Nonparametric Empirical Bayes Estimation On Heterogeneous Data

Trambak Banerjee[1], Luella J. Fu[2], Gareth M. James[3], Wenguang Sun[3]

Abstract

The simultaneous estimation of many parameters based on data collected from corresponding studies is a key research problem that has received renewed attention in the high-dimensional setting. Many practical situations involve heterogeneous data where heterogeneity is captured by a nuisance parameter. Effectively pooling information across samples while correctly accounting for heterogeneity presents a significant challenge in large-scale estimation problems. We address this issue by introducing the "Nonparametric Empirical Bayes Structural Tweedie" (NEST) estimator, which efficiently estimates the unknown effect sizes and properly adjusts for heterogeneity via a generalized version of Tweedie's formula. For the normal means problem, NEST simultaneously handles the two main selection biases introduced by heterogeneity: one, the selection bias in the mean, which cannot be effectively corrected without also correcting for, two, selection bias in the variance. Our theoretical results show that NEST has strong asymptotic properties without requiring explicit assumptions about the prior. Extensions to other two-parameter members of the exponential family are discussed. Simulation studies show that NEST outperforms competing methods, with much efficiency gains in many settings. The proposed method is demonstrated on estimating the batting averages of baseball players and Sharpe ratios of mutual fund returns.

1 Introduction

Suppose that we are interested in estimating a vector of parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ based on the summary statistics $Y_1, \ldots, Y_n$ from n study units. The setting where $Y_i \mid \mu_i \sim N(\mu_i, \sigma^2)$ is the most well–known example, but the broader scope includes the compound estimation of Poisson parameters λ_i, Binomial parameters p_i, and other members of the exponential family.

In modern large-scale applications it is often of interest to perform simultaneous and selective inference (Benjamini and Yekutieli 2011, Berk et al. 2013, Weinstein et al. 2013), which has called for solving the compound estimation problem in new ways and for new purposes. For example, there has been recent work on how to construct valid simultaneous confidence intervals of η_i's after a selection procedure is applied (Lee et al. 2016). In multiple testing, as well as related ranking and selection problems, it is often desirable to incorporate estimates of the effect sizes η_i in the decision process to prioritize the selection of more scientifically meaningful hypotheses (Benjamini and Hochberg 1997, Sun and McLain 2012, He et al. 2015, Henderson and Newton 2016, Basu et al. 2017).

However, the simultaneous inference of thousands of means, or other parameters, is challenging because, as described in Efron (2011), the large scale of the problem introduces selection bias, wherein some data points are large merely by chance, causing traditional estimators to overestimate the corresponding means. Shrinkage estimation, exemplified by the seminal work of James and Stein (1961), has been widely used in simultaneous inference. There are several popular classes of methods, including linear shrinkage estimators (James and Stein 1961, Efron and Morris 1975, Berger 1976), non–linear thresholding–based estimators motivated by sparse priors (Donoho and Jonhstone 1994, Johnstone and Silverman 2004, Abramovich et al. 2006), and both Bayes or empirical Bayes estimators with unspecified priors (Brown and Greenshtein 2009, Jiang and Zhang 2009, Castillo and van der Vaart 2012). This article focuses on a class of estimators based on Tweedie's formula (Robbins 1956)[1]. The formula is an ele-

---

[1] Tweedie's formula appears even earlier in the astronomy literature (Dyson 1926, Eddington 1940), wherein Frank Dyson credits the formula to Sir Arthur Eddington.

gant shrinkage estimator, for distributions from the exponential family, that has recently received renewed interest (Brown and Greenshtein 2009, Efron 2011, Koenker and Mizera 2014). Tweedie's formula is simple and intuitive, and its implementation under the f-modeling strategy for empirical Bayes estimation only requires estimating the marginal distribution of Y_i. This property is particularly appealing for large–scale estimation problems where such estimates can be easily constructed from the observed data. The resultant empirical Bayes estimator enjoys optimality properties (Brown and Greenshtein 2009) and delivers superior numerical performance. The work of Efron (2011) further convincingly demonstrates that Tweedie's formula provides an effective bias correction tool when estimating thousands of parameters simultaneously.

