Empirical Bayes Quantile-Prediction
aka E-B Prediction under Check-loss;

Lawrence D. Brown
Wharton School, Univ. of Pennsylvania

BFF4, May 2, 2017

Joint work with
Gourab Mukherjee and Paat Rusmevichientong

Preprint available on ArXiv.

NOTE: In several places the symbol \( \nu \) should read \( \nu^2 \). This correction should be clear from the context.
Multiple Independent Gaussian Problems

• $n$ independent problems, indexed by $i = 1, \ldots, n$

• Observe $X_i \sim N\left(\theta_i, \nu_{p,i}^2 \right)$

  $[\theta_i$ unknown; $\nu_{p,i}$ known$]$

  [In applications, $X_i$ might be $X_i = \bar{X}_i$. with]

  $X_{ik} \sim \text{iid } N\left(\theta_i, \sigma_i^2 \right), \ k = 1, \ldots, K_i, \ \nu_{p,i}^2 = \sigma_i^2 / K_i$.]
Quantile Prediction

• For each $i \exists$ potential future observation $Y_i \sim N(\theta_i, \nu_{F,i})$.
  [Again this might be average of several iid var’s.]
• Note that for each $i$, the mean $\theta_i$ is same for both Past and Future.
• Fix $b_i \in (0,1), i=1,\ldots,n$. Goal is to predict the $b_i$-th quantile of dist. of $Y_i$ for each index, $i$. This quantile is
  $$q_{0,i} = \theta_i + \sqrt{\nu_{F,i}} \Phi^{-1}(b_i).$$
• The naïve coordinate-wise predictor is
  $$\hat{q}_{0,i} = X_i + \sqrt{\nu_{F,i}} \Phi^{-1}(b_i).$$

NOTE: This is formal Bayes for uniform prior
Shrinkage Estimation

- Generally beneficial in multi-mean problems (and many other multivariate settings).
- But the shrinkage needs to be properly implemented.
- For homoscedastic problems this is well understood. (Stein estimation)
- But the current problem is quite heteroskedastic: \( v_{p,i}, v_{F,i}, b_i \) can all depend on \( i \).
- For such problems minimax shrinkage may dramatically differ from Empirical Bayes shrinkage;
- And, well-implemented E-B shrinkage is usually preferable

Parametric Empirical Bayes

• We follow the hierarchical conjugate-prior paradigm
  Efron and Morris (op. cit.), Stein (1962), Lindley (1962), Lindley and Smith (1972) ...

\[ \theta_1, \ldots, \theta_n \sim_{iid} N(\eta, \tau^2) \]

• \( \eta, \tau^2 \) are unknown hyper-parameters to be estimated from the data.
• Let \( \hat{\eta}, \hat{\tau}^2 \) denote the estimates.

For simplicity and notational simplicity:
• Consider (for the talk) the case where \( \eta = 0 \) is known.
  (Paper allows unknown \( \eta \) and estimates it along with \( \tau \).)
Quantile Prediction for Known Hyper-parameters

- If $\tau$ is known then posterior distribution of $\tilde{\theta}$ known (& Gaussian)
- Yields known, Gaussian predictive distribution for each future $Y_i$: Let $\alpha_i = \frac{\tau^2}{(\tau^2 + v_{p,i})}$, then
  \[ F_{Y_i}(y | \tau^2, X_i) = N\left(\alpha_i X_i, v_{F,i} + \alpha_i v_{p,i}\right). \]
  
  So quantile prediction is
  \[ \hat{q}(X_i; \tau^2) = \alpha_i X_i + \left(v_{F,i} + \alpha_i v_{p,i}\right) \Phi^{-1}(b_i). \]

- The role of the E-B prior structure is only to motivate this family of quantile predictors.
“Direct” E-B quantile prediction
• Could use data + any plausible estimate of $\tau^2$, the hyper-parameter(s), and plug in.
• i.e., could use marginal MLE $\hat{\tau}^2$ and get prediction
  $$q(X_i; \hat{\tau}^2).$$
• Or could similarly use M of M estimate instead of MLE.
• These (and other) plausible estimates of $\tau^2$ are different & give different predictors.
• They’re not “bad”.
• BUT none of these needs to be the best choice.
Formulation with Loss Function

To investigate optimal choice of $\hat{\tau}^2$ in $q\left(X_i;\hat{\tau}^2\right)$ impose a \textbf{predictive loss function} under which $\hat{q}_\tau = q\left(X_i;\tau^2\right)$ is the natural predictor of $q$ when $\tau^2$ is known.

