OPTIMAL SHRINKAGE ESTIMATION OF MEAN PARAMETERS IN FAMILY OF DISTRIBUTIONS WITH QUADRATIC VARIANCE

BY XIANCHAO XIE*,§, S. C. KOU†,¶ and LAWRENCE BROWN‡,∥

Two Sigma Investments LLC§, Harvard University¶ and University of Pennsylvania∥

This paper discusses the simultaneous inference of mean parameters in a family of distributions with quadratic variance function. We first introduce a class of semi-parametric/parametric shrinkage estimators and establish their asymptotic optimality properties. Two specific cases, the location-scale family and the natural exponential family with quadratic variance function, are then studied in detail. We conduct a comprehensive simulation study to compare the performance of the proposed methods with existing shrinkage estimators. We also apply the method to real data and obtain encouraging results.

1. Introduction. The simultaneous inference of several mean parameters has interested statisticians since the 1950s and has been widely applied in many scientific and engineering problems ever since. Stein (1956) and James and Stein (1961) discussed the homoscedastic (equal variance) normal model and proved that shrinkage estimators can have uniformly smaller risk compared to the ordinary maximum likelihood estimate. This seminal work inspired a broad interest in the study of shrinkage estimators in hierarchical normal models. Efron and Morris (1972, 1973) studied the James-Stein estimators in an empirical Bayes framework and proposed several competing shrinkage estimators. Berger and Strawderman (1996) discussed this problem from a hierarchical Bayesian perspective. For applications of shrinkage techniques in practice, see Efron and Morris (1975), Rubin (1981), Morris (1983), Green and Strawderman (1985), Jones (1991) and Brown (2008).

There has also been substantial discussion on simultaneous inference for non-Gaussian cases. Brown (1966) studied the admissibility of invariance es-

In this article, we focus on the simultaneous inference of the mean parameters for families of distributions with quadratic variance function. These distributions include many common ones, such as the normal, Poisson, binomial, negative-binomial and gamma distributions; they also include location-scale families, such as the $t$, logistic, uniform, Laplace, Pareto and extreme value distributions. We ask the question: among all the estimators that estimate the mean parameters by shrinking the within-group sample mean toward a central location, is there an optimal one, subject to the intuitive constraint that more shrinkage is applied to observations with larger variances (or smaller sample sizes)? We propose a class of semi-parametric/parametric shrinkage estimators and show that there is indeed an asymptotically optimal shrinkage estimator; this estimator is explicitly obtained by minimizing an unbiased estimate of the risk. We note that similar types of estimators are found in Xie, Kou and Brown (2012) in the context of the heteroscedastic (unequal variance) hierarchical normal model. The treatment in this article, however, is far more general, as it covers a much wider range of distributions. We illustrate the performance of our shrinkage estimators by a comprehensive simulation study on both exponential and location-scale families. We apply our shrinkage estimators on the baseball data obtained by Brown (2008), observing quite encouraging results.

The remainder of the article is organized as follows: In Section 2, we introduce the general class of semi-parametric URE shrinkage estimators, identify the asymptotically optimal one, and discuss its properties. We then study the special case of location-scale families in Section 3 and the case of natural exponential families with quadratic variance (NEF-QVF) in Section 4. A systematic simulation study is conducted in Section 5, along with the application of the proposed methods to the baseball data set in Section 6. We give a brief discussion in Section 7 and the technical proofs are placed
2. Semi-parametric estimation of mean parameters of distributions with quadratic variance function. Consider simultaneously estimating the mean parameters of $p$ independent observations $Y_i$, $i = 1, \ldots, p$. Assume that the observation $Y_i$ comes from a distribution with quadratic variance function, that is, $E(Y_i) = \theta_i \in \Theta$ and $\text{Var}(Y_i) = V(\theta_i)/\tau_i$ such that

$$V(\theta_i) = \nu_0 + \nu_1 \theta_i + \nu_2 \theta_i^2$$

with $\nu_k$ ($k = 0, 1, 2$) being known constants. The set $\Theta$ of allowable parameters is a subset of $\{\theta : V(\theta) \geq 0\}$. $\tau_i$ is assumed to be known here and can be interpreted as the within-group sample size (i.e., when $Y_i$ is the sample average of the $i$-th group) or as the (square root) inverse-scale of $Y_i$. It is worth emphasizing that distributions with quadratic variance function include many common ones, such as the normal, Poisson, binomial, negative-binomial and gamma distributions as well as location-scale families. We introduce the general theory on distributions with quadratic variance function in this section and will specifically treat the cases of location-scale family and exponential family in the next two sections.

In a simultaneous inference problem, hierarchical models are often used to achieve partial pooling of information among different groups. For example, in the famous normal-normal hierarchical model $Y_i \overset{\text{iid}}{\sim} N(\theta_i, 1/\tau_i)$, one often puts a conjugate prior distribution $\theta_i \overset{\text{i.i.d.}}{\sim} N(\mu, \lambda)$ and uses the posterior mean

$$(2.1) \quad E(\theta_i | \mathbf{Y}; \mu, \lambda) = \frac{\tau_i}{\tau_i + 1/\lambda} \cdot Y_i + \frac{1/\lambda}{\tau_i + 1/\lambda} \cdot \mu$$

to estimate $\theta_i$. Similarly, if $Y_i$ represents the within-group average of Poisson observations $Y_i \overset{\text{iid}}{\sim} \text{Poisson}(\tau_i \theta_i)$, then with a conjugate gamma prior distribution $\theta_i \overset{\text{i.i.d.}}{\sim} \Gamma(\alpha, \lambda)$, the posterior mean

$$(2.2) \quad E(\theta_i | \mathbf{Y}; \alpha, \lambda) = \frac{\tau_i}{\tau_i + 1/\lambda} \cdot Y_i + \frac{1/\lambda}{\tau_i + 1/\lambda} \cdot \alpha \lambda$$

is often used to estimate $\theta_i$. The hyper-parameters, $(\mu, \lambda)$ or $(\alpha, \lambda)$ above, are usually first estimated from the marginal distribution of $Y_i$ and then plugged into the above formulae to form an empirical Bayes estimate.

One potential drawback of the formal parametric empirical Bayes approaches lies in its explicit parametric assumption on the prior distribution. It can lead to undesirable results if the explicit parametric assumption is
violated in real applications – we will see a real-data example in Section 6. Aiming to provide more flexible and, at the same time, efficient simultaneous estimation procedures, we consider in this section a class of semi-parametric shrinkage estimators.

To motivate these estimators, let us go back to the normal and Poisson examples (2.1) and (2.2). It is seen that the Bayes estimate of each mean parameter \( \theta_i \) is the weighted average of \( Y_i \) and the prior mean \( \mu \) (or \( \alpha \lambda \)). In other words, \( \theta_i \) is estimated by shrinking \( Y_i \) toward a central location (\( \mu \) or \( \alpha \lambda \)). It is also noteworthy that the amount of shrinkage is governed by \( \tau_i \), the sample size: the larger the sample size, the less is the shrinkage toward the central location. This feature makes intuitive sense. We will see in Section 4.2 that in fact these observations hold not only for normal and Poisson distributions, but also for general natural exponential families.

With these observations in mind, we consider in this section shrinkage estimators of the following form

\[
\hat{\theta}_i^{b,\mu} = (1 - b_i) \cdot Y_i + b_i \cdot \mu
\]

with \( b_i \in [0, 1] \) satisfying

\[
0 < b_i < b_j \quad \text{for any } i \text{ and } j \text{ such that } \tau_i \geq \tau_j.
\]

Requirement (MON) asks the estimator to shrink the group mean with a larger sample size (or smaller variance) less toward the central location. Other than this intuitive requirement, we do not put on any restriction on \( b_i \). Therefore, this class of estimators is semi-parametric in nature.

The question we want to investigate is, for such a general estimator \( \hat{\theta}_i^{b,\mu} \), whether there exists an optimal choice of \( b \) and \( \mu \). Note that the two parametric estimates (2.1) and (2.2) are special cases of the general class with \( b_i = \frac{1}{\tau_i + 1/\lambda} \). We will see shortly that such an optimal choice indeed exists asymptotically (i.e., as \( p \to \infty \)) and this asymptotically optimal choice is specified by an unbiased risk estimate (URE).

For a general estimator \( \hat{\theta}_i^{b,\mu} \) with fixed \( b \) and \( \mu \), under the sum of squared-error loss

\[
l_p(\theta, \hat{\theta}_i^{b,\mu}) = \frac{1}{p} \sum_{i=1}^{p} (\hat{\theta}_i^{b,\mu} - \theta_i)^2,
\]

an unbiased estimate of its risk \( R_p(\theta, \hat{\theta}_i^{b,\mu}) = E[l_p(\theta, \hat{\theta}_i^{b,\mu})] \) is given by

\[
\text{URE}(b, \mu) = \frac{1}{p} \sum_{i=1}^{p} \left[ b_i^2 \cdot (Y_i - \mu)^2 + (1 - 2b_i) \cdot \frac{V(Y_i)}{\tau_i + \nu_2} \right],
\]
because

\[
E[\text{URE}(b, \mu)] = \frac{1}{p} \sum_{i=1}^{p} \{b_i^2[\text{Var}(Y_i) + (\theta_i - \mu)^2] + (1 - 2b_i)\text{Var}(Y_i)\} = \frac{1}{p} \sum_{i=1}^{p} [(1 - b_i)^2\text{Var}(Y_i) + b_i^2(\theta_i - \mu)^2] = R_p(\theta, \hat{\theta}^{b,\mu}).
\]

Note that the idea and results below can be easily extended to the case of weighted quadratic loss, with the only difference being that the regularity conditions will then involve the corresponding weight sequence.

Ideally the “best” choice of \(b\) and \(\mu\) is the one that minimizes \(R_p(\theta, \hat{\theta}^{b,\mu})\), which is, however, unobtainable as the risk depends on the unknown \(\theta\). To bypass this impracticability, we minimize URE, the unbiased estimate, with respect to \((b, \mu)\) instead. This gives our semi-parametric URE shrinkage estimator:

\[
(2.5) \quad \hat{\theta}^{SM}_i = (1 - \hat{b}_i) \cdot Y_i + \hat{b}_i \cdot \hat{\mu}^{SM}
\]

where

\[
(\hat{b}^{SM}, \hat{\mu}^{SM}) = \text{minimizer of } \text{URE}(b, \mu)
\]

subject to \(b_i \in [0, 1], \mu \in [-\max_i |Y_i|, \max_i |Y_i|] \cap \Theta\) and Requirement (MON).

We require \(|\mu| \leq \max_i |Y_i|\), since no sensible estimators shrink the observations to a location completely outside the range of the data. Intuitively, the URE shrinkage estimator would behave well if URE\((b, \mu)\) is close to the risk \(R_p(\theta, \hat{\theta}^{b,\mu})\).

To investigate the properties of the semi-parametric URE shrinkage estimator, we now introduce the following regularity conditions:

(A) \(\lim sup \frac{1}{p} \sum_{i=1}^{p} \text{Var}(Y_i) < \infty\);

(B) \(\lim sup \frac{1}{p} \sum_{i=1}^{p} \text{Var}(Y_i) \cdot \theta_i^2 < \infty\);

(C) \(\lim sup \frac{1}{p} \sum_{i=1}^{p} \text{Var}(Y_i^2) < \infty\);

(D) \(\sup_i \left(\frac{\tau_i}{\tau_i + \nu_2}\right)^2 < \infty, \sup_i \left(\frac{\nu_i}{\tau_i + \nu_2}\right)^2 < \infty\);

(E) \(\lim sup \frac{1}{p^{1+\varepsilon}} E(\max_{1 \leq i \leq p} Y_i^2) < \infty\) for some \(\varepsilon > 0\).