1.1   Issues with heterogeneous data

Most of the research in this area has been restricted to models of the form $f(y_i \mid \eta_i)$ where the distribution of Y_i is solely a function of η_i. In situations involving a nuisance parameter θ it is generally assumed to be known and identical for all Y_i. For example, homoscedastic Gaussian models of the form $Y_i \mid \mu_i, \sigma \overset{ind}{\sim} N(\mu_i, \sigma^2)$ involve a common nuisance parameter $\theta = \sigma^2$ for all i. However, in large-scale studies when the data are collected from heterogeneous sources, the nuisance parameters may vary over the n study units. Perhaps the most common example, and the setting we concentrate most on, involves heteroscedastic errors, where σ^2 varies over Y_i. Microarray data (Erickson and Sabatti 2005, Chiaretti et al. 2004), returns on mutual funds (Brown et al. 1992), and the state-wide school performance gaps (Sun and McLain 2012) are all instances of large-scale data where genes, funds, or schools have heterogeneous variances. Heteroscedastic errors also arise in analysis of variance and linear regression settings (Weinstein et al. 2018). Moreover, in compound binomial problems, heterogeneity arises through unequal sample sizes across different study units. Unfortunately, the conventional Tweedie's formula assumes identical nuisance parameters across study units and so cannot eliminate selection bias for heterogeneous data. Moreover, various works show that failing to account for heterogeneity leads to inefficient shrinkage estimators

(Weinstein et al. 2018), methods with invalid false discovery rates (Efron 2008, Cai and Sun 2009), unstable multiple testing procedures (Tusher et al. 2001) and suboptimal ranking and selection algorithms (Henderson and Newton 2016), exacerbating the replicability crisis in large-scale studies. Few methodologies are available to address this issue.

For Gaussian data, a common goal is to find the estimator, or make the decision, that minimizes the expected squared error loss. A plausible–seeming solution might be to scale each Y_i by its estimated standard deviation S_i so that a homoscedastic method could be applied to $X_i = Y_i/S_i$, before undoing the scaling on the final estimate of μ_i. Indeed, this is essentially the approach taken whenever we compute standardized test statistics, such as t–values and z–values. A similar standardization is performed in the Binomial setting when we compute $\hat{p}_i = X_i/m_i$, where X_i is the number of successes and m_i the number of trials. However, this approach, which disregards important structural information, can be highly inefficient. More advanced methods have been developed, but all suffer from various limitations. For instance, the methods proposed by Xie et al. (2012), Tan (2015), Jing et al. (2016), Kou and Yang (2017), and Zhang and Bhattacharya (2017) are designed for heteroscedastic data but assume a parametric Gaussian prior or semi-parametric Gaussian mixture prior, which leads to loss of efficiency when the prior is misspecified. Moreover, existing methods, such as Xie et al. (2012) and Weinstein et al. (2018), often assume that the nuisance parameters are known and use a consistent estimator for implementation. However, when a large number of units are investigated simultaneously, traditional sample variance estimators may similarly suffer from selection bias, which often leads to severe deterioration in the MSE for estimating the means.

1.2 The proposed approach and main contributions

In the homogeneous setting, Tweedie's formula estimates η_i using the score function of Y_i, but that approach does not immediately extend to heterogeneous data. Instead, this article proposes a two–step approach, "Nonparametric Empirical Bayes Structural Tweedie" (NEST), which first estimates the bivariate score function for data from a two–parameter

exponential family, and second predicts η_i using a generalized version of Tweedie's formula that effectively incorporates the structural information encoded in the variances. A significant challenge in the heterogeneous setting is how to pool information from different study units effectively while accounting for the heterogeneity captured by the possibly unknown nuisance parameters. NEST addresses the issue by proposing a double shrinkage method that simultaneously incorporates the structural information encoded in both the primary (e.g. Y_i) and auxiliary (e.g. S_i) data.