Then find ‘best’ (or, just ‘good’) $\hat{\tau}^2 = \hat{\tau}^2\left(\bar{X}\right)$ under this loss. Use $\hat{q} = q\left(X_i;\hat{\tau}^2\right)$ for each $i$.

- \textbf{Check-Loss} \textit{(aka, pinball loss or quantile loss)}. Let $b,h > 0$; normalize by $b + h = 1$ (w.l.o.g).

\begin{equation*}
i\text{-th component of predictive loss is } \ell_i\left(Y_i,q\right) = b_i\left(q - Y_i\right)^+ + h_i\left(Y_i - q\right)^+\end{equation*}
Note that $b_i, h_i$ are known and allowed to depend on $i$.
This loss has been introduced here as a device to evaluate potential Quantile-predictors. But it has motivation in various applications. For example:
The Classical News Vendor Problem

Consider Only One Product

Demand distribution for the product: \( X \sim f \)

Stock Inventory quantity: \( \hat{q}(X) \)

Lost Sales Cost: \( b \times (X - \hat{q}(X))^+ \)

Inventory Cost: \( h \times (\hat{q}(X) - X)^+ \)

Total Cost:

\[
 b (X - \hat{q}(X))^+ + h (\hat{q}(X) - X)^+
\]

Cost asymmetric in:
- Over-estimation
- Under-estimation

Typically: \( b >> h \) for non-perishable commodities hence it is a highly non-symmetric loss function.

\( b \): per unit lost sales/ reputation cost

\( h \): per unit storage cost/ depreciation
“Improved” E-B quantile prediction

Conceptual Idea

• Define average predictive risk as

$$\overline{R}(\hat{\theta}, \tilde{q}) = n^{-1} \sum E_{\theta_i} \left( \ell_i \left( Y_i, q_i \left( \tilde{X} \right) \right) \right).$$

• When using a particular hyper-parameter estimator, \( \tilde{\tau}^2 = \tilde{\tau}^2 \left( \tilde{X} \right) \), this is

$$\overline{R}(\hat{\theta}; \tilde{\tau}^2) = n^{-1} \sum E_{\theta_i} \left( \ell_i \left( Y_i, q_i \left( X_i; \tilde{\tau}^2 \left( \tilde{X} \right) \right) \right) \right).$$

• Goal is to create a good/best estimator \( \tilde{\tau} \).
• **KEY** is to find a good and (nearly) unbiased estimator of
\[ n^{-1} \sum E_{\theta_i} \left( \ell_i \left( Y_i, q_i \left( X_i, \tau^2 \right) \right) \right) \] [Note what happened to \( \tau^2 \).]

• Call it \( \hat{RE} \left( \bar{X}; \tau^2 \right) \). [\( \hat{RE} \) is a fnct. of \( \bar{X} \), but not of \( \bar{Y} \).]

• This will have

\[(1) \quad E_{\theta_i} \left( \hat{RE} \left( \bar{X}; \tau^2 \right) \right) \approx n^{-1} \sum E_{\theta_i} \left( \ell_i \left( Y_i, q_i \left( X_i, \tau^2 \right) \right) \right) . \]

• Then choose

\[ \hat{\tau}^2_{RE} \left( \bar{X} \right) = \arg\min_{\tau^2} \left\{ \hat{RE} \left( \bar{X}; \tau^2 \right) \right\} . \]
Why does this yield the best $\hat{\tau}^2$ (as $n \to \infty$)?