The theorem below shows that URE\((b, \mu)\) not only unbiasedly estimates the risk, but also serves as a good approximation of the actual loss \(l_p(\theta, \hat{\theta}^{b,\mu})\),
which is a much stronger property. In fact, \( \text{URE}(b, \mu) \) is asymptotically uniformly close to the actual loss. Therefore, we expect that minimizing \( \text{URE}(b, \mu) \) would lead to an estimate with competitive risk properties.

**Theorem 2.1.** Assuming regularity conditions (A)-(E), we have

\[
\sup \left| \text{URE}(b, \mu) - l_p(\theta, \hat{\theta}^{b,\mu}) \right| \to 0 \text{ in } L^1 \text{ and in probability, as } p \to \infty.
\]

where the supremum is taken over \( b \in [0,1], |\mu| \leq \max_i |Y_i| \) and Requirement (MON).

The following result compares the asymptotic behavior of our URE shrinkage estimator (2.5) with other shrinkage estimators from the general class. It establishes the asymptotic optimality of our URE shrinkage estimator.

**Theorem 2.2.** Assume regularity conditions (A)-(E). Then for any shrinkage estimator \( \hat{\theta}^{b,\mu} = (1 - \hat{b}) \cdot Y + \hat{b} \cdot \hat{\mu} \), where \( \hat{b} \in [0,1] \) satisfies Requirement (MON), and \( |\hat{\mu}| \leq \max_i |Y_i| \), we always have

\[
\lim_{p \to \infty} P \left( l_p(\theta, \hat{\theta}^{SM}) \geq l_p(\theta, \hat{\theta}^{b,\mu}) + \varepsilon \right) = 0 \text{ for any } \varepsilon > 0
\]

and

\[
\limsup_{p \to \infty} \left[ R(\theta, \hat{\theta}^{SM}) - R(\theta, \hat{\theta}^{b,\mu}) \right] \leq 0.
\]

As a special case of Theorem 2.2, the semi-parametric URE shrinkage estimator asymptotically dominates the parametric empirical Bayes estimators, like (2.1) or (2.2). It is worth noting that the asymptotic optimality of our semi-parametric URE shrinkage estimators does not assume any prior distribution on the mean parameters \( \theta_i \), nor does it assume any parametric form on the distribution of \( Y \) (other than the quadratic variance function). Therefore, the results enjoy a large extent of robustness. In fact, the individual \( Y_i \)'s do not even have to be from the same distribution family as long as the regularity conditions (A)-(E) are met. (However, whether shrinkage estimation in that case is a good idea or not becomes debatable.)

2.1. Shrinking toward the grand mean. In the previous development, the central shrinkage location \( \mu \) is determined by minimizing URE. The joint minimization of \( b \) and \( \mu \) offers asymptotic optimality in the class of estimators. For small or moderate \( p \) (the number of \( Y_i \)'s), however, it is not necessarily true that the semi-parametric URE shrinkage estimator will always be the optimal one. In this setting, it might be beneficial to set \( \mu \) by
a predetermined rule and only optimize $b$, as it might reduce the variability of the resulting estimate. In this subsection, we consider shrinking toward the grand mean:

$$\hat{\mu} = \bar{Y} = \frac{1}{p} \sum_{i=1}^{p} Y_i.$$  

The particular reason why the grand average is chosen instead of the weighted average $\bar{Y}_w = (\sum_{i=1}^{p} \tau_i Y_i) / (\sum_{i=1}^{p} \tau_i)$ is that the latter might be biased when $\theta_i$ and $\tau_i$ are dependent. In the case where such dependence is not a concern, the idea and results obtained below can be similarly derived.

The corresponding estimator becomes

$$\hat{Y}_i^{b,\bar{Y}} = (1 - b_i) Y_i + b_i \bar{Y},$$

where $b_i \in [0, 1]$ satisfies Requirement (MON). To find the asymptotically optimal choice of $b$, we start from an unbiased estimate of the risk of $\hat{Y}_i^{b,\bar{Y}}$.

It is straightforward to verify that for fixed $b$ an unbiased estimate of the risk of $\hat{Y}_i^{b,\bar{Y}}$ is

$$\text{URE}_G(b) = \frac{1}{p} \sum_{i=1}^{p} \left[ b_i^2 (Y_i - \bar{Y})^2 + \left( 1 - 2(1 - \frac{1}{p}) b_i \right) \frac{V(Y_i)}{\tau_i + \nu_2} \right].$$

Note that we use the superscript “$G$”, which stands for “grand mean”, to distinguish it from the previous URE$(b, \mu)$. Like what we did previously, minimizing the URE$^G$ with respect to $b$ then leads to our semi-parametric URE grand-mean shrinkage estimator

$$\hat{Y}_i^{SG} = (1 - \hat{b}_i^{SG}) \cdot Y_i + \hat{b}_i^{SG} \cdot \bar{Y},$$

where

$$\hat{b}_i^{SG} = \text{minimizer of } \text{URE}_G(b)$$

subject to $b_i \in [0, 1]$ and Requirement (MON).

Again, we expect that the URE estimator $\hat{Y}_i^{SG}$ would be competitive if URE$^G$ is close to the risk function or the loss function. The next theorem confirms the uniform closeness.

**Theorem 2.3.** Under regularity conditions (A)-(E), we have

$$\sup |\text{URE}_G(b) - l_p(\theta, \hat{Y}_i^{b,\bar{Y}})| \to 0 \text{ in } L^1 \text{ and in probability, as } p \to \infty.$$

where the supremum is taken over $b_i \in [0, 1]$ and Requirement (MON).
Consequently, $\hat{\Theta}^{SG}$ is asymptotically optimal among all shrinkage estimators $\hat{\Theta}^{b,Y}$ that shrink toward the grand mean $\bar{Y}$, as shown in the next theorem.

**Theorem 2.4.** Assume regularity conditions (A)-(E). Then for any shrinkage estimator $\hat{\Theta}^{b,Y} = (1 - \hat{b}) \cdot Y + \hat{b} \cdot \bar{Y}$, where $\hat{b} \in [0,1]$ satisfies Requirement (MON), we have

$$\lim_{p \to \infty} P \left( l_p(\Theta, \hat{\Theta}^{SG}) \geq l_p(\Theta, \hat{\Theta}^{b,Y}) + \varepsilon \right) = 0 \quad \text{for any } \varepsilon > 0$$

and

$$\limsup_{p \to \infty} \left[ R(\Theta, \hat{\Theta}^{SG}) - R(\Theta, \hat{\Theta}^{b,Y}) \right] \leq 0.$$

3. Simultaneous estimation of mean parameters in location-scale families. In this section, we focus on location-scale families, which are a special case of distributions with quadratic variance functions. We show how the regularity conditions can be simplified in this case.

For a location-scale family, we can write

$$Y_i = \theta_i + Z_i / \sqrt{\tau_i}$$

where the standard variates $Z_i$ are i.i.d. with mean zero and variance $\nu_0$. The constants $\nu_1$ and $\nu_2$ in the quadratic variance function $V(\theta_i) = \nu_0 + \nu_1 \theta_i + \nu_2 \theta_i^2$ are zero for the location-scale family. $1 / \sqrt{\tau_i}$ is the scale of $Y_i$.

The next lemma simplifies the regularity conditions for a location-scale family.

**Lemma 3.1.** For $Y_i, i = 1, \ldots, p$, independently from a location-scale family (3.1), the following four conditions imply the regularity conditions (A)-(E) in Section 2:

(i) $\limsup_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\tau_i^2} < \infty$;

(ii) $\limsup_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \theta_i^2 / \tau_i < \infty$;

(iii) $\limsup_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} |\theta_i|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$;

(iv) the standard variate $Z$ satisfies $P(|Z| > t) \leq Dt^{-\alpha}$ for constants $D > 0, \alpha > 4$.

Note that (i)-(iv) in Lemma 3.1 covers the common location-scale families, including the $t$ (degree of freedom $> 4$), normal, uniform, logistic, Laplace, Pareto ($\alpha > 4$) and extreme value distributions.
Lemma 3.1, together with Theorems 2.1 and 2.2, immediately yields the following corollaries: the semi-parametric URE shrinkage estimator is asymptotically optimal for location-scale families.

**Corollary 3.2.** For \( Y_i, i = 1, \ldots, p \), independently from a location-scale family (3.1), under conditions (i)-(iv) in Lemma 3.1, we have
\[
\sup \left| \text{URE}(b, \mu) - l_p(\theta, \hat{\theta}_p) \right| \to 0 \text{ in } L^1 \text{ and in probability, as } p \to \infty.
\]
where the supremum is taken over \( b \in [0, 1], |\mu| \leq \max_i |Y_i| \) and Requirement (MON).

**Corollary 3.3.** Let \( Y_i, i = 1, \ldots, p \), be independent from a location-scale family (3.1). Assume conditions (i)-(iv) in Lemma 3.1. Then for any shrinkage estimator \( \hat{\theta}_p = (1 - \hat{b}) \cdot \bar{Y} + \hat{b} \cdot \hat{\mu} \), where \( \hat{b} \in [0, 1] \) satisfies Requirement (MON), and \( |\hat{\mu}| \leq \max_i |Y_i| \), we always have
\[
\lim_{p \to \infty} P \left( l_p(\theta, \hat{\theta}_p) \geq l_p(\theta, \hat{\theta}_p) + \varepsilon \right) = 0 \text{ for any } \varepsilon > 0
\]
and
\[
\limsup_{p \to \infty} \left[ R(\theta, \hat{\theta}_p) - R(\theta, \hat{\theta}_p) \right] \leq 0.
\]

In the case of shrinking toward the grand mean \( \bar{Y} \), the corresponding semi-parametric URE grand-mean shrinkage estimator is also asymptotically optimal.

**Corollary 3.4.** For \( Y_i, i = 1, \ldots, p \), independently from a location-scale family (3.1), under conditions (i)-(iv) in Lemma 3.1, we have
\[
\sup \left| \text{URE}^G(b) - l_p(\theta, \hat{\theta}_p) \right| \to 0 \text{ in } L^1 \text{ and in probability, as } p \to \infty.
\]
where the supremum is taken over \( b \in [0, 1] \) and Requirement (MON).

**Corollary 3.5.** Let \( Y_i, i = 1, \ldots, p \), be independent from a location-scale family (3.1). Assume conditions (i)-(iv) in Lemma 3.1. Then for any shrinkage estimator \( \hat{\theta}_p = (1 - \hat{b}) \cdot \bar{Y} + \hat{b} \cdot \bar{Y} \), where \( \hat{b} \in [0, 1] \) satisfies Requirement (MON), we have
\[
\lim_{p \to \infty} P \left( l_p(\theta, \hat{\theta}_p) \geq l_p(\theta, \hat{\theta}_p) + \varepsilon \right) = 0 \text{ for any } \varepsilon > 0
\]
and
\[
\limsup_{p \to \infty} \left[ R(\theta, \hat{\theta}_p) - R(\theta, \hat{\theta}_p) \right] \leq 0.
\]

4.1. Semi-parametric URE shrinkage estimators. In this section, we focus on natural exponential families with quadratic variance functions (NEF-QVF), as they incorporate the most common distributions that one encounters in practice. We show how the regularity conditions (A)-(E) can be significantly simplified and offer concrete examples.