NEST has several clear advantages. First, it simultaneously handles the two main selection biases introduced by heterogeneity: one, selection bias in the primary parameter (mean), which cannot be effectively corrected without also correcting for, two, selection bias in the nuisance parameter (variance). By producing more accurate estimates for the nuisance parameters, NEST in general renders improved shrinkage factors for estimating the primary parameters. Second, NEST makes no explicit assumptions about the prior since it uses a nonparametric method to directly estimate the bivariate score function. Third, NEST exploits the structure of the entire sample and avoids the information loss that occurs in the discretization step used in grouping methods (Weinstein et al. 2018). Fourth, NEST only requires a few simple assumptions to achieve strong asymptotic properties. Finally, NEST provides a general estimation framework for members of the two-parameter exponential family. It is fast to implement, produces stable estimates, and is robust against model mis-specification. We demonstrate numerically that NEST can provide high levels of estimation accuracy relative to a host of benchmark methods.

1.3 Organization

The rest of the paper is structured as follows. In Section 2 we first introduce our hierarchical Gaussian model where both the mean and variance parameters are unknown, and then discuss a version of Tweedie's formula that uses sample variances in Section 2.1. Section 2.2 presents the generalized Tweedie's formula for our hierarchical model which is subsequently used to introduce the oracle NEST estimator in Section 2.3. We then develop a convex optimization

6

approach in Section 3 for estimating the unknown shrinkage factors in the oracle NEST formula. In Section 4, we describe the theoretical setup, justify the optimization criterion, and finally establish asymptotic theories for the proposed NEST estimator. Simulation studies are carried out in Section 5 to compare NEST to competing methods. Two data applications are presented in Section 6. The proofs, as well as additional theoretical and numerical results, are provided in the supplementary material.

2 Double Shrinkage Estimation on Heteroscedastic Normal Data

In the main text of this article, we focus on the normal means problem; extensions of the methodology to other members of the two-parameter exponential family are discussed in Section B of the supplementary material.

Suppose we collect m_i observations for the ith study unit, $i = 1, \ldots, n$. The data are normally distributed obeying the following hierarchical model:

$$
\begin{aligned}
Y_{ij} \mid \mu_i, \tau_i \quad &\overset{i.i.d}{\sim} \quad N(\mu_i, 1/\tau_i), \quad j = 1, \ldots, m_i, \\
\mu_i \mid \tau_i \quad &\overset{ind}{\sim} \quad G_\mu(\cdot | \tau_i), \quad \tau_i \overset{i.i.d}{\sim} H_\tau(\cdot), \quad i = 1, \ldots, n.
\end{aligned}
\tag{2.1}
$$

We view μ_i and τ_i, both of which are unknown, as the primary and nuisance parameters, respectively. The prior distributions $G_\mu(\cdot | \tau_i)$ and $H_\tau(\cdot)$ are unspecified. When the precisions τ_i are known in Model (2.1), compound estimation of the means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ under the squared error loss has received significant attention in recent years (see for example Weinstein et al. (2018), Xie et al. (2012), Cai et al. (2021), Soloff et al. (2021) and the references therein). Here, we develop a nonparametric empirical Bayes method for estimating $\boldsymbol{\mu}$ under the squared error loss when τ_i are unknown.

2.1 Tweedie's formula with estimated sample variances

Let $Y_i = m_i^{-1} \sum_{j=1}^{m_i} Y_{ij}$ and $S_i^2 = (m_i - 1)^{-1} \sum_{j=1}^{m_i} (Y_{ij} - Y_i)^2$ respectively denote the sample mean and sample variance under Model (2.1). Further, let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{S} = (S_1^2, \ldots, S_n^2)$ be the vectors of summary statistics, and $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{s} = (s_1^2, \ldots, s_n^2)$ the observed values. Denote $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ an estimator for $\boldsymbol{\mu}$ based on $(\boldsymbol{Y}, \boldsymbol{S})$. Consider the squared error loss $l_n(\boldsymbol{\mu}, \boldsymbol{\delta}) = n^{-1} \sum_{i=1}^{n} \ell(\mu_i, \delta_i)$, where $\ell(\mu_i, \delta_i) = (\delta_i - \mu_i)^2$. The compound Bayes risk is

