda \to 0$): Bates-Granger-Ramanathan

Full shrinkage ($\lambda \to \infty$): Equal weights

"Selects in the wrong direction, shrinks in the right direction"

Partially-Egalitarian LASSO

$$\hat{w}_{pLASSO} = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_i w_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} |w_i| \left| w_i - \frac{1}{p(w)} \right| \right),$$

where $p(w)$ is the number of non-zero elements of w.

– Selects to 0, shrinks toward $1/k$

No shrinkage ($\lambda \to 0$): Bates-Granger-Ramanathan

Full shrinkage ($\lambda \to \infty$): Sophisticated trimmed average

"Selects in the right direction, shrinks in the right direction"

Problem: Challenging optimization

Penn
University of Pennsylvania

Two-Step Partially-Egalitarian LASSO

Step 1 (Select to 0): Using standard LASSO, select k forecasts from among the full set of K forecasts.

$$\hat{w}_1 = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} w_i f_{it} \right)^2 + \lambda_1 \sum_{i=1}^{K} |w_i| \right)$$

Step 2 (Shrink/Select to $1/k$): Using egailtarian LASSO, shrink/select the weights on the k survivors toward $1/k$.

$$\hat{w}_2 = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{k} w_i f_{it} \right)^2 + \lambda_2 \sum_{i=1}^{k} \left| w_i - \frac{1}{k} \right| \right)$$

Feasibility

– To make the partially-egalitarian LASSO feasible,
we need a way to select the penalty parameters λ_1 and λ_2

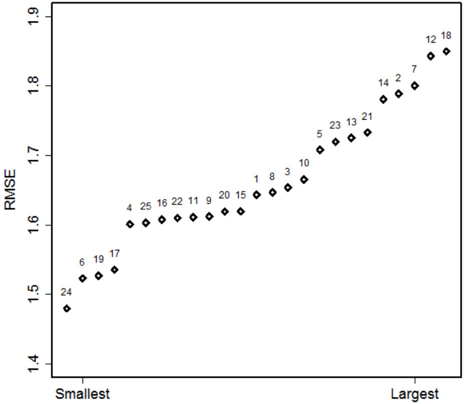
– This is a real issue. We will return to it.

Combining Survey Forecasts

Basic Framework – ECB-SPF

- Euro-area real GDP growth
- Quarterly 1-year-ahead survey of professional forecasts
- 25 forecasters in the pool continuously ($K = 25$)
- 20-quarter rolling estimation window ($T = 20$)
- Errors based on realizations from summer 2014 vintage
- Forecast evaluation period 2000Q4-2014Q1 (54 obs.)

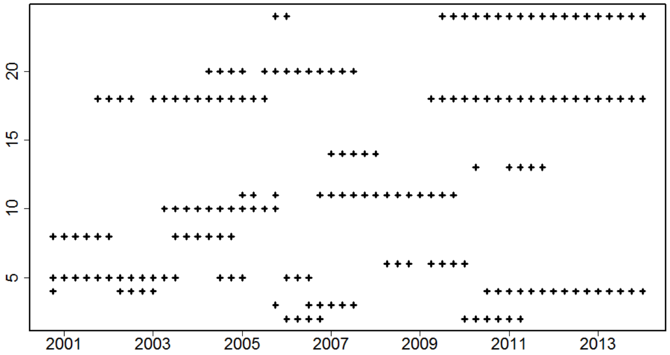
RMSE's of 25 Forecasters – ECB-SPF



Relative RMSE's for pLASSO Combinations ECB-SPF Euro-Zone Real Growth (Infeasible – Based on Ex-Post Optimal λ's)