It is well known that there are in total six distinct distributions that belong to NEF-QVF (Morris, 1982): the normal, binomial, Poisson, negative-binomial, Gamma and generalized hyperbolic secant (GHS) distributions. We represent in general an NEF-QVF as

\[ Y_i \sim \text{NEF-QVF}[	heta_i, V(\theta_i)/\tau_i] \]

where \( \theta_i = E(Y_i) \in \Theta \) is the mean parameter and \( \tau_i \) is the convolution parameter (or within-group sample size). For example, in the binomial case \( Y_i \sim \text{Bin}(n_i, p_i)/n_i \), \( \theta_i = p_i \), \( V(\theta_i) = \theta_i(1 - \theta_i) \) and \( \tau_i = n_i \).

The next result provides easy-to-check conditions that considerably simplify those in Section 2. As the case of heteroscedastic normal data has been studied in Xie, Kou and Brown (2012), we concentrate on the other five NEF-QVF distributions.

**Lemma 4.1.** For the five non-Gaussian NEF-QVF distributions, Table 1 lists the respective conditions, under which regularity conditions (A)-(E) in Section 2 are satisfied. For example, for the binomial distribution, the condition is \( \tau_i = n_i \geq 2 \) for all \( i = 1, \ldots, p \).

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Data ( (\nu_0, \nu_1, \nu_2) )</th>
<th>Note</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>binomial</td>
<td>( Y_i \sim \text{Bin}(n_i, p_i)/n_i ) ( (0, 1, -1) ) ( \tau_i = n_i )</td>
<td>( n_i \geq 2 ) for all ( i = 1, \ldots, p )</td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>( Y_i \sim \text{Poi}(\tau_i \theta_i)/\tau_i ) ( (0, 1, 0) )</td>
<td>( \theta_i = p_i )</td>
<td>(i) ( \inf \tau_i &gt; 0 ), ( \inf \tau_i (\tau_i \theta_i) &gt; 0 ) (ii) ( \sum \theta_i^2 = O(p) )</td>
</tr>
<tr>
<td>neg-binomial</td>
<td>( Y_i \sim \text{NBin}(n_i, p_i)/n_i ) ( (0, 1, 1) ) ( \tau_i = n_i ) ( \theta_i = \frac{n_i p_i}{p_i} )</td>
<td>( \theta_i = \alpha \lambda_i )</td>
<td>(i) ( \inf \tau_i &gt; 0 ), ( \inf (\tau_i \theta_i) &gt; 0 ) (ii) ( \sum \lambda_i^4 = O(p) )</td>
</tr>
<tr>
<td>Gamma</td>
<td>( Y_i \sim \Gamma(\tau_i \alpha, \lambda_i)/\tau_i ) ( (0, 0, 1/\alpha) ) ( \theta_i = \alpha \lambda_i )</td>
<td></td>
<td>(i) ( \inf \tau_i &gt; 0 ) (ii) ( \sum \lambda_i^4 = O(p) )</td>
</tr>
<tr>
<td>GHS</td>
<td>( Y_i \sim \text{GHS}(\tau_i \alpha, \lambda_i)/\tau_i ) ( (\alpha, 0, 1/\alpha) ) ( \theta_i = \alpha \lambda_i )</td>
<td></td>
<td>(i) ( \inf \tau_i &gt; 0 ) (ii) ( \sum \lambda_i^4 = O(p) )</td>
</tr>
</tbody>
</table>

**Table 1**

The conditions for the five non-normal NEF-QVF distributions to guarantee the regularity conditions (A)-(E) in Section 2.
Lemma 4.1 and the general theory in Section 2 yield the following optimality results for our semi-parametric URE shrinkage estimator in the case of NEF-QVF.

**Corollary 4.2.** Let \( Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}\left(\theta_i, V(\theta_i)/\tau_i\right), \ i = 1, \ldots, p, \) be non-Gaussian. Under the respective conditions listed in Table 1, we have

\[
\sup \left| \text{URE}(b, \mu) - l_p(\theta, \hat{b}^{\mu}) \right| \to 0 \text{ in } L^1 \text{ and in probability, as } p \to \infty,
\]

where the supremum is taken over \( b_i \in [0, 1], \ |\mu| \leq \max_i |Y_i| \) and Requirement (MON).

**Corollary 4.3.** Let \( Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}\left(\theta_i, V(\theta_i)/\tau_i\right), \ i = 1, \ldots, p, \) be non-Gaussian. Assume the respective conditions listed in Table 1. Then for any shrinkage estimator \( \hat{b}^{\mu} = (1 - \hat{b}) \cdot \hat{Y} + \hat{b} \cdot \hat{\mu}, \) where \( \hat{b} \in [0, 1] \) satisfies Requirement (MON), and \( |\mu| \leq \max_i |Y_i|, \) we always have

\[
\lim_{p \to \infty} P \left( l_p(\theta, \hat{b}^{\mu}) \geq l_p(\theta, \hat{b}^{\mu} + \varepsilon) \right) = 0 \quad \text{for any } \varepsilon > 0
\]

and

\[
\limsup_{p \to \infty} \left[ R(\theta, \hat{b}^{\mu}) - R(\theta, \hat{b}^{\mu}) \right] \leq 0.
\]

For shrinking toward the grand mean \( \hat{Y}, \) the corresponding semi-parametric URE grand mean shrinkage estimator is also asymptotically optimal (within the smaller class).

**Corollary 4.4.** Let \( Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}\left(\theta_i, V(\theta_i)/\tau_i\right), \ i = 1, \ldots, p, \) be non-Gaussian. Under the respective conditions listed in Table 1, we have

\[
\sup \left| \text{URE}^G(b) - l_p(\theta, \hat{b}^{\mu} \cdot \hat{Y}) \right| \to 0 \text{ in } L^1 \text{ and in probability, as } p \to \infty,
\]

where the supremum is taken over \( b_i \in [0, 1] \) and Requirement (MON).

**Corollary 4.5.** Let \( Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}\left(\theta_i, V(\theta_i)/\tau_i\right), \ i = 1, \ldots, p, \) be non-Gaussian. Assume the respective conditions listed in Table 1. Then for any shrinkage estimator \( \hat{b}^{\mu} = (1 - \hat{b}) \cdot \hat{Y} + \hat{b} \cdot \hat{\mu}, \) where \( \hat{b} \in [0, 1] \) satisfies Requirement (MON), we have

\[
\lim_{p \to \infty} P \left( l_p(\theta, \hat{b}^{SG}) \geq l_p(\theta, \hat{b}^{\mu} \cdot \hat{Y} + \varepsilon) \right) = 0 \quad \text{for any } \varepsilon > 0
\]

and

\[
\limsup_{p \to \infty} \left[ R(\theta, \hat{b}^{SG}) - R(\theta, \hat{b}^{\mu} \cdot \hat{Y}) \right] \leq 0.
\]
4.2. Parametric URE shrinkage estimators and conjugate priors. For \( Y_i \) from an exponential family, hierarchical models based on the conjugate prior distributions are often used; the hyperparameters in the prior distribution are often estimated from the marginal distribution of \( Y_i \) in an empirical Bayes way. Two questions arise naturally in this scenario. First, are there other choices to estimate the hyperparameters? Second, is there an optimal choice? We will show in this subsection that our URE shrinkage idea applies to the parametric conjugate prior case; the resulting parametric URE shrinkage estimators are asymptotically optimal and thus asymptotically dominate the traditional empirical Bayes estimators.

Let \( Y_i \sim \text{NEF-QVF}[\theta_i, V(\theta_i)/\tau_i], i = 1, \ldots, p \), be independent. If \( \theta_i \) are i.i.d. from the conjugate prior, the Bayesian estimate of \( \theta_i \) is then

\[
\hat{\theta}_i = \frac{\tau_i}{\tau_i + \gamma} \cdot Y_i + \frac{\gamma}{\tau_i + \gamma} \cdot \mu,
\]

where \( \gamma \) and \( \mu \) are functions of the hyperparameters in the prior distribution. Table 2 details the conjugate priors for the five well known NEF-QVF distributions – the normal, binomial, Poisson, negative-binomial and Gamma distributions – and the corresponding expressions of \( \gamma \) and \( \mu \) in terms of the hyperparameters. For example, in the binomial case, \( Y_i \overset{\text{iid}}{\sim} \text{Bin}(\tau_i, \theta_i)/\tau_i \) and the conjugate prior is \( \theta_i \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta); \gamma = \alpha + \beta, \mu = \alpha/(\alpha + \beta) \).

Although it can be shown that for the sixth NEF-QVF distribution – the GHS distribution – taking a conjugate prior also gives (4.1), the conjugate prior distribution does not have a clean expression and is rarely encountered in practice. We thus omit it from Table 2.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Data</th>
<th>Conjugate prior</th>
<th>( \gamma ) and ( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>( Y_i \sim N(\theta_i, 1/\tau_i) )</td>
<td>( \theta_i \overset{\text{iid}}{\sim} N(\mu, \lambda) )</td>
<td>( \gamma = 1/\lambda )</td>
</tr>
<tr>
<td>binomial</td>
<td>( Y_i \sim \text{Bin}(\tau_i, \theta_i)/\tau_i )</td>
<td>( \theta_i \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta) )</td>
<td>( \gamma = \alpha + \beta, \mu = \frac{\alpha}{\alpha + \beta} )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( Y_i \sim \text{Poi}(\tau_i, \theta_i)/\tau_i )</td>
<td>( \theta_i \overset{\text{iid}}{\sim} \text{Gamma} (\alpha, \lambda) )</td>
<td>( \gamma = 1/\lambda, \mu = \alpha \lambda )</td>
</tr>
<tr>
<td>neg-binomial</td>
<td>( Y_i \sim \text{NBin}(\tau_i, p_i)/\tau_i )</td>
<td>( p_i \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta), \theta_i = \frac{p_i}{1 - p_i} )</td>
<td>( \gamma = \beta - 1, \mu = \frac{\alpha}{\beta - 1} )</td>
</tr>
<tr>
<td>Gamma</td>
<td>( Y_i \sim \Gamma(\tau_i, \lambda_i)/\tau_i )</td>
<td>( \lambda_i \overset{\text{iid}}{\sim} \text{inv-Gamma} (\alpha_0, \beta_0), \theta_i = \alpha \lambda_i )</td>
<td>( \gamma = \frac{\alpha_0 - 1}{\alpha}, \mu = \frac{\alpha_0 \beta_0}{\alpha_0 - 1} )</td>
</tr>
</tbody>
</table>

Table 2: Conjugate priors for the five well known NEF-QVF distributions.

We now apply our URE idea to formula (4.1) to estimate \( (\gamma, \mu) \), in contrast to the conventional empirical Bayes method that determines the hyperparameters through the marginal distribution. For fixed \( \gamma \) and \( \mu \), an unbiased
estimate for the risk of $\hat{\gamma}^{\gamma,\mu}$ is given by

$$\text{URE}^P(\gamma, \mu) = \frac{1}{p} \sum_{i=1}^{p} \left[ \frac{\gamma^2}{(\tau_i + \gamma)^2} \cdot (Y_i - \mu)^2 + \frac{\tau_i - \gamma}{\tau_i + \gamma} \cdot \frac{V(Y_i)}{\tau_i + \nu_2} \right],$$

where we use the superscript “$P$” to stand for “parametric”. Minimizing $\text{URE}^P(\gamma, \mu)$ leads to our parametric URE shrinkage estimator

$$(4.2) \quad \hat{\gamma}^{PM}_i = \frac{\tau_i}{\tau_i + \hat{\gamma}^{PM}} \cdot Y_i + \frac{\hat{\gamma}^{PM}}{\tau_i + \hat{\gamma}^{PM}} \cdot \hat{\mu}^{PM},$$

where

$$\left(\hat{\gamma}^{PM}, \hat{\mu}^{PM}\right) = \arg\min \text{URE}^P(\gamma, \mu) \text{ over } \{0 \leq \gamma \leq \infty, |\mu| \leq \max_i |Y_i|, \mu \in \Theta\}.$$ 

Parallel to Theorems 2.1 and 2.2, the next two results show that the parametric URE shrinkage estimator gives the asymptotically optimal choice of $(\gamma, \mu)$, if one wants to use estimators of the form (4.1).