$$r(\boldsymbol{\delta}, \mathcal{G}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{\ell(\mu_i, \delta_i)\} = \frac{1}{n} \sum_{i=1}^{n} \int \int \int \ell(\mu_i, \delta_i) f(y, s^2 \mid \boldsymbol{\psi}_i) dy ds^2 d\mathcal{G}(\boldsymbol{\psi}_i), \qquad (2.2)$$

where $\boldsymbol{\psi}_i = (\mu_i, \tau_i)$, $\mathcal{G}(\boldsymbol{\psi}_i) = G_\mu(\mu_i|\tau_i) H_\tau(\tau_i)$, and $f(y, s^2 \mid \boldsymbol{\psi}_i)$ is the likelihood function of (Y_i, S_i^2). The Bayes estimator that minimizes (2.2) is given by $\boldsymbol{\delta}^\pi = (\delta_1^\pi, \ldots, \delta_n^\pi)$, where

$$\delta_i^\pi \coloneqq \delta^\pi(y_i, s_i^2, m_i) = \mathbb{E}(\mu_i \mid Y_i = y_i, S_i^2 = s_i^2, m_i). \qquad (2.3)$$

When the precisions are known, the estimator minimizing the expected squared error loss is,

$$\mathbb{E}(\mu_i \mid Y_i = y_i, \tau_i, m_i) = y_i + \frac{1}{m_i \tau_i} w_1(y_i; m_i, \tau_i), \qquad (2.4)$$

where $w_1(y; m, \tau) \coloneqq \frac{\partial}{\partial y} \log f_{m,\tau}(y)$ and $f_{m,\tau}(\cdot)$ is the pdf of the marginal distribution of Y. Equation (2.4) is the celebrated Tweedie's formula with known variances (Efron 2011) which forms the basis for f-modeling strategies for estimating μ_i, and only requires estimation of the score functions $w_1(y_i; m_i, \tau_i)$ in order to compute the estimator. This is particularly appealing in large-scale studies where one observes thousands of (Y_i, τ_i), making it possible to obtain an accurate estimate of $w_1(y_i; m_i, \tau_i)$ (see Section B of the supplementary material for more details). However, when the precisions are unknown, Model (2.1) in its full generality does not allow a closed-form expression for δ_i^π given in (2.3). In this setting a Tweedie-type formula, such as the one given by (2.4), is not readily available. In such a scenario existing methods, such as Xie et al. (2012), Weinstein et al. (2018), rely on the sample variance S_i^2, a

consistent estimator of the unknown population variance $1/\tau_i$, for practical implementation. For instance, Definition 1 presents the oracle Tweedie's formula with sample variance which is a natural counterpart to Equation (2.4) when the variance is unknown.

Definition 1 *(Oracle Tweedie's formula with sample variances) Consider the hierarchical Model (2.1) and let $f_m(y, s^2) = \int f_m(y, s^2 \mid \psi) d\mathcal{G}(\psi)$ denote the joint density of (Y, S^2). Then an estimator for μ_i is δ_i^{TF} where,*

$$\delta_i^{\mathsf{TF}} := \delta^{\mathsf{TF}}(y_i, s_i^2, m_i) = y_i + \frac{s_i^2}{m_i} w_1(y_i, s_i^2; m_i),$$

and $w_1(y, s^2; m) := \dfrac{\partial}{\partial y} \log f_m(y, s^2)$.

In Definition (1), δ_i^{TF} involves the score function $w_{1i} := w_1(y_i, s_i^2; m_i)$ which is unknown in practice. We discuss its estimation in Section 3. While δ_i^{TF} is an approximation to δ_i^π, Proposition 1 establishes that the oracle Tweedie's estimator with sample variances dominates the sample mean estimator under a squared error loss function.