	Avg #	RMSE/Med	DM	RMSE/Avg	DM
2-Step (Step 2 Average)	3.31	0.91	0.92 (0.18)	0.92	1.15 (0.13)
2-Step (Step 2 eLASSO)	3.31	0.91	0.92 (0.18)	0.92	1.15 (0.13)
Best Individual	1	0.91	0.65 (0.26)	0.92	0.71 (0.24)
Median Individual	1	1.00	N/A	1.00	-0.17 (0.57)
Worst Individual	1	1.14	-1.10 (0.86)	1.15	-1.05 (0.85)
Simple Average	25	1.00	0.17 (0.43)	1.00	N/A

Individual Forecasters Selected (ECB-SPF)



– Selected sets are small
– Serial correlation in selected individuals
– Best individual not always included in the selected set
– Worst individual sometimes included in the selected set

Broad Lessons So Far

Substantial (Ex Post) Gains From Partially-Egalitarian LASSO

Selection:

- Selection penalty should be harsh so selected set is small ($k \approx 3$)
- Selected set evolves gradually over time

Shrinkage:

- Selected forecasts should be shrunken toward a simple average
- Shrinkage penalty should be harsh, so that forecasts are simply averaged

Making Partially-Egalitarian LASSO Feasible

- Optimal LASSO penalties λ_1 and λ_2 are unknown ex ante and must be estimated in real time
- Standard "leave-one-out" cross validation performs poorly (No surprise: small samples, serially-correlated data, ...)
- But the structure of our earlier infeasible solution holds the key...

Direct Averaging Approaches

- "Average-Best N"

  – Average the recently best-performing N forecasters

  – Computationally simple. But there is an issue of how to define "recently best-performing". We want sophisticated trimming.

- $\boxed{\text{"Best } N\text{-Average"}}$

  – Examine all ${}_{25}C_N$ N-averages.
  Use the recently best-performing.

  – Computationally more burdensome, but still simple.
  No issue of how to define "recently best-performing".
  Sophisticated trimming automatically embedded.

- Simple extensions:

  – Average-best $\leq N_{max}$

  – $\boxed{\text{Best} \leq N_{max}\text{-average}}$

Penn
University of Pennsylvania

Best N-Average Combinations ECB-SPF Euro-Zone Real Growth

	Avg #	RMSE/Med	DM	RMSE/Avg	DM
$N = 1$	1	0.95	0.49 (0.31)	0.95	0.56 (0.29)
$N = 2$	2	0.93	0.68 (0.25)	0.94	0.81 (0.21)
$N = 3$	3	0.93	0.77 (0.22)	0.93	0.91 (0.18)
$N = 4$	4	0.93	0.80 (0.21)	0.94	0.97 (0.17)
$N = 5$	5	0.94	0.88 (0.19)	0.94	1.11 (0.14)
$N = 6$	6	0.94	0.90 (0.19)	0.95	1.20 (0.12)
Best Individual	1	0.91	0.65 (0.26)	0.92	0.71 (0.24)
Median Individual	1	1.00	N/A	1.00	-0.17 (0.57)
Worst Individual	1	1.14	-1.10 (0.86)	1.15	-1.05 (0.85)
Simple Average	25	1.00	0.17 (0.43)	1.00	N/A

Best $\leq N_{max}$-Average Combinations
ECB-SPF Euro-Zone Real Growth

	Avg #	RMSE/Med	DM	RMSE/Avg	DM
$N_{max} = 1$	1.00	0.95	0.49 (0.31)	0.95	0.56 (0.29)
$N_{max} = 2$	1.52	0.93	0.67 (0.25)	0.94	0.79 (0.22)
$N_{max} = 3$	1.87	0.93	0.71 (0.24)	0.94	0.84 (0.20)
$N_{max} = 4$	2.00	0.93	0.70 (0.24)	0.94	0.83 (0.21)
$N_{max} = 5$	2.00	0.93	0.70 (0.24)	0.94	0.83 (0.21)
$N_{max} = 6$	2.00	0.93	0.70 (0.24)	0.94	0.83 (0.21)
Best Individual	1	0.91	0.65 (0.26)	0.92	0.71 (0.24)
Median Individual	1	1.00	N/A	1.00	-0.17 (0.57)
Worst Individual	1	1.14	-1.10 (0.86)	1.15	-1.05 (0.85)
Simple Average	25	1.00	0.17 (0.43)	1.00	N/A