**Theorem 4.6.** Let $Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}[\theta_i, V(\theta_i)/\tau_i], i = 1, \ldots, p$, be non-Gaussian. Under the respective conditions listed in Table 1, we have

$$\sup \left| \text{URE}^P(\gamma, \mu) - l_p(\theta, \hat{\theta}^{\gamma,\mu}) \right| \to \text{in } L^1 \text{ and in probability, as } p \to \infty$$

where the supremum is taken over $\{0 \leq \gamma \leq \infty, |\mu| \leq \max_i |Y_i|, \mu \in \Theta\}$.

**Theorem 4.7.** Let $Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}[\theta_i, V(\theta_i)/\tau_i], i = 1, \ldots, p$, be non-Gaussian. Assume the respective conditions listed in Table 1. Then for any estimator $\hat{\gamma}^{\hat{\gamma},\hat{\mu}} = \frac{\tau_i}{\tau_i + \hat{\gamma}} Y_i + \frac{\hat{\gamma}}{\tau_i + \hat{\gamma}} \hat{\mu}$, where $\hat{\gamma} \geq 0$ and $|\hat{\mu}| \leq \max_i |Y_i|$, we always have

$$\lim_{p \to \infty} P \left( l_p(\theta, \hat{\theta}^{PM}) \geq l_p(\theta, \hat{\theta}^{\gamma,\mu}) + \varepsilon \right) = 0 \quad \text{for any } \varepsilon > 0$$

and

$$\limsup_{p \to \infty} \left[ R_p(\theta, \hat{\theta}^{PM}) - R_p(\theta, \hat{\theta}^{\gamma,\mu}) \right] \leq 0.$$

In the case of shrinking toward the grand mean $\bar{Y}$ (when $p$, the number of $Y_i$'s, is small or moderate), we have the following parametric results parallel to the semi-parametric ones.

First, for the grand-mean shrinkage estimator

$$(4.3) \quad \hat{\gamma}^{\gamma,\bar{Y}} = \frac{\tau_i}{\tau_i + \gamma} \cdot Y_i + \frac{\gamma}{\tau_i + \gamma} \cdot \bar{Y},$$
with a fixed $\gamma$, an unbiased estimate of its risk is

$$\text{URE}^{PG}(\gamma) = \frac{1}{p} \sum_{i=1}^{p} \left[ \frac{\gamma^2}{(\tau_i + \gamma)^2} (Y_i - \bar{Y})^2 + \left(1 - 2 \frac{1}{p} \frac{\gamma}{\tau_i + \gamma} \right) V(Y_i) \right].$$

Minimizing it yields our parametric URE grand-mean shrinkage estimator

$$(4.4) \quad \hat{\theta}^{PG}_i = \frac{\tau_i}{\tau_i + \hat{\gamma}^{PG}} \cdot Y_i + \frac{\hat{\gamma}^{PG}}{\tau_i + \hat{\gamma}^{PG}} \cdot \bar{Y},$$

where

$$\hat{\gamma}^{PG} = \arg \min_{0 \leq \gamma \leq \infty} \text{URE}^{PG}(\gamma).$$

Similar to Theorems 4.6 and 4.7, the next two results show that in the case of shrinking toward the grand mean under the formula (4.3), the parametric URE grand-mean shrinkage estimator is asymptotically optimal.

**Theorem 4.8.** Let $Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}[\theta_i, V(\theta_i)/\tau_i], i = 1, \ldots, p$, be non-Gaussian. Under the respective conditions listed in Table 1, we have

$$\sup_{0 \leq \gamma < \infty} \left| \text{URE}^{PG}(\gamma) - l_p(\theta, \hat{\theta}; \bar{Y}) \right| \rightarrow \text{in } L^1 \text{ and in probability, as } p \rightarrow \infty.$$

**Theorem 4.9.** Let $Y_i \overset{\text{ind}}{\sim} \text{NEF-QVF}[\theta_i, V(\theta_i)/\tau_i], i = 1, \ldots, p$, be non-Gaussian. Assume the respective conditions listed in Table 1. Then for any estimator $\hat{\theta}; \bar{Y} = \frac{\tau}{\tau + \hat{\gamma}} Y + \frac{\hat{\gamma}}{\tau + \hat{\gamma}} \bar{Y},$ where $\hat{\gamma} \geq 0$, we have

$$\lim_{p \rightarrow \infty} P \left( l_p(\theta, \hat{\theta}; \bar{Y}) \geq l_p(\theta, \hat{\theta}; \bar{Y}) + \varepsilon \right) = 0 \quad \text{for any } \varepsilon > 0$$

and

$$\lim \inf_{p \rightarrow \infty} \left[ R_p(\theta, \hat{\theta}^{PG}) - R_p(\theta, \hat{\theta}; \bar{Y}) \right] \leq 0.$$

5. **Simulation Study.** In this section, we conduct a number of simulations to investigate the performance of the URE estimators and compare their performance to that of other existing shrinkage estimators. For each simulation, we first draw $(\theta_i, \tau_i), i = 1, \ldots, p$, independently from a distribution and then generate $Y_i$ given $(\theta_i, \tau_i)$. This process is repeated a large number of times ($N = 100,000$) to obtain an accurate estimate of the risk for each estimator. The sample size $p$ is chosen to vary from 20 to 500 at an interval of length 20. For notational convenience, in this section, we write $A_i = 1/\tau_i$ so that $A_i$ is (essentially) the variance.
5.1. Location-scale family. For the location-scale families, we consider three non-Gaussian cases: the Laplace (where the standard variate $Z$ has density $f(z) = \frac{1}{2} \exp(-|z|)$), logistic (where the standard variate $Z$ has density $f(z) = e^{-z}/(1+e^{-z})^2$) and student-t distributions with 7 degrees of freedom. To evaluate the performance of the semi-parametric URE estimator $\hat{\theta}^{SM}$, we compare it to the naive estimator

$$\hat{\theta}_i^{Naive} = Y_i$$

and the extended James-Stein estimator

$$(5.1) \quad \hat{\theta}_i^{JS+} = \mu^{JS+} + \left(1 - \frac{(p-3)}{\sum_{i=1}^{p} (Y_i - \mu^{JS+})^2/A_i} \right)^{+} Y_i$$

where $A_i = 1/\tau_i$ and $\mu^{JS+} = \sum_{i=1}^{p} (X_i/A_i)/\sum_{i=1}^{p} 1/A_i$.

For each of the three distributions we study four different setups to generate $(\theta_i, A_i)$ for $i = 1, \cdots, p$. We then generate $Y_i$ via (3.1) except for scenario (4) below.

**Scenario (1).** $(\theta_i, A_i)$ are drawn from $A_i \sim \text{Unif}(0,1)$ and $\theta_i \sim N(0,1)$ independently. In this scenario, the location and scale are independent of each other. Panels (a) in Figures 1-3 plot the performance of the three estimators. The risk function of the naive estimator $\hat{\theta}_i^{Naive}$, being a constant for all $p$, is way above the other two. The risk of the semi-parametric URE estimator is significantly smaller than that of both the extended James-Stein estimator and the naive estimator, particularly so when the sample size $p > 40$.

**Scenario (2).** $(\theta_i, A_i)$ are drawn from $A_i \sim \text{Unif}(0,1)$ and $\theta_i = A_i$. This scenario tests the case that the location and scale have a strong correlation. Panels (b) in Figures 1-3 show the performance of the three estimators. The risk of the semi-parametric URE estimator is significantly smaller than that of both the extended James-Stein estimator and the naive estimator. The naive estimator $\hat{\theta}_i^{Naive}$, with a constant risk, performs the worst. This example indicates that the semi-parametric URE estimator behaves robustly well even when there is a strong correlation between the location and the scale. This is because the semi-parametric URE estimator does not make any assumption on the relationship between $\theta_i$ and $\tau_i$.

**Scenario (3).** $(\theta_i, A_i)$ are drawn such that $A_i \sim \frac{1}{2} \cdot 1_{\{A_i=0.1\}} + \frac{1}{2} \cdot 1_{\{A_i=0.5\}}$ – that is, $A_i$ is 0.1 or 0.5 with 50% probability each – and that conditioning on $A_i$ being 0.1 or 0.5, $\theta_i | A_i = 0.1 \sim N(2,0.1)$; $\theta_i | A_i = 0.5 \sim N(0,0.5)$. This scenario tests the case that there are two underlying groups in the data.
Panels (c) in Figures 1-3 compare the performance of the semi-parametric URE estimator to that of the native estimator and the extended James-Stein estimator. The semi-parametric URE estimator is seen to significantly outperform the other two estimators.

**Scenario (4).** \( (\theta_i, A_i) \) are drawn from \( A_i \sim \text{Unif}(0,1) \) and \( \theta_i = A_i \). Given \( \theta_i \) and \( A_i \), \( Y_i \) are generated from \( Y_i \sim \text{Unif}(\theta_i - \sqrt{3A_i}\sigma, \theta + \sqrt{3A_i}\sigma) \), where \( \sigma \) is the standard deviation of the standard variate \( Z \), i.e., \( \sigma = \sqrt{2}, \pi/\sqrt{3} \) and \( \sqrt{7/5} \) for the Laplace, logistic and \( t \)-distribution (\( df = 7 \)), respectively. This scenario tests the case of model mis-specification and, hence, the robustness of the estimators. Note that \( Y_i \) is drawn from a uniform distribution, not from the Laplace, logistic or \( t \) distribution. Panels (d) in Figures 1-3 show the performance of the three estimators. It is seen that the naive estimator behaves the worst and that the semi-parametric URE estimator clearly outperforms the other two. This example indicates the robust performance of the semi-parametric URE estimator even when the model is incorrectly specified. This because URE estimator essentially only involves the first two moments of \( Y_i \); it does not rely on the specific density function of the distribution.

5.2. **Exponential family.** We consider exponential family in this subsection, conducting simulation evaluations on the beta-binomial and Poisson-gamma models.

