"Comparing Possibly Misspecified Forecasts" by Patton

Minchul Shin
University of Pennsylvania

November 17, 2014

Background

Forecast users ask forecast providers to generate forecasts about some economic variables

- Forecast survey asks future state about economic variable.
- Sometimes survey specifies **"forecasting target"** explicitly (mean, quantile, distribution). Sometimes not.
- It is rare that survey asks the **loss function** that forecast providers used.

Background: Consistent loss function

- A loss function is said to be **"consistent"** for a given statistical functional (mean, median, etc.), if the expected loss is minimized when the given function is used as the forecast.

Background: Consistent loss function

- A loss function is said to be **"consistent"** for a given statistical functional (mean, median, etc.), if the expected loss is minimized when the given function is used as the forecast.

- Quick question:
  - How many loss functions are consistent for mean?
  - In other words, how many $L(\cdot, \cdot)$ that satisfies

$$E[y_t] = \arg\min_{\widehat{y}} E_{\widehat{P}}\left[L(y_t, \widehat{y})\right]$$

  - We know that the quadratic loss function $L(y_t, \widehat{y}) = (y_t - \widehat{y})^2$ satisfies above.

Background: Consistent loss function

- A loss function is said to be **"consistent"** for a given statistical functional (mean, median, etc.), if the expected loss is minimized when the given function is used as the forecast.

- Quick question:
  - How many loss functions are consistent for mean?
  - In other words, how many $L(\cdot, \cdot)$ that satisfies

$$E[y_t] = \arg\min_{\widehat{y}} E_{\widehat{P}}[L(y_t, \widehat{y})]$$

  - We know that the quadratic loss function $L(y_t, \widehat{y}) = (y_t - \widehat{y})^2$ satisfies above.

- Answer: Infinitely many.
- The class of loss functions that is consistent for the mean is known as the **Bregman class of loss functions** (Savage, 1971, Banerjee, et al., 2005, and Bregman, 1967)

Example

Set a forecasting target as “conditional mean”.

- Forecast user asks professionals to provide “conditional mean”
- Then, individual professionals will use their own loss function that is consistent with the conditional mean. (any loss function in a set of Bregman loss functions)
- Forecast user collects those individual forecasts (**note: individual forecasts are possibly based on different loss functions**).

Forecast user needs to pick the best (or rank them) among forecasts from professionals based on his loss function not based on professionals’ loss functions.

Problem: Forecast user’s loss function can be different from professionals’ loss.

Contribution of the paper

Patton's question: "Does ranking based on forecast user's loss function (MSE) agree with ranking based on forecast provider's loss function with Bregman loss functions?"

- If we can consistently rank them based on some criterion, say MSE, forecast user does not need to worry about the fact that professionals use different loss function.
- Then, asking for the forecasting target (mean, quantile,...) is sufficient.

Contribution of the paper

Patton's question: "Does ranking based on forecast user's loss function (MSE) agree with ranking based on forecast provider's loss function with Bregman loss functions?"

- If we can consistently rank them based on some criterion, say MSE, forecast user does not need to worry about the fact that professionals use different loss function.
- Then, asking for the forecasting target (mean, quantile,...) is sufficient.

Patton's answer: Not in general.

- He shows under what conditions the ranking by MSE is sufficient.
- These conditions are very restrictive.

Contribution of the paper

Patton's question: "Does ranking based on forecast user's loss function (MSE) agree with ranking based on forecast provider's loss function with Bregman loss functions?"

- If we can consistently rank them based on some criterion, say MSE, forecast user does not need to worry about the fact that professionals use different loss function.
- Then, asking for the forecasting target (mean, quantile,...) is sufficient.

Patton's answer: Not in general.

- He shows under what conditions the ranking by MSE is sufficient.
- These conditions are very restrictive.

Patton's suggestion: *"Best practice for point forecasting is to declare the specific loss function that will be used to evaluate forecasts, and to make that loss function consistent for the target functional of interest to the forecast user."*