Because there is a best $\hat{\tau}^2$ (as $n \to \infty$).

Oracle “Predictor”

- For every $\bar{\theta}$ let
  \[ \hat{\tau}_{OR}^2 = \hat{\tau}_{OR}^2(\bar{X};\bar{\theta}) = \arg\min_{\tau^2} R(\theta;\tau^2). \]

- Then show [as $n \to \infty$ for every (L1 b’nded) seq $\bar{\theta}$]
  \[
  (2) \quad \hat{\tau}_{RE}^2(\bar{X}) \to \hat{\tau}_{OR}^2(\bar{X};\bar{\theta}) \text{ in prob, } \& \\
  (3) \quad R(\bar{\theta};\hat{\tau}_{RE}^2) \to R(\bar{\theta};\hat{\tau}_{OR}^2).
  \]

- This conceptual scheme involves two major steps:
Two Major Steps

- Step 1: Create the (asymptotic) risk estimator $\hat{RE}(\tau^2)$ to yield a suitable approximation (1).
- Step 2: Prove it has the desired convergence and risk properties (2) and (3).

The two steps are interrelated:

- $\hat{RE}(\tau^2)$ needs to produce a satisfactory approximation in (1), and be computationally feasible.
- AND it also needs to enable Step 2.
Step 1: Asymptotic Risk Estimator

- Need to satisfy

$$E_{\theta_i}\left(\hat{RE}\left(\bar{X};\tau^2\right)\right) \approx n^{-1} \sum E_{\theta_i}\left(\ell_i\left(Y_i, q_i\left(X_i;\tau^2\right)\right)\right),$$

where $\ell_i$ is check-loss.

- If $\ell_i$ were squared error then can use SURE.
- But $\ell_i$ is not differentiable. So something like SURE seems a big stretch!
• HOWEVER, 

\[
\ell_i^* (\theta_i, q) \triangleq E_{\theta_i}^Y (\ell_i (Y, q)) = \sqrt{\nu_{F,i}} G \left( \frac{q - \theta_i}{\sqrt{\nu_{F,i}}}, b_i \right) \quad \exists \\

G(w, b) = \varphi(w) + w\Phi(w) - bw .
\]

• \( G \) is a smooth, \( C^\infty \), function. And \( \ell_i^* (\theta_i, q) \geq 0 \) plays the role of a conventional loss function.

• Here’s a picture of \( G \) for \( b = 0.7 \):
Plot of the function $G$
Creation of the Asymptotic Risk Estimator \( \text{ARE} \)

For \( v_{p,i} = 1 \) (for simplicity):

(a) Approximate \( G(w,b_i) \) by Taylor expansion ("TE") to \( K(i) \) terms.

(b) Substitute \( q_i(X_i;\tau) - \theta_i = w \).

(c) The resulting expectation is, say, \( E_{\theta_i}(TE(X_i,\theta_i)) \). It includes powers of \( \theta_i \) times functions of \( X_i \).

(d) Integrate by parts to create an equivalent expectation that contains only functions of \( X_i \) & no terms with \( \theta_i \).

(e) Step (d) could be understood as a version of SURE generalized to higher powers of \( \theta_i \).
(f) Actual computation is greatly facilitated by use of Hermite polynomials.

(g) The resulting expectand is a function of $X_i$ that is an unbiased estimate of the expected Taylor approximation. That’s the core of our $\text{ARE}$.

(h) There are some additional (clever) steps needed to handle large values of $W$, for which Taylor expansion is not good.