5.2.1. **Beta-binomial hierarchical model.** For binomial observations \( Y_i \overset{i.i.d.}{\sim} \text{Bin}(\tau_i, \theta_i)/\tau_i \), classical hierarchical inference typically assumes the conjugate prior \( \theta_i \overset{i.i.d.}{\sim} \text{Beta}(\alpha, \beta) \). The marginal distribution of \( Y_i \) is used by classical empirical Bayes methods to estimate the hyperparameters. Plugging the estimate of these hyperparameters into the posterior mean of \( \theta_i \) given \( Y_i \) yields the empirical Bayes estimate of \( \theta_i \). In this subsection, we consider both the semi-parametric URE estimator \( \hat{\theta}^{SM} \) (equation (2.5)) and the parametric URE estimator \( \hat{\theta}^{PM} \) (equation (4.2)), and compare them with the parametric empirical Bayes maximum likelihood estimator \( \hat{\theta}^{ML} \) and the parametric empirical Bayes method-of-moment estimator \( \hat{\theta}^{MM} \). The empirical Bayes maximum likelihood estimator \( \hat{\theta}^{ML} \) is given by

\[
\hat{\theta}_i^{ML} = \frac{\tau_i}{\tau_i + \hat{\gamma}_i^{ML}} \cdot Y_i + \frac{\hat{\gamma}_i^{ML}}{\tau_i + \hat{\gamma}_i^{ML}} \cdot \hat{\mu}^{ML},
\]

where \((\hat{\gamma}_i^{ML}, \hat{\mu}^{ML})\) maximizes the marginal likelihood of \( Y_i \):

\[
(\hat{\gamma}_i^{ML}, \hat{\mu}^{ML}) = \arg \max_{\gamma > 0, \mu} \prod_i \frac{\Gamma(\gamma\mu + \tau_i y_i)\Gamma(\gamma(1 - \mu) + (1 - y_i)\tau_i)\Gamma(\gamma)}{\Gamma(\tau_i)\Gamma(\gamma(1 - \mu))},
\]
where \( \mu = \alpha / (\alpha + \beta) \) and \( \gamma = \alpha + \beta \) as in Table 2. Likewise, the empirical Bayes method-of-moment estimator \( \hat{\theta}^{MM} \) is given by

\[
\hat{\theta}^{MM}_i = \frac{\tau_i}{\tau_i + \hat{\gamma}^{MM}_i} \cdot \bar{Y}_i + \frac{\hat{\gamma}^{MM}_i}{\tau_i + \hat{\gamma}^{MM}_i} \cdot \hat{\mu}^{MM},
\]

where

\[
\hat{\mu}^{MM} = \bar{Y} = \frac{1}{p} \sum_{i=1}^{p} Y_i,
\]

\[
\hat{\gamma}^{MM} = \frac{\bar{Y}(1 - \bar{Y}) \cdot \sum_{i=1}^{p} (1 - 1/\tau_i)}{[\sum_{i=1}^{p} \left(Y_i^2 - \bar{Y}/\tau_i - \bar{Y}^2(1 - 1/\tau_i)\right)]^+}.
\]

There are in total four different simulation setups in which we study the
Fig 2. Comparison of the risks of different shrinkage estimators for the logistic case.

four different estimators. In addition, in each case, we also calculate the oracle risk "estimator" $\tilde{\theta}^{OR}$, defined as

$$
\tilde{\theta}^{OR}_i = \frac{\tau_i}{\tau_i + \tilde{\gamma}^{OR}} Y_i + \frac{\tilde{\gamma}^{OR}}{\tau_i + \tilde{\gamma}^{OR}} \tilde{\mu}^{OR},
$$

where

$$
(\tilde{\gamma}^{OR}, \tilde{\mu}^{OR}) = \arg \min_{\gamma \geq 0, \mu} R_p(\theta, \tilde{\theta}^{\gamma, \mu})
$$

$$
= \arg \min_{\gamma \geq 0, \mu} \sum_{i=1}^p \frac{1}{p} E \left[ \left( \frac{\tau_i}{\tau_i + \gamma} Y_i + \frac{\gamma}{\tau_i + \gamma} \mu - \theta_i \right)^2 \right].
$$

Clearly the oracle risk estimator $\tilde{\theta}^{OR}$ cannot be used in practice, since it
depends on the unknown $\theta$, but it does provide a sensible lower bound of the risk achievable by any shrinkage estimator with the given parametric form.

**Example 1.** We generate $\tau_i \sim \text{Poisson}(3) + 2$ and $\theta_i \sim \text{Beta}(1, 1)$ independently, and draw $Y_i \sim \text{Bin}(\tau_i, \theta_i)/\tau_i$. The oracle estimator $\tilde{\theta}^{OR}$ is found to have $\tilde{\gamma}^{OR} = 2$ and $\tilde{\mu}^{OR} = 0.5$. The corresponding risk for the oracle estimator is numerically found to be $R_p(\theta, \tilde{\theta}^{OR}) \approx 0.0253$. The plot in Figure 4(a) shows the risks of the five shrinkage estimators as the sample size $p$ varies. It is seen that the performance of all four shrinkage estimators approaches that of the parametric oracle estimator, the “best estimator” one can hope to get under the parametric form. Note that the two empirical Bayes esti-
mators converges to the oracle estimator faster than the two URE shrinkage estimators. This is because the hierarchical distribution on \( i \) and \( j \) are exactly the one assumed by the empirical Bayes estimators. In contrast, the URE estimators do not make any assumption on the hierarchical distribution but still achieve rather competitive performance. When the sample size is moderately large, all four estimators well approach the limit given by the parametric oracle estimator.

Example 2. We generate \( \tau_i \sim \text{Poisson}(3) + 2 \) and \( \theta_i \sim \frac{1}{2}\text{Beta}(1,3) + \frac{1}{2}\text{Beta}(3,1) \) independently, and draw \( Y_i \sim \text{Bin}(\tau_i, \theta_i)/\tau_i \). In this example, \( \theta_i \) no longer comes from a beta distribution, but \( \theta_i \) and \( \tau_i \) are still independent. The oracle estimator is found to have \( \tilde{\gamma}^{\text{OR}} \approx 1.5 \) and \( \tilde{\mu}^{\text{OR}} = 0.5 \). The corresponding risk for the oracle estimator \( \tilde{\theta}^{\gamma,\theta} \) is \( R_p(\theta, \tilde{\theta}^{\text{OR}}) \approx 0.0248 \). The plot in Figure 4(b) shows the risks of the five shrinkage estimators as the sample size \( p \) varies. Again, as \( p \) gets large, the performance of all shrinkage estimators eventually approaches that of the oracle estimator. This observation indicates that the parametric form of the prior on \( \theta_i \) is not crucial as long as \( \tau_i \) and \( \theta_i \) are independent.

Example 3. We generate \( \tau_i \sim \text{Poisson}(3) + 2 \) and let \( \theta_i = 1/\tau_i \), and then we draw \( Y_i \sim \text{Bin}(\tau_i, \theta_i)/\tau_i \). In this case, there is a (negative) correlation between \( \theta_i \) and \( \tau_i \). The parametric oracle estimator is found to have \( \tilde{\gamma}^{\text{OR}} \approx 23.0898 \) and \( \tilde{\mu}^{\text{OR}} \approx 0.2377 \) numerically; the corresponding risk is \( R_p(\mu, \tilde{\theta}^{\text{OR}}) \approx 0.0069 \). The plot in Figure 4(c) shows the risks of the five shrinkage estimators as functions of the sample size \( p \). Unlike the previous examples, the two empirical Bayes estimators no longer converge to the parametric oracle estimator, i.e., the limit of their risk (as \( p \to \infty \)) is strictly above the risk of the parametric oracle estimator. On the other hand, the risk of the parametric URE estimator \( \tilde{\theta}^{PM} \) still converges to the risk of the parametric oracle estimator. It is interesting to note that the limiting risk of the semi-parametric URE estimators \( \tilde{\theta}^{SM} \) is actually strictly smaller than the risk of the parametric oracle estimator (although the difference between the two is not easy to spot due to the scale of the plot).

Example 4. We generate \((\tau_i, \theta_i, Y_i)\) as follows. First, we draw \( I_i \) from \( \text{Bernoulli}(1/2) \), and then generate \( \tau_i \sim I_i \cdot \text{Poisson}(10) + (1-I_i) \cdot \text{Poisson}(1) + 2 \) and \( \theta_i \sim I_i \cdot \text{Beta}(1,3) + (1-I_i) \cdot \text{Beta}(3,1) \). Given \((\tau_i, \theta_i)\), we draw \( Y_i \sim \text{Bin}(\tau_i, \theta_i)/\tau_i \). In this example, there exist two groups in the data (indexed by \( I_i \)). It thus serves to test the different estimators in the presence of grouping. The parametric oracle estimator is found to have \( \tilde{\gamma}^{\text{OR}} \approx 0.3108 \) and \( \tilde{\mu}^{\text{OR}} \approx 2.0426 \); the corresponding risk is \( R_p(\mu, \tilde{\theta}^{\text{OR}}) \approx 0.0201 \). Figure
4(d) plots the risks of the five shrinkage estimators versus the sample size $p$. The two empirical Bayes estimators clearly encounter much greater risk than the URE estimators, and the limiting risks of the two empirical Bayes estimators are significantly larger than the risk of the parametric oracle estimator. The risk of the parametric URE estimator $\hat{\theta}^{PM}$ converges to that of the parametric oracle estimator. It is quite noteworthy that the semi-parametric URE estimator $\hat{\theta}^{SM}$ achieves a significant improvement over the parametric oracle one.

![Graphs showing risk comparison](image)

(a) $\tau \sim \text{Poisson}(3) + 2$, $\theta \sim \text{Beta}(1, 1)$ independently; $Y \sim \text{Bin}(\tau, \theta)/\tau$.

(b) $\tau \sim \text{Poisson}(3) + 2$, $\theta \sim \frac{1}{2} \cdot \text{Beta}(1, 3) + \frac{1}{2} \cdot \text{Beta}(3, 1)$ independently; $Y \sim \text{Bin}(\tau, \theta)/\tau$.

(c) $\tau \sim \text{Poisson}(3) + 2$, $\theta = 1/\tau$, $Y \sim \text{Bin}(\tau, \theta)/\tau$.

(d) $I \sim \text{Bern}(1/2)$, $\tau \sim I \cdot \text{Poisson}(10) + (1 - I) \cdot \text{Poisson}(1) + 2$, $\theta \sim I \cdot \text{Beta}(1, 3) + (1 - I) \cdot \text{Beta}(3, 1)$; $Y \sim \text{Bin}(\tau, \theta)/\tau$.

**Fig 4.** Comparison of the risks of shrinkage estimators in the Beta-Binomial hierarchical models.
5.2.2. Poisson-Gamma hierarchical model. For Poisson observations $Y_i \overset{i.i.d.}{\sim} \text{Poisson}(\tau_i \theta_i)/\tau_i$, the conjugate prior is $\theta_i \overset{i.i.d.}{\sim} \Gamma(\alpha, \lambda)$. Like in the previous subsection, we compare five estimators: the empirical Bayes maximum likelihood estimator, the empirical Bayes method-of-moment estimator, the parametric and semi-parametric URE estimators and the parametric oracle “estimator” (5.2). The empirical Bayes maximum likelihood estimator $\hat{\theta}^{ML}$ is given by

$$
\hat{\theta}^{ML}_i = \frac{\tau_i}{\tau_i + \hat{\tau}^{ML}} \cdot Y_i + \frac{\hat{\gamma}^{ML}}{\tau_i + \hat{\gamma}^{ML}} \cdot \hat{\mu}^{ML},
$$

where $(\hat{\gamma}^{ML}, \hat{\mu}^{ML})$ maximizes the marginal likelihood of $Y_i$:

$$(\hat{\gamma}^{ML}, \hat{\mu}^{ML}) = \arg \max_{\gamma \geq 0, \mu} \prod_i \frac{\gamma^{\gamma \mu} \Gamma(\gamma \mu + \gamma \mu_i)}{(\tau_i + \gamma)^{\gamma \mu_i + \gamma \mu} \Gamma(\gamma \mu)},$$

where $\mu = \alpha \lambda$ and $\gamma = 1/\lambda$ as in Table 2. The empirical Bayes method-of-moment estimator $\hat{\theta}^{MM}$ is given by

$$
\hat{\theta}^{MM}_i = \frac{\tau_i}{\tau_i + \hat{\tau}^{MM}} \cdot Y_i + \frac{\hat{\gamma}^{MM}}{\tau_i + \hat{\gamma}^{MM}} \cdot \hat{\mu}^{MM},
$$

where

$$
\hat{\mu}^{MM} = \bar{Y} = \frac{1}{p} \sum_{i=1}^p Y_i
$$

and

$$
\hat{\gamma}^{MM} = \frac{p \cdot \bar{Y}}{\left[ \sum_{i=1}^p (Y_i^2 - \bar{Y}/\tau_i - \bar{Y}^2) \right]^+}.
$$

We consider four different simulation settings.