He provides similar results for

1) quantile (including median) forecast
2) density forecast

Mean forecasts and Bregman loss functions

Elements of the **Bregman class** of loss functions, denoted $\mathscr{L}_{Bregman}$, take the form:

$$L(y, \hat{y}) = \phi(y) - \phi(\hat{y}) - \phi'(\hat{y})(y - \hat{y})$$

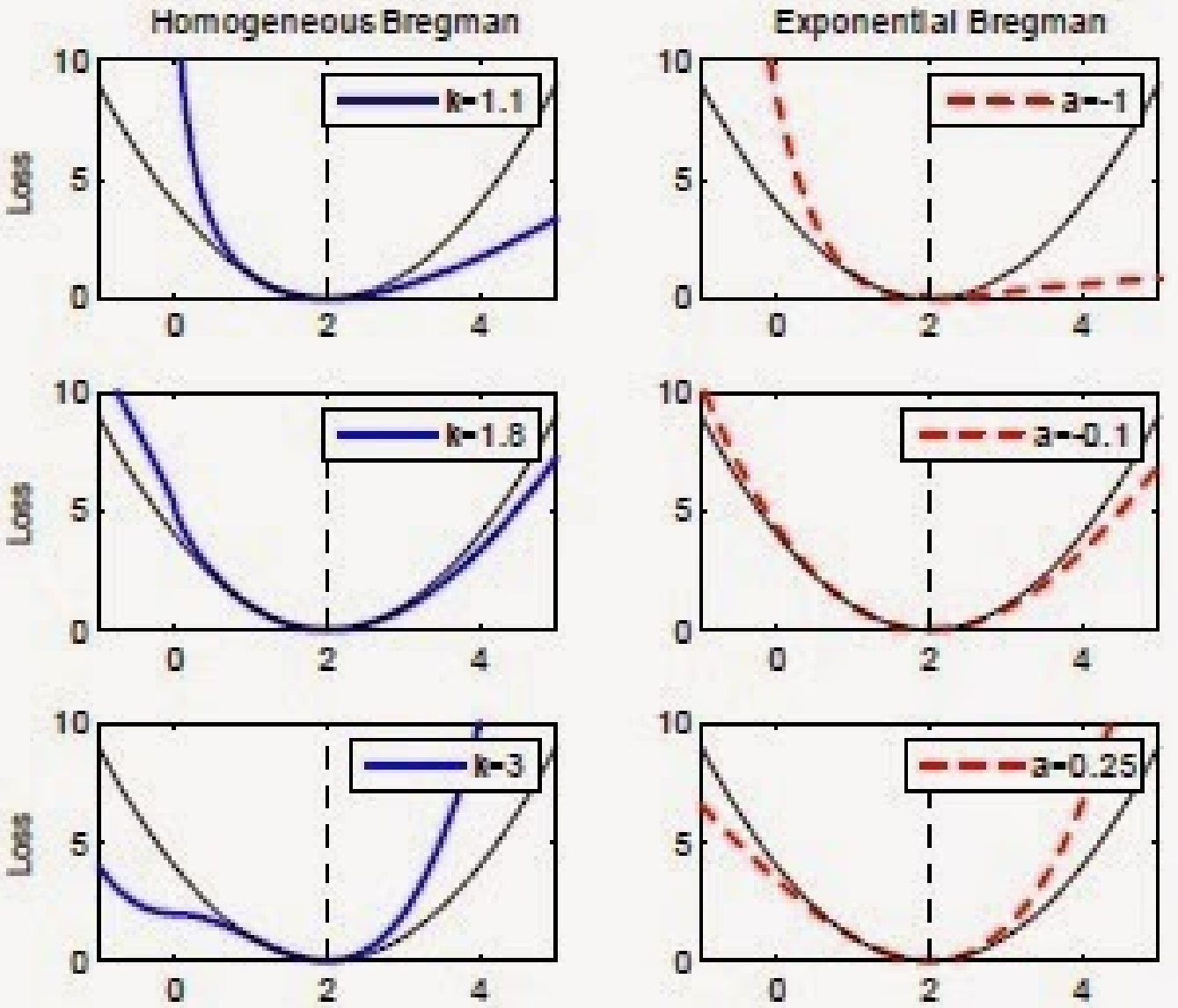
where $\phi : \mathscr{Y} \to R$ is any strictly convex function, and $\mathscr{Y}$ is the support of y. Moreover, this class of loss functions is also **necessary** for conditional mean forecasts.

Note that

- Bregman loss functions do not have to be symmetric.
- Bregman optimal forecasts are the conditional mean. This implies that under correct specification, the optimal forecast is unbiased,

$$E[\widehat{y}] = E\Big[E[y|\mathscr{F}]\Big] = E[y], \quad \text{for any } \mathscr{F}$$

Figure: Various Bregman loss functions



Something fishy?

At first glance, it was quite surprising to me. Why?