More

- Different window widths
- Variable window widths
  $W_t \in \{W_1, W_2, ..., W_m\}$
  e.g., $W_t \in \{4, 8, 12, 16, 20, 24, 28, 32, 36\}$
- g-group clustering: $C_t = w_1 \bar{f}_1 + w_2 \bar{f}_2 + ... + w_g \bar{f}_g$
  e.g., two groups: $C_t = .75\bar{f}_1 + .25\bar{f}_2$
- Other regions (U.S.)
- DENSITY FORECASTS ...

Penn
University of Pennsylvania

Log Probability Score for a Single Density Forecast

$$LPS_i = -\sum_{t=1}^{T} \log p_{it}(y_t)$$

where:

p_{it} is the time-t forecast

y_t is the time-t realization

T is the number of periods

– (Negative of) predictive (log) likelihood

– Minimizing *LPS* analogous to minimizing *SSE* for a point forecast

Log Probability Score For a Mixture Density Forecast

$$LPS(w) = -\sum_{t=1}^{T} \log p_t(y_t)$$

where:

$p_t = \sum_{i=1}^{K} w_i p_{it}$ is the time-t mixture forecast

w_i is the mixture weight on density forecaster i

K is the number of individual forecasters

y_t is the time-t realization

T is the number of periods

Amisano and Geweke (2017, *REStat*)

A Problem with LPS...

SPF density forecasts can look like this:

$$p(y \in I_j) = \begin{cases} 0 & y \in (-\infty, 0] \\ 0 & y \in (0, 0.5] \\ 0 & y \in (0.5, 1.0] \\ 0 & y \in (1.0, 1.5] \\ 0.3 & y \in (1.5, 2.0] \\ 0.5 & y \in (2.0, 2.5] \\ 0.2 & y \in (2.5, 3.0] \\ 0 & y \in (3.0, 3.5] \\ 0 & y \in (3.5, 4.0] \\ 0 & y \in (4.0, \infty] \end{cases}$$

Consider a realization $y = 1.2$.

Then $LPS = \infty$.

Ranked Probability Score
For a Single Density Forecast

$$RPS_i = \sum_{t=1}^{T} \left(\sum_{j=1}^{J} \left\{ P_{ijt} - 1(y_t \leq b_j) \right\}^2 \right)$$

where:

$P_{ijt} = \sum_{h=1}^{j} p_{it}(I_h)$ is the cdf of density forecast p_{it} defined on intervals $I_j = [a_j, b_j]$, $j = 1, ..., J$

Czado, Gneiting and Held (2009, *Biometrics*)

Ranked Probability Score
For a Mixture Density Forecast

$$RPS(w) = \sum_{t=1}^{T} \left(\sum_{j=1}^{J} \left\{ P_{jt} - 1(y_t \leq b_j) \right\}^2 \right)$$

where:

$P_{jt} = \sum_{h=1}^{j} p_t(I_h)$ is the cdf of density forecast p_t
defined on intervals $I_j = [a_j, b_j]$, $j = 1, ..., J$

$p_t = \sum_{i=1}^{K} w_i p_{it}$ is the time-t mixture forecast

w_i is the mixture weight on density forecast i

K is the number of individual forecasts

Partially-Egalitarian LASSO for Density Forecasts

Recall partially-egalitarian LASSO for combining point forecasts:

$$\hat{w}_{pLASSO} = argmin_w \Big(SSE(w) + Penalty_{pLASSO}(w) \Big)$$

Now, for combining density forecasts:

$$\hat{w}_{pLASSO_{LPS}} = argmin_w \Big(LPS(w) + Penalty_{pLASSO}(w) \Big)$$

$$\hat{w}_{pLASSO_{RPS}} = argmin_w \Big(RPS(w) + Penalty_{pLASSO}(w) \Big)$$

Partially-Egalitarian LASSO for Density Forecasts

Recall partially-egalitarian LASSO for combining point forecasts:

$$\hat{w}_{pLASSO} = argmin_w \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} w_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left| w_i \right| \left| w_i - \frac{1}{p(w)} \right| \right)$$