**Example 1.** We generate $\tau_i \sim \text{Poisson}(3) + 2$ and $\theta_i \sim \Gamma(1, 1)$ independently, and draw $Y_i \sim \text{Poisson}(\tau_i \theta_i)/\tau_i$. The plot in Figure 5(a) shows the risks of the five shrinkage estimators as the sample size $p$ varies. Clearly, the performance of all shrinkage estimators approaches that of the parametric oracle estimator. As in the beta-binomial case, the two empirical Bayes estimators converges to the oracle estimator faster than the two URE shrinkage estimators. Again, this is because the hierarchical distribution on $\tau_i$ and $\theta_i$ are exactly the one assumed by the empirical Bayes estimators. The URE estimators, without making any assumption on the hierarchical distribution, still achieve rather competitive performance.

**Example 2.** We generate $\tau_i \sim \text{Poisson}(3) + 2$ and $\theta_i \sim \text{Unif}(0, 1)$ independently, and draw $Y_i \sim \text{Poisson}(\tau_i \theta_i)/\tau_i$. In this setting, $\theta_i$ no longer comes
from a gamma distribution, but \( \theta_i \) and \( \tau_i \) are still independent. The plot in Figure 5(b) shows the risks of the five shrinkage estimators as the sample size \( p \) varies. As \( p \) gets large, the performance of all shrinkage estimators eventually approaches that of the oracle estimator. Like in the beta-binomial case, the picture indicates that the parametric form of the prior on \( \theta_i \) is not crucial as long as \( \tau_i \) and \( \theta_i \) are independent.

**Example 3.** We generate \( \tau_i \sim \text{Poisson}(3) + 2 \) and let \( \theta_i = 1/\tau_i \), and then we draw \( Y_i \sim \text{Poisson}(\tau_i\theta_i)/\tau_i \). In this setting, there is a (negative) correlation between \( \theta_i \) and \( \tau_i \). The plot in Figure 5(c) shows the risks of the five shrinkage estimators as functions of the sample size \( p \). Unlike the previous two examples, the two empirical Bayes estimators no longer converge to the parametric oracle estimator – the limit of their risk is strictly above the risk of the parametric oracle estimator. The risk of the parametric URE estimator \( \hat{\theta}^{PM} \), on the other hand, still converges to the risk of the parametric oracle estimator. The limiting risk of the semi-parametric URE estimator \( \hat{\theta}^{SM} \) is actually strictly smaller than the risk of the parametric oracle estimator (although it is not easy to spot it on the plot).

**Example 4.** We generate \((\tau_i, \theta_i)\) by first drawing \( I_i \sim \text{Bernoulli}(1/2) \) and then \( \tau_i \sim I_i \cdot \text{Poisson}(10) + (1 - I_i) \cdot \text{Poisson}(1) + 2 \) and \( \theta_i \sim I_i \cdot \Gamma(1, 1) + (1 - I_i) \cdot \Gamma(5, 1) \). With \((\tau_i, \theta_i)\) obtained, we draw \( Y_i \sim \text{Poisson}(\tau_i\theta_i)/\tau_i \). This setting tests the case that there is grouping in the data. Figure 5(d) plots the risks of the five shrinkage estimators versus the sample size \( p \). It is seen that the two empirical Bayes estimators have the largest risk, and that the parametric URE estimator \( \hat{\theta}^{PM} \) achieves the risk of the parametric oracle estimator in the limit. The semi-parametric URE estimator \( \hat{\theta}^{SM} \) notably outperforms the parametric oracle estimator, when \( p > 100 \).

6. **Application to the prediction of batting average.** In this section, we apply the URE shrinkage estimators to a baseball data set, collected and discussed in Brown (2008). This data set consists of the batting records for all the Major League Baseball players in the season of 2005. Following Brown (2008), the data are divided into two half seasons; the goal is to use the data from the first half season to predict the players’ batting average in the second half season. The prediction can then be compared against the actual record of the second half season. The performance of different estimators can thus be directly evaluated.

For each player, let the number of at-bats be \( N \) and the successful number of batting be \( H \); we then have

\[ H_{ij} \sim \text{Binomial}(N_{ij}, p_j), \]
where $i = 1, 2$ is the season indicator, $j = 1, 2, \cdots$, is the player indicator, and $p_j$ corresponds to the player’s hitting ability. Let $Y_{ij}$ be the observed proportion:

$$Y_{ij} = H_{ij}/N_{ij}.$$  

For this binomial setup, we apply our method to obtain the semi-parametric URE estimators $\hat{\theta}_{SM}$ and $\hat{\theta}_{SG}$, defined in (2.5) and (2.7) respectively, and the parametric URE estimators $\hat{\theta}_{PM}$ and $\hat{\theta}_{PG}$, defined in (4.2) and (4.4) respectively.

To compare the prediction accuracy of different estimators, we note that most shrinkage estimators in the literature assume normality of the under-
lying data. Therefore, for sensible evaluation of different methods, we can apply the following variance-stabilizing transformation as discussed in Brown (2008):

\[ X_{ij} = \arcsin \sqrt{\frac{H_{ij} + 1/4}{N_{ij} + 1/2}}, \]

which gives

\[ X_{ij} \sim N(\theta_j, \frac{1}{4N_{ij}}), \quad \theta_j = \arcsin(\sqrt{\hat{p}_j}). \]

To evaluate an estimator \( \hat{\theta} \) based on the transformed \( X_{ij} \), we measure the total sum of squared prediction errors (TSE) as

\[ \text{TSE}(\hat{\theta}) = \sum_j (X_{2j} - \hat{\theta}_j)^2 - \sum_j \frac{1}{4N_{2j}}. \]

To conform to the above transformation (as used by most shrinkage estimators), we calculate \( \hat{\theta}_j = \arcsin(\sqrt{\hat{p}_j}) \), where \( \hat{p}_j \) is a URE estimator of the binomial probability, so that the TSE of the URE estimators can be calculated and compared with other (normality based) shrinkage estimators.

Table 3 below summarizes the numerical results of our URE estimators with a collection of competing shrinkage estimators. The values reported are the ratios of the error of a given estimator to that of the benchmark naive estimator, which simply uses the first half season \( X_{1j} \) to predict the second half \( X_{2j} \). All shrinkage estimators are applied three times — to all the baseball players, the pitchers only, and the non-pitchers only. The first group of shrinkage estimators in Table 3 are the classical ones based on normal theory: two empirical Bayes methods (applied to \( X_{1j} \)), the grand mean and the extended James-Stein estimator (5.1). The second group includes a number of more recently developed methods: the nonparametric shrinkage methods in Brown and Greenshtein (2009), the binomial mixture model in Muralidharan (2010) and the weighted least squares and general maximum likelihood estimators (with or without the covariate of at-bats effect) in Jiang and Zhang (2009, 2010). The numerical results for these methods are from Brown (2008), Muralidharan (2010) and Jiang and Zhang (2009, 2010). The last group corresponds to the results from our binomial URE methods: the first two are the parametric methods and the last two are the semi-parametric ones.

It is seen that our URE shrinkage estimators, especially the semi-parametric ones, achieve very competitive prediction result among all the estimators. We think the primary reason is that the baseball data contain unique features that violate the underlying assumptions of the classical empirical Bayes
methods. Both the normal prior assumption and the implicit assumption of
the uncorrelatedness between the binomial probability $p$ and the sample size
$\tau$ are not justified here. To illustrate the last point, we present a scatter plot
of $\log_{10}$ (number of at bats) versus the observed batting average $y$ for the
non-pitcher group in Figure 6.

<table>
<thead>
<tr>
<th></th>
<th>ALL</th>
<th>Pitchers</th>
<th>NonPitchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>naive</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Grand mean $\hat{X}_1$</td>
<td>0.852</td>
<td>0.127</td>
<td>0.378</td>
</tr>
<tr>
<td>Parametric EB-MM</td>
<td>0.593</td>
<td>0.129</td>
<td>0.387</td>
</tr>
<tr>
<td>Parametric EB-ML</td>
<td>0.902</td>
<td>0.117</td>
<td>0.398</td>
</tr>
<tr>
<td>Extended James-Stein</td>
<td>0.525</td>
<td>0.164</td>
<td>0.359</td>
</tr>
<tr>
<td>Nonparametric EB</td>
<td>0.508</td>
<td>0.212</td>
<td>0.372</td>
</tr>
<tr>
<td>Binomial mixture</td>
<td>0.588</td>
<td>0.156</td>
<td>0.314</td>
</tr>
<tr>
<td>Weighted least square (Null)</td>
<td>1.074</td>
<td>0.127</td>
<td>0.468</td>
</tr>
<tr>
<td>Weighted generalized MLE (Null)</td>
<td>0.306</td>
<td>0.173</td>
<td>0.326</td>
</tr>
<tr>
<td>Weighted least square (AB)</td>
<td>0.537</td>
<td>0.087</td>
<td>0.290</td>
</tr>
<tr>
<td>Weighted generalized MLE (AB)</td>
<td>0.301</td>
<td>0.141</td>
<td>0.261</td>
</tr>
<tr>
<td>Parametric URE $\hat{\theta}^{PG}$</td>
<td>0.515</td>
<td>0.105</td>
<td>0.278</td>
</tr>
<tr>
<td>Parametric URE $\hat{\theta}^{PM}$</td>
<td>0.421</td>
<td>0.105</td>
<td>0.276</td>
</tr>
<tr>
<td>Semi-parametric URE $\hat{\theta}^{SG}$</td>
<td>0.414</td>
<td>0.045</td>
<td>0.259</td>
</tr>
<tr>
<td>Semi-parametric URE $\hat{\theta}^{SM}$</td>
<td>0.422</td>
<td>0.041</td>
<td>0.273</td>
</tr>
</tbody>
</table>

Table 3
Prediction errors of batting averages for the baseball data.

7. Summary. In this paper, we develop a general theory of URE shrinkage estimation in family of distributions with quadratic variance function. We first discuss a class of semi-parametric URE estimator and establish their optimality property. Two specific cases are then carefully studied: the location-scale family and the natural exponential families with quadratic variance function. In the latter case, we also study a class of parametric URE estimators, whose forms are derived from the classical conjugate hierarchical model. We show that each URE shrinkage estimator is asymptotically optimal in its own class and their asymptotic optimality do not depend on the specific distribution assumptions, and more importantly, do not depend on the implicit assumption that the group mean $\theta$ and the sample size $\tau$ is uncorrelated, which underlies many classical shrinkage estimators. The URE estimators are evaluated in comprehensive simulation studies and one real data set. It is found that the URE estimators offer numerically superior performance compared to the classical empirical Bayes and many other com-
Fig 6. Scatter plot of $\log_{10}$ (number of at bats) versus observed batting average

peting shrinkage estimators. The semi-parametric URE estimators appear to be particularly competitive.

It is worth emphasizing that the optimality properties of the URE estimators is not in contradiction with well established results on the non-existence of Stein’s paradox in simultaneous inference problems with finite sample space (Gutmann, 1982), since the results we obtained here are asymptotic ones when $p$ approaches infinity. A question that naturally arises here is then how large $p$ needs to be for the URE estimators to become superior compared with their competitors. Even though we did not develop a formal finite-sample theory for such comparison, our comprehensive simulation indicates that $p$ usually does not need to be large — $p$ can be as small as 100 — for the URE estimators to achieve competitive performance.