In Econ 705/706, we were always told that the conditional mean is optimal under the quadratic loss function. (seldom talk about the opposite direction)

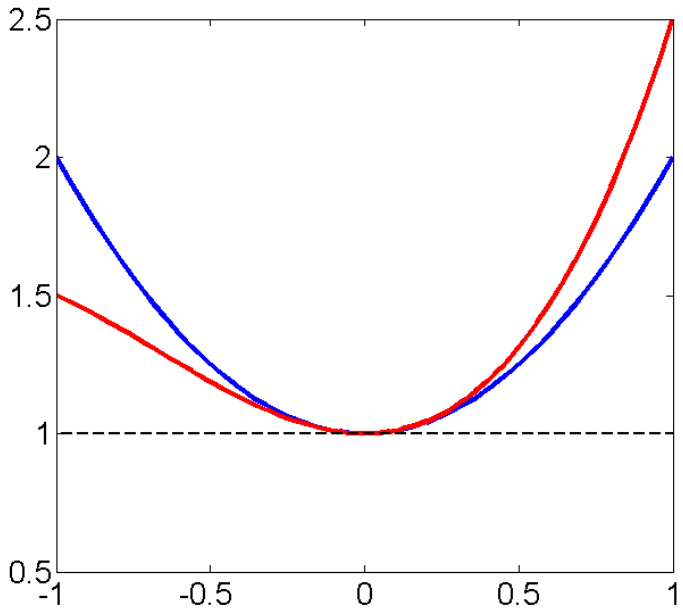
Our instinct says that

- To get a mid-point as an optimizer, something should be symmetric. (like quadratic function)
- But the loss function can be asymmetric.
- Moreover, the distribution of y can be asymmetric.

Expected loss function over $\widehat{y}$

The following figure presents expected losses based on **two** different loss functions,

$$E_{\widehat{P}}[L(y, \widehat{y})]$$



- Blue is symmetric and Red is asymmetric. But minimizer is the same.
- What really matters to get the conditional mean as a loss minimizer is whether the loss minimizer is at $E[y]$ or not (local property)

Example continued

Optimal forecast is a minimizer of the following objective function

$$E[L(y,\hat{y})] = \int L(y,\hat{y})\, f(y|\mathscr{F})\, dy.$$

We will get the conditional mean as an optimal forecast (expected loss minimizer) as long as the first order condition looks like

$$FOC: \quad c(\hat{y})(E[y|\mathscr{F}] - \hat{y}) = 0$$

where $c(\cdot)$ is a function returns non-zero number.

Doesn't matter whether the loss function is symmetric or not.

Recall the form of Bregman loss function,

$$L(y, \hat{y}) = \phi(y) - \phi(\hat{y}) - \phi'(\hat{y})(y - \hat{y})$$

Optimal forecast is a minimizer of the following objective function

$$E[L(y, \hat{y})] = \int L(y, \hat{y})\, f(y|\mathscr{F})\, dy.$$

Derivative with respect to $\hat{y}$ is

$$\begin{aligned}
\frac{\partial}{\partial \hat{y}} E[L(y, \hat{y})] &= \frac{\partial}{\partial \hat{y}} \int L(y, \hat{y})\, f(y|\mathscr{F})\, dy \\
&= \int \frac{\partial}{\partial \hat{y}} L(y, \hat{y})\, f(y|\mathscr{F})\, dy \\
&= -\phi''(\hat{y})(E[y|\mathscr{F}] - \hat{y})
\end{aligned}$$

Hence, the first order condition reads

$$-\phi''(\hat{y})(E[y|\mathscr{F}] - \hat{y}) = 0$$

This is what we need!

Remarks

Bregman class,

$$L(y, \hat{y}) = \phi(y) - \phi(\hat{y}) - \phi'(\hat{y})(y - \hat{y})$$

- For loss functions of the form $L(e)$, the quadratic loss function is the only function in the Bregman class (Savage, 1971).
- Remark 1 (State-dependency)
  - $L(e)$ is not state dependent.
  - $L(y, \widehat{y})$ can be state (y) dependent (I feel more bad for wrong weather prediction if today is rainy day).
  - The asymmetry only comes through this type of state-dependency.
- Remark 2 (Calibrating a parametric forecasting model)
  - Correctly specified models: We get a consistent (Bregman) estimator for any $L(\cdot)$.
  - Misspecified models: Different Bregman loss function estimators yield different probability limits.
  - More on this issue in section 2.4.

Proposition 1

Assume that

(i) The information sets of the two forecasters are nested (and their model is correctly specified)

(ii) Forecasts A and B are optimal under some Bregman loss function.

Then the ranking of these forecasts by MSE is sufficient for their ranking by any Bregman loss function.