Now, for combining density forecasts:

$$\hat{w}_{pLASSO_{LPS}} = argmin_w \left(- \sum_{t=1}^{T} \log p_t(y_t) + \lambda \sum_{i=1}^{K} \left| w_i \right| \left| w_i - \frac{1}{p(w)} \right| \right)$$

$$\hat{w}_{pLASSO_{RPS}} = argmin_w \left(\sum_{t=1}^{T} \left(\sum_{j=1}^{J} \left\{ P_{jt} - 1(y_t \leq b_j) \right\}^2 \right) + \lambda \sum_{i=1}^{K} \left| w_i \right| \left| w_i - \frac{1}{p(w)} \right| \right)$$

Framework: ECB-SPF

- Quarterly 1-year-ahead density forecasts for Euro-area real GDP growth
- 17 forecasters in the pool continuously
  (as opposed to 25 in DS point prediction application)
- 20-quarter rolling estimation window
- Realizations based on the 2017/11/18 vintage
- Sample period
  - Survey dates: 1999Q1 – 2017Q1 (73 obs.)
  - Target dates: 1999Q3 – 2017Q3 (73 obs.)
  - For example, in the 1991Q1 survey forecasters were asked to generate predictions (point, density) for real GDP growth between 1998Q4–1999Q3. This is because by the time 1999Q1 survey was conducted, real GDP data were available up to 1998Q4.

Best N-Mixture
ECB-SPF Euro-Zone Real Growth

	Avg #	RPS/Med	DM	RPS/Avg	DM
$N = 1$	1	0.92	-0.94(0.17)	1.02	0.28(0.61)
$N = 2$	2	0.86	-1.8(0.04)	0.95	-0.9(0.18)
$N = 3$	3	0.87	-1.93(0.03)	0.96	-0.92(0.18)
$N = 4$	4	0.86	-2.17(0.02)	0.96	-1.02(0.15)
$N = 5$	5	0.87	-2.21(0.01)	0.96	-1.18(0.12)
$N = 6$	6	0.87	-2.22(0.01)	0.96	-1.41(0.08)
Best Individual	1	0.87	-1.55(0.06)	0.97	-0.84(0.20)
Median Individual	1	1.00	NA	1.11	1.71(0.96)
Worst Individual	1	1.27	2.64(0.99)	1.41	3.30(0.99)
Equal-Weight Mixture (Avg)	17	0.90	-1.71(0.04)	1.00	NA

Best $\leq N_{max}$-Mixture ECB-SPF Euro-Zone Real Growth

	Avg #	RPS/Med	DM	RPS/Avg	DM
$N_{max} = 1$	1.00	0.92	-0.94(0.17)	1.02	0.28(0.61)
$N_{max} = 2$	1.56	0.76	-2.60(0.01)	0.85	-2.12(0.02)
$N_{max} = 3$	2.03	0.73	-2.94(0.01)	0.81	-2.67(0.01)
$N_{max} = 4$	2.49	0.71	-3.19(0.01)	0.79	-2.96(0.01)
$N_{max} = 5$	2.85	0.69	-3.39(0.01)	0.77	-3.37(0.01)
$N_{max} = 6$	3.04	0.68	-3.54(0.01)	0.76	-3.58(0.01)
Best Individual	1	0.87	-1.55(0.06)	0.97	-0.84(0.20)
Median Individual	1	1.00	NA	1.11	1.71(0.96)
Worst Individual	1	1.27	2.64(0.99)	1.41	3.30(0.99)
Equal-Weight Mixture (Avg)	17	0.90	-1.71(0.04)	1.00	NA