The theory here extends the one on the normal hierarchical models in Xie, Kou and Brown (2012). There are three critical features that make the generalization of previous results possible here: (i) the use of quadratic risk; (ii) the linear form of the shrinkage estimator and (iii) the quadratic variance function of the distribution. The three features together guarantee the existence of an unbiased risk estimate. For the hierarchical models where
an unbiased risk estimate does not exist, similar idea can still be applied to some estimate of risk, e.g., the bootstrap estimate (Efron, 2004). However, a theory is in demand to justify the performance of the resulting shrinkage estimators.

It would also be an important area of research to study confidence intervals for the URE estimators obtained here. Understanding whether the optimality of the URE estimators implies any optimality property of the estimators of the hyper-parameters under certain conditions is an interesting related question. However, such topics are out of scope for the current paper and we will need to address them in future research.

Appendix: Proofs. Proof of Theorem 2.1. Throughout this proof, when a supremum is taken, it is over \( b_i \in [0, 1], |\mu| \leq \max_i |Y_i| \) and Requirement (MON), unless explicitly stated otherwise. Since

\[
\text{URE}(b, \mu) - l(\theta, \hat{\theta}^{b, \mu})
\]

\[
= \frac{1}{p} \sum_{i=1}^{p} (1 - 2b_i) \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) + \frac{2}{p} \sum_{i=1}^{p} b_i(Y_i - \theta_i)(\theta_i - \mu),
\]

it follows that

\[
\sup \left| \text{URE}(b, \mu) - l(\theta, \hat{\theta}^{b, \mu}) \right| \leq \frac{1}{p} \left| \sum_{i=1}^{p} \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \right| + \frac{2}{p} \sup \left| \sum_{i=1}^{p} b_i \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \right|
\]

\[
\phantom{=} + \frac{2}{p} \sup \left| \sum_{i=1}^{p} b_i(Y_i - \theta_i)(\theta_i - \mu) \right|.
\]

(A.1)

For the first term on the right hand side, we note that

\[
\frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 = - \frac{\tau_i}{\tau_i + \nu_2} \left( Y_i^2 - EY_i^2 \right) + (2\theta_i - \frac{\nu_1}{\tau_i + \nu_2})(Y_i - \theta_i).
\]

Thus,

\[
E \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \leq 2 \left( \frac{\tau_i}{\tau_i + \nu_2} \right)^2 \text{Var}(Y_i^2) + (2\theta_i - \frac{\nu_1}{\tau_i + \nu_2})^2 \text{Var}(Y_i)
\]

\[
\leq 2 \left( \frac{\tau_i}{\tau_i + \nu_2} \right)^2 \text{Var}(Y_i^2) + 16\theta_i^2 \text{Var}(Y_i) + 4 \left( \frac{\nu_1}{\tau_i + \nu_2} \right)^2 \text{Var}(Y_i).
\]
It follows that by Conditions (A)-(D)

\[(A.2) \quad \frac{1}{p} \sum_{i=1}^{p} \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \to 0 \text{ in } L^2 \text{ as } p \to \infty. \]

For the term \( \frac{2}{p} \sup \sum_i b_i \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \) in (A.1), without loss of generality, let us assume \( \tau_1 \leq \cdots \leq \tau_p \); we then know from Requirement (MON) that \( b_1 \geq \cdots \geq b_2 \). As in Lemma 2.1 in Li (1986), we observe that

\[
\sup \frac{2}{p} \sum_{i=1}^{p} b_i \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) = \left( \max_{1 \leq j \leq p} \frac{2}{p} \sum_{i=1}^{j} \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \right). 
\]

Let \( M_j = \sum_{i=1}^{j} \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \). Then \( \{M_j; j = 1, 2, \ldots\} \) forms a martingale. The \( L^p \) maximum inequality implies

\[
E(\max_{1 \leq j \leq p} M_j^2) \leq 4E(M_p^2) = 4\sum_{i=1}^{p} E \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right)^2,
\]

which implies by (A.2) that

\[(A.3) \quad \sup \frac{2}{p} \sum_{i=1}^{p} b_i \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \to 0 \text{ in } L^2 \text{ as } p \to \infty. \]

For the last term \( \frac{2}{p} \sup |\sum_i b_i(Y_i - \theta_i)(\theta_i - \mu)| \) in (A.1), we note that

\[
\frac{1}{p} \sum_{i=1}^{p} b_i(Y_i - \theta_i)(\theta_i - \mu) = \frac{1}{p} \sum_{i=1}^{p} b_i\theta_i(Y_i - \theta_i) - \frac{\mu}{p} \sum_{i=1}^{p} b_i(Y_i - \theta_i).
\]

Using the same argument as in the proof of (A.3), we can show that

\[
\sup \frac{1}{p} \left| \sum_i b_i\theta_i(Y_i - \theta_i) \right| \to 0 \text{ in } L^2,
\]

\[
E \left( \sup \left| \sum_i b_i(Y_i - \theta_i)^2 \right| \right) = O(p).
\]
Applying Condition (E) and the Cauchy-Schwarz inequality, we obtain

\[ \frac{1}{p} E \left( \sup_{i} | \mu \sum_{i=1}^{p} b_{i}(Y_{i} - \theta_{i})| \right) \]

\[ = \frac{1}{p} E \left( \max_{1 \leq i \leq p} |Y_{i}| \cdot \sup_{i} | \sum_{i} b_{i}(Y_{i} - \theta_{i})| \right) \]

\[ \leq \frac{1}{p} \left( E \left( \max_{1 \leq i \leq p} Y_{i}^{2} \right) \cdot E \left( \sup_{i} | \sum_{i} b_{i}(Y_{i} - \theta_{i})|^{2} \right) \right)^{1/2} \]

\[ = O(p^{(1-\varepsilon)/2-1}) = O(p^{-\varepsilon/2}). \]

Therefore,

\[ \frac{1}{p} \sup_{i} | \sum_{i=1}^{p} b_{i}(Y_{i} - \theta_{i})| \to 0 \text{ in } L^{1}. \]

This completes the proof, since each term on the right hand side of (A.1) converges to zero in \( L^{1} \). \( \square \)

**Proof of Theorem 2.2.** Throughout this proof, when a supremum is taken, it is over \( b_{i} \in [0, 1] \), \( |\mu| \leq \max_{i} |Y_{i}| \) and Requirement (MON). Note that

\[ \text{URE}(\hat{b}^{SM}, \hat{\mu}^{SM}) \leq \text{URE}(\hat{b}, \hat{\mu}) \]

and we know from Theorem 2.1 that

\[ \sup | \text{URE}(b, \mu) - l_{p}(\theta, \hat{b}^{b, \mu}) | \to 0 \text{ in probability.} \]

It follows that for any \( \varepsilon > 0 \)

\[ P \left( l_{p}(\theta, \hat{b}^{SM}) \geq l_{p}(\theta, \hat{b}^{b, \mu}) + \varepsilon \right) \]

\[ \leq P \left( l_{p}(\theta, \hat{b}^{SM}) - \text{URE}(\hat{b}^{SM}, \hat{\mu}^{SM}) \geq l_{p}(\theta, \hat{b}^{b, \mu}) - \text{URE}(\hat{b}, \hat{\mu}) + \varepsilon \right) \]

\[ \leq P \left( |l_{p}(\theta, \hat{b}^{SM}) - \text{URE}(\hat{b}^{SM}, \hat{\mu}^{SM})| \geq \frac{\varepsilon}{2} \right) \]

\[ + P \left( |l_{p}(\theta, \hat{b}^{b, \mu}) - \text{URE}(\hat{b}, \hat{\mu})| \geq \frac{\varepsilon}{2} \right) \to 0. \]

Next, to show that

\[ \limsup_{p \to \infty} \left[ R_{p}(\theta, \hat{b}^{SM}) - R_{p}(\theta, \hat{b}^{b, \mu}) \right] \leq 0, \]
we note that
\[
l_p(\theta, \hat{\theta}^{SM}) - l_p(\theta, \hat{\theta}^{b, \hat{\mu}}) = \left( l_p(\theta, \hat{\theta}^{SM} - URE(\hat{b}^{SM}, \hat{\mu}^{SM})) + (URE(\hat{b}^{SM}, \hat{\mu}^{SM}) - URE(\hat{b}, \hat{\mu})) \right) + (URE(\hat{b}, \hat{\mu}) - l_p(\theta, \hat{\theta}^{b, \hat{\mu}})) \leq 2 \sup |URE(\hat{b}, \hat{\mu}) - l_p(\theta, \hat{\theta}^{b, \hat{\mu}})|.
\]

Theorem 2.1 then implies that
\[
\limsup_{p \to \infty} \left[ R(\theta, \hat{\theta}^{SM}) - R(\theta, \hat{\theta}^{b, \hat{\mu}}) \right] \leq 0. \quad \Box
\]

**Proof of Theorem 2.3.** Throughout this proof, when a supremum is taken, it is over \( b_i \in [0, 1] \) and Requirement (MON). Since
\[
URE^G(b) - l_p(\theta, \hat{\theta}^{b, \bar{Y}}) = \frac{1}{p} \sum_{i=1}^{p} (1 - 2(1 - \frac{1}{p})b_i) \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) + \frac{2}{p} \sum_{i=1}^{p} b_i \left( \theta_i(Y_i - \theta_i) + \frac{1}{p}(Y_i - \theta_i)^2 - (Y_i - \theta_i)\bar{Y} \right),
\]
it follows that
\[
\sup |URE^G(b) - l_p(\theta, \hat{\theta}^{b, \bar{Y}})| \leq \frac{1}{p} \left| \sum_i \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \right| + \frac{2}{p} \left( 1 - \frac{1}{p} \right) \sup \left| \sum_i b_i \left( \frac{V(Y_i)}{\tau_i + \nu_2} - (Y_i - \theta_i)^2 \right) \right| + \frac{2}{p} \sup \left| \sum_i b_i(Y_i - \theta_i) \right| + \frac{2}{p^2} \sum_i (Y_i - \theta_i)^2 + \frac{2}{p} |\bar{Y}| \cdot \sup \left| \sum_i b_i(Y_i - \theta_i) \right| \quad (A.4)
\]

We have already shown in the proof of Theorem 2.1 that the first three terms on the right hand side converge to zero in \( L^2 \). It only remains to manage the last two terms.
\[
E \left( \frac{2}{p^2} \sum_i (Y_i - \theta_i)^2 \right) = \frac{2}{p^2} \sum_i Var(Y_i) \to 0
\]
by regularity condition (A).

\[ \frac{1}{p} E \left( |\bar{Y}| \cdot \sup \left\{ \sum_i b_i (Y_i - \theta_i) \right\} \right) \leq \frac{1}{p} E \left( \max_{1 \leq i \leq p} |Y_i| \cdot \sup \left\{ \sum_i b_i (Y_i - \theta_i) \right\} \right) \to 0, \]

as was shown in the proof of Theorem 2.1. Therefore, the last two terms of (A.4) converge to zero in \( L^1 \), and this completes the proof. □

**Proof of Theorem 2.4.** With Theorem 2.3 established, the proof is almost identical to that of Theorem 2.2. □

**Proof of Lemma 3.1.** It is straightforward to check that (i)-(iii) imply conditions (A)-(D) in Section 2, so we only need to verify condition (E).