Proposition 1

Assume that

(i) The information sets of the two forecasters are nested (and their model is correctly specified)

(ii) Forecasts A and B are optimal under some Bregman loss function.

Then the ranking of these forecasts by MSE is sufficient for their ranking by any Bregman loss function.

Proof by words

- Under this environment, the only way to get different forecasts, $A \neq B$, is to have different predictive distributions. (otherwise, they will give the same answer which is conditional mean of the predictive distribution)
- Under this environment, the only way to get different predictive distribution is to have different information sets.
- Due to the nested information set assumption, if two predictive distributions are different, then it must be the case that

$$\mathscr{F}_A \subset \mathscr{F}_B \iff var(\hat{y}|\mathscr{F}_A) > var(\hat{y}|\mathscr{F}_B)$$

- Hence, MSE comparison is sufficient.

Proposition 2

Assume that

1) The information sets of the two forecasters are non-nested.
2) Or, at least one of the forecasts is based on a misspecified model.
3) Forecasts A and B are optimal under some Bregman loss function

Then the ranking of these forecasts is, in general, sensitive to the choice of Bregman loss function.

Conclusion: *"Best practice for point forecasting is to declare the specific loss function that will be used to evaluate forecasts, and to make that loss function consistent for the target functional of interest to the forecast user."*

My comments/discussions I

Implications for forecasting combination: The paper is manly about evaluating forecasts. But results also have implications for forecast combination.

- For example, Proposition 1 implies that the MSE-optimal (linearly-) combined forecast is also optimal in terms of any Bregman loss function, so longs as the individual forecasts satisfy your Proposition 1 assumptions.
- Related, it seems that ranking any linear combination of Bregman optimal conditional mean forecasts by MSE is sufficient for their ranking by any Bregman loss function.

My comments/discussions II

Bregman-loss-robust forecasts?

- The paper suggests a good practice for forecast users ("tell your loss function to forecast providers"). This is a forecast user's problem.
- Now, suppose the following situation (forecast provider's problem):
  - Forecast provider is asked to provide a conditional mean.
  - He knows that it might not be optimal for the forecast user if he just optimizes his loss function.
  - But, he wants to provide good forecasts for the forecast user without knowing user's loss function.
- What should the forecaster provide? In other words, what would be the best rule for the forecast provider when he doesn't know forecast users' loss function?
- To answer this question, we need to define a concept of "best" (it should be some kind of robustness to misspecification of loss functions, models, etc).

Last comments

It is a very insightful paper.

- I recommend you all to read the paper.
- I didn't cover results for density forecasts and quantile forecasts.
- Similar results (and wisdom) can be drawn for density forecasts and quantile forecasts.

The End

Quantile forecasts

There is a Bregman type class of loss functions for quantile forecasts.

The loss function that is necessary and sufficient for quantile forecasts is called a **"generalized piecewise linear" (GPS) loss** function, denoted by $\mathscr{L}^{\alpha}_{GPL}$:

$$L(y, \hat{y}; \alpha) = (1\{y \leq \hat{y}\} - \alpha)(g(\hat{y}) - g(y))$$

where g is a nondecreasing function, and $\alpha \in (0, 1)$ indicates the quantile of interest.

- It is lin-lin loss function when the function g is identity function, $g(x) = x$.

Andrew shows the same result as Prop 1 and Prop 2 for this class of loss functions. (Prop 4 and Prop 5)

Density forecasts

There is a Bregman type class of loss functions for density forecasts.

- As we know, it is called “proper scoring rule”.
- Gneiting and Raftery (2007) show that if L is a proper scoring rule then it must be of the form:

$$L(F, y) = \Psi(F) + \Psi^*(F, y) - \int \Psi^*(F, y) dF(y)$$

where Ψ is a convex, real-valued function, and Ψ^* is a subtangent of Ψ at the point $F \in \mathscr{P}$.

Andrew shows the same result as Prop 1 and Prop 2 for this class of loss functions (Prop 7 and Prop 8).

Mean forecasts of symmetric random variables

Previous results for conditional mean (prop 1 and prop 2) can be extended to any convex combination of a Bregman and a $GPL^{1/2}$ loss function,

$$\mathscr{L}_{Breg \times GPL} = \lambda \mathscr{L}_{Bregman} + (1-\lambda)\mathscr{L}_{GPL}^{1/2}, \quad \lambda \in [0,1]$$

when forecasters optimize forecast with respect to **a symmetric continuous distribution**.

Note that this is expected result because under a symmetric continuous distribution, median and mean are the same.