(i) Just to impress you, here’s the core of $\text{ARE}$ without the additional steps in (h): (In the following $U_i(\tau)$ is a minor truncation/modification of $X_i$ and $d_i(\tau)$ is a simple rational function of $\tau, v_{p,i}, b_i$. This is copied from the paper, which uses the notation “$\tau$” in place of “$\tau^2$”. )
\[ G(0, \tilde{b}_i) + G'(0, \tilde{b}_i)U_i(\tau) \]
\[ + \phi(0) \sum_{k=0}^{K_n(i)-2} \frac{(-1)^k H_k(0)}{(k+2)!} \left( \frac{2\nu_{p,i} d_i^2(\tau)}{H_k} \right)^{(k+2)/2} H_{k+2} \left( \frac{U_i(\tau)}{\left( 2\nu_{p,i} d_i^2(\tau) \right)^{1/2}} \right) \]

(j) Choose \( K_n(i) \), the number of terms in the Taylor expansion. This varies with \( i \), but is \( O(\log n) \).

(k) Then add over \( i \) to get \( \widehat{ARE} \).

(l) Finally, find the desired \( \text{arg min}_{\tau} \left( \widehat{ARE} \right) \) via a discrete grid search.
Step 2: Prove the desired asymptotic properties

There are some clues in Xie, Kou and Brown (2012). But parts of the proof here differ in key respects from what’s there, and the details here are much more complex.
Simulation #1
Homoscedastic ($\sigma^2 = 1$). Two types of $\left(\theta_i, b_i\right)$ pairs:

- $\left(0.58, 0.51\right)$ with prob 0.9
- $\left(5.1, 0.9\right)$ with prob 0.1

Comparison of pred. risks of three pred. methods

<table>
<thead>
<tr>
<th>#coord’s→</th>
<th>n=20</th>
<th>n=20</th>
<th>n=50</th>
<th>n=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method ↓</td>
<td>Efficiency</td>
<td>Ave. of $\hat{\tau}^2$</td>
<td>Efficiency %</td>
<td>Ave. of $\hat{\tau}^2$</td>
</tr>
<tr>
<td>ARE</td>
<td>86%</td>
<td>1.21</td>
<td>99%</td>
<td>0.344</td>
</tr>
<tr>
<td>ML/MM</td>
<td>68%</td>
<td>0.037</td>
<td>69%</td>
<td>0.037</td>
</tr>
<tr>
<td>Oracle</td>
<td>100%</td>
<td>0.296</td>
<td>69%</td>
<td>0.296</td>
</tr>
</tbody>
</table>

Note: Because model is homoscedastic, MaxLik and MethMom are the same
Simulations #3
This was a set of 6 simulations involving different scenarios, all with heteroscedastic data and random choices for parameters and $b_i$. E.g., in the first setup (the simplest)

$v_p^2 \sim U(0.1, 0.33), v_F^2 = 1, \theta \sim U(0, 1), b \sim U(0.5, 0.99)$, all indep.

In these 6 scenarios the efficiency of our ARE predictor relative to the oracle was

For $n=20$: $88\% < \text{Eff.} < 98\%$
For $n=100$: $91\% < \text{Eff.} < 99\%$.

[Note: The last scenario involved uniformly distributed observations, rather than normally distributed ones. But it isn’t reported in the manuscript whether the oracle knew and used that fact. I’ll need to ask Gourab and Paat.]
Recap

• Problem involves multivariate normal means
• Goal is quantile predictions of future observations
• A shrinkage predictor is produced
  o Shrinkage is produced by a hierarchical conjugate setup.
  o Quality of quantile prediction measured through predictive check-loss, which is the natural loss for quantile prediction.
• Methodology involves creation and use of a complicated but computable Asymptotic Risk Estimate (ARE) involving Hermite polynomials.
• Method is asymptotically justified. And
• Appears to work well in simulations for moderate $n$. 
Concluding Remarks for BFF4

1. Quantile predictor methodology is motivated from a Bayesian hierarchical structure.

2. The form of the predictor then drives the methodology.

3. The ARE method is a CONDITIONAL FREQUENTIST notion.

4. BUT: The proposed ARE predictor is NOT BAYESIAN – It doesn’t result from any prior I can think of. not even asymptotically.

5. Well known, but often overlooked – Prediction is different from parameter estimation. An estimate of the sample quantile doesn’t directly lead to quantile prediction.