Since \( Y_2^i = Z_2^i = i^2 + 2i + \frac{1}{2}i^2 \), we know that

(A.5) \[ \max_{1 \leq i \leq p} Y_2^i \leq \max_{1 \leq i \leq p} \frac{1}{\tau_i} \cdot \max_{1 \leq i \leq p} Z_2^i \max_{1 \leq i \leq p} \theta_i^2 + 2 \max_{1 \leq i \leq p} |\theta_i| \cdot \max_{1 \leq i \leq p} |Z_i|. \]

(i) and (ii) imply \( \max_{1 \leq i \leq p} 1/\tau_i = O(p^{1/2}) \) and \( \max_{1 \leq i \leq p} |\theta_i|/\sqrt{\tau_i} = O(p^{1/2}). \) (iii) gives \( \max_{1 \leq i \leq p} \theta_i^2 = O(p^{2(2+\epsilon)}). \) We next bound \( E(\max_{1 \leq i \leq p} |Z_i|^k) \) for \( k = 1, 2, \) (iv) implies that for \( k = 1, 2, \)

\[ E(\max_{1 \leq i \leq p} |Z_i|^k) \]

\[ = \int_0^\infty kt^{k-1} P(\max_{1 \leq i \leq p} |Z_i| > t) dt \]

\[ \leq \int_0^\infty kt^{k-1}(1 - (1 - Dt^{-\alpha})^p) dt \]

\[ = \int_0^{p^{1/\alpha}} kt^{k-1}(1 - (1 - Dt^{-\alpha})^p) dt + \int_{p^{1/\alpha}}^\infty kt^{k-1}(1 - (1 - Dt^{-\alpha})^p) dt \]

\[ = O(p^{k/\alpha}) + p^{k/\alpha} \int_1^\infty k z^{k-1} \left( 1 - (1 - \frac{1}{p} Dz^{-\alpha})^p \right) dz, \]

where a change of variable \( z = t/p^{1/\alpha} \) is applied. We know by the monotone convergence theorem that for \( k \geq 1, \) as \( p \to \infty, \)

\[ \int_1^\infty z^{k-1} \left( 1 - (1 - \frac{1}{p} Dz^{-\alpha})^p \right) dz \to \int_1^\infty z^{k-1} (1 - \exp(-Dz^{-\alpha})) dz < \infty. \]

It then follows that

\[ E(\max_{1 \leq i \leq p} |Z_i|^k) = O(p^{k/\alpha}), \text{ for } k = 1, 2. \]
Taking it back to (A.5) gives

$$E(\max_{1 \leq i \leq p} Y_i^2) = O(p^{1/2 + 2/\alpha}) + O(p^{2/2+\varepsilon}) + O(p^{1/2+1/\alpha}),$$

which verifies condition (E). □

To prove Lemma 4.1, we need the following lemma.

**Lemma A.1.** Let $Y_i$ be independent from one of the six NEF-QVF. Then Condition (B) in Section 2 and

- $(F)$ $\limsup_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} |\theta_i|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$;
- $(G)$ $\limsup_{p \to \infty} \frac{1}{p} \sum_{i=1}^{p} \text{Var}(Y_i) < \infty$;
- $(H)$ $\sup_{i} \text{skew}(Y_i) = \sup_{i} \frac{\nu_i + 2\nu_2 \theta_i}{(\nu_0 + \nu_1 \theta_1 + \nu_2 \theta_2)^{1/2}} < \infty$.

imply Condition (E).

**Proof of Lemma A.1.** Let us denote $\sigma_i^2 = \text{Var}(Y_i)$. We can write $Y_i = \sigma_i Z_i + \theta_i$, where $Z_i$ are independent with mean zero and variance one. It follows from $Y_i^2 = \sigma_i^2 Z_i^2 + \theta_i^2 + 2\sigma_i \theta_i Z_i$ that

$$\max_{1 \leq i \leq p} Y_i^2 \leq \max_{i} \sigma_i^2 \cdot \max_{i} Z_i^2 + \max_{i} \theta_i^2 + 2 \max_{i} \sigma_i |\theta_i| \cdot \max_{i} |Z_i|.$$

Condition (B) implies $\max_{i} \sigma_i^2 \theta_i^2 \leq \sum_{i} \sigma_i^2 \theta_i^2 = O(p)$. Thus, $\max_{i} \sigma_i |\theta_i| = O(p^{1/2})$. Similarly, Condition (G) implies that $\max_{i} \sigma_i^2 = O(p^{1/2})$. Condition (F) implies that $\max_{i} |\theta_i|^{2+\varepsilon} \leq \sum_{i} |\theta_i|^{2+\varepsilon} = O(p)$, which gives $\max_{i} \theta_i^2 = O((p^{2/2+\varepsilon}))$. If we can show that

$$E(\max_{1 \leq i \leq p} |Z_i|) = O(\log p), \quad E(\max_{1 \leq i \leq p} Z_i^2) = O(\log^2 p),$$

then we establish (E), since

$$E(\max_{1 \leq i \leq p} Y_i^2) = O(p^{1/2 \log^2 p} + p^{2/(2+\varepsilon)} + p^{1/2 \log p}) = O(p^{2/(2+\varepsilon^*)}),$$

where $\varepsilon^* = \min(\varepsilon, 1)$. To prove (A.7), we begin from

$$E(\max_{i} |Z_i|) = k \int_{0}^{\infty} t^{k-1} P(\max_{i} |Z_i| > t) dt, \quad \text{for all } k > 0.$$

The large deviation results for NEF-QVF in Morris (1982, Section 9) and Condition (H) (i.e., $Y_i$ have bounded skewness) imply that for all $t > 1$, the
tail probabilities $P(|Z_i| > t)$ are uniformly bounded exponentially: there exists a constant $c_0 > 0$ such that

$$P(|Z_i| > t) \leq e^{-c_0 t} \text{ for all } i.$$  

Taking it into (A.8), we have

$$E(\max_{1 \leq i \leq p} |Z_i|^k) \leq \int_0^\infty k t^{k-1} (1 - (1 - e^{-c_0 t})^p) dt$$

$$= \int_0^{\log p/c_0} k t^{k-1} (1 - (1 - e^{-c_0 t})^p) dt$$

$$+ \int_{\log p/c_0}^\infty k t^{k-1} (1 - (1 - e^{-c_0 t})^p) dt$$

(A.9) \hspace{1cm} = O(\log^k p) + \int_0^\infty k(z + 1/c_0) k^{k-1} \left(1 - (1 - \frac{1}{p} e^{-c_0 z})^p\right) dz,$$

where in the last line a change of variable $z = t - \log p/c_0$ is applied. We know by the monotone convergence theorem that for $k \geq 1$, as $p \to \infty$,

$$\int_0^\infty z^{k-1} \left(1 - (1 - \frac{1}{p} e^{-c_0 z})^p\right) dz \to \int_0^\infty z^{k-1} (1 - \exp(-e^{-c_0 z})) dz < \infty.$$

It then follows from (A.9) that

$$E(\max_{1 \leq i \leq p} |Z_i|^k) = O(\log^k p), \text{ for } k = 1, 2,$$

which completes our proof. \[\square\]

**Proof of Lemma 4.1.** We go over the five distributions one by one.

1. **Binomial.** Since $\theta_i = p_i$, $\mathbb{V}ar(Y_i) = p_i(1 - p_i)/n_i$, and $\mathbb{V}ar(Y_i^2) \leq EY_i^4 \leq 1$, it is straightforward to verify that $n_i \geq 2$ for all $i$ guarantees Conditions (A)-(E) in Section 2.

2. **Poisson.** $\mathbb{V}ar(Y_i) = \theta_i/\tau_i$, and $\mathbb{V}ar(Y_i^2) = (4\tau_i^2 \theta_i^3 + 6\tau_i \theta_i^2 + \theta_i)/\tau_i^2$. It is straightforward to verify that $\inf_i \tau_i > 0$, $\inf_i \tau_i \theta_i > 0$ and $\sum_i \theta_i^4 = O(p)$ imply Conditions (A)-(D) in Section 2 and Conditions (F)-(H) in Lemma A.1.

3. **Negative-binomial.** $\theta_i = \frac{p_i}{1-p_i}$, $\mathbb{V}ar(Y_i) = \frac{1}{n_i(1-p_i)^2} = \frac{1}{n_i} (\theta_i + \theta_i^2)$, so $\nu_0 = 0$, $\nu_1 = \nu_2 = 1$. $\mathbb{V}ar(Y_i^2) = \frac{1}{n_i^2(1-p_i)^4} (p_i + 4p_i^2 + 6n_i p_i^2 + p_i^3 + 4n_i p_i^3 + 4n_i^2 p_i^3)$. From these, we know that $\sum_{i=1}^p \mathbb{V}ar(Y_i) = \sum_{i=1}^p \frac{1}{n_i(1-p_i)^4} = O(\sum_i (\frac{p_i}{1-p_i})^4) = O(p)$, which verifies Conditions (A) and (G). $\sum_{i=1}^p \mathbb{V}ar(Y_i) \theta_i^2 = \sum_{i=1}^p \frac{1}{n_i p_i} (\frac{p_i}{1-p_i})^4 = O(\sum_i (\frac{p_i}{1-p_i})^4) = O(p)$, which verifies Condition (B). For
Condition (C), since $n_i \geq 1$ and $0 \leq p_i \leq 1$, we only need to verify that $\sum_i \frac{1}{n_i(1-p_i)^2}(p_i + 6n_i p_i^2) = O(p)$. This is true, since $\sum_i \frac{1}{n_i(1-p_i)^2}(p_i + 6n_i p_i^2) = \sum_i \left( \frac{1}{(n_i p_i)^2} \right)^4 + \frac{6}{(n_i p_i)^2} \left( \frac{p_i}{1-p_i} \right)^4 = O(\sum_i \left( (\frac{p_i}{1-p_i})^4 \right) = O(p)$. Condition (D) is automatically satisfied. For Condition (F), consider $\sum_i \left( \frac{p_i}{1-p_i} \right)^4 = O(p)$. For Condition (H), note that $\text{skew}(Y_i) = \frac{1}{\sqrt{n_i p_i + 1}} \leq 1 + 1/\sqrt{n_i p_i}$. Thus, $\sup_i \text{skew}(Y_i) < \infty$ by (i).

(4) Gamma. $\theta_i = \alpha \lambda_i$, $\text{Var}(Y_i) = \alpha \lambda_i^2 / \tau_i$, so $v_0 = v_1 = 0$, $v_2 = 1/\alpha$, $\text{skew}(Y_i) = 2 / \sqrt{\tau_i \alpha}$. $\text{Var}(Y_i^2) = \frac{2}{\tau_i} \lambda_i^4 (\alpha + \frac{1}{\tau_i}) (4 \alpha + \frac{6}{\tau_i})$. It is straightforward to verify that (i) and (ii) imply Conditions (A)-(D) in Section 2 and Conditions (F)-(H) in Lemma A.1.

(5) GHS. $\theta_i = \alpha \lambda_i$, $\text{Var}(Y_i) = \alpha (1 + \lambda_i^2) / \tau_i$, so $v_0 = v_1 = 0$, $v_2 = 1/\alpha$, $\text{skew}(Y_i) = 2 / \sqrt{\tau_i \alpha}$. $\text{Var}(Y_i^2) = \frac{2}{\tau_i} (1 + \lambda_i^2) (\alpha + \frac{1}{\tau_i}) (\frac{1}{\tau_i} + \frac{3}{\tau_i} \lambda_i^2 + 2 \alpha \lambda_i^4)$. It is then straightforward to verify that (i) and (ii) imply Conditions (A)-(D) in Section 2 and Conditions (F)-(H) in Lemma A.1. □

Proof of Theorem 4.6. We note that the set over which the supremum is taken is a subset of that of Theorem 2.1. The desired result thus automatically holds. □

Proof of Theorem 4.7. With Theorem 4.6 established, the proof is almost identical to that of Theorem 2.2. □

Proof of Theorem 4.8. We note that the set over which the supremum is taken is a subset of that of Theorem 2.3. The desired result thus holds. □

Proof of Theorem 4.9. The proof essentially follows the same steps in that of Theorem 2.2. □

References.