Proposition 1 *Suppose $f_m(y, s^2)$ is a log-concave density and $w_1(y, s^2; m)$ is a non-decreasing function of s^2. Then under Model (2.1) and for $m_i > 3$, we have,*

$$r(\boldsymbol{\delta}^\pi, \mathcal{G}) < r(\boldsymbol{\delta}^{\mathsf{TF}}, \mathcal{G}) < r(\boldsymbol{Y}, \mathcal{G}).$$

However, when a large number of units are investigated simultaneously, traditional sample variance estimators may suffer from selection bias (Jing et al. 2016, Kwon and Zhao 2018), and their direct use may lead to severe deterioration in the MSE for estimating the means. By contrast, our approach relies on carefully constructed shrinkage factors for both the sample mean and sample variance, which ultimately provide a much improved approximation to δ_i^π. The key idea is to exploit the exponential family representation of the posterior distribution of (μ_i, τ_i) to first derive a generalized Tweedie's formula for the canonical parameters $(\tau_i \mu_i, \tau_i)$ (Section 2.2) and then employ the canonical formulas to construct our oracle NEST estimator

9

(Section 2.3).

2.2 Generalized Tweedie's formula for a heteroscedastic Normal model

In this section we present the generalized Tweedie's formula for the canonical parameters $(\tau_i \mu_i, \tau_i)$ under Model (2.1). To this end we exploit the properties of the two-parameter exponential family to form a new representation of the posterior distribution of (μ_i, τ_i) in Proposition 2. This useful representation is subsequently employed in Corollary 1 to construct a generalized Tweedie's formula for the canonical parameters. Finally, Definition 2 (Section 2.3) introduces the oracle version of our proposed NEST estimator of μ_i.

Proposition 2 *Consider hierarchical Model* (2.1) *with* $m_i > 3$. *Let* $f_m(y, s^2) = \int f_m(y, s^2 \mid \boldsymbol{\psi}) d\mathcal{G}(\boldsymbol{\psi})$ *denote the joint density of* (Y, S^2). *The posterior distribution of* (μ_i, τ_i) *belongs to a two-parameter exponential family with density*

$$f_{m_i}(\mu_i, \tau_i | y_i, s_i^2) \propto \exp\left\{\boldsymbol{\eta}_i^T \boldsymbol{T}(\mu_i, \tau_i) - A(\boldsymbol{\eta}_i)\right\} g_\mu(\mu_i|\tau_i) h_\tau(\tau_i),$$

where $\boldsymbol{T}(\mu_i, \tau_i) = (\tau_i \mu_i, \tau_i/2), \boldsymbol{\eta}_i = \left\{m_i y_i, -m_i y_i^2 - (m_i - 1)s_i^2\right\} := (\eta_{1i}, \eta_{2i}), \gamma(\eta_{1i}, \eta_{2i}) = \frac{-\eta_{2i} - m_i^{-1}\eta_{1i}^2}{m_i - 1}$ *and* $A(\boldsymbol{\eta}_i) = -\frac{1}{2}(m_i - 3)\log\gamma(\eta_{1i}, \eta_{2i}) + \log f_{m_i}\left\{m_i^{-1}\eta_{1i}, \gamma(\eta_{1i}, \eta_{2i})\right\}$.

The two-parameter exponential family representation of the joint posterior distribution of (μ_i, τ_i) in Proposition 2 is particularly useful because, as shown in Corollary 1, it allows one to explicity compute the Bayes estimators of the canonical parameters $\zeta_i := \tau_i \mu_i$ and τ_i.

Corollary 1 (Generalized Tweedie's formula for the canonical parameters) *Under the hierarchical Model* (2.1) *with* $m_i > 3$, *the Bayes estimators of* (ζ_i, τ_i) *under squared error loss are, respectively,*

$$\hat{\zeta}_i^\pi := \hat{\zeta}^\pi(y_i, s_i^2, m_i) = \mathbb{E}(\zeta_i|y_i, s_i^2, m_i) = y_i \mathbb{E}(\tau_i|y_i, s_i^2, m_i) + m_i^{-1} w_1(y_i, s_i^2; m_i), \quad (2.5)$$

$$\hat{\tau}_i^\pi := \hat{\tau}^\pi(y_i, s_i^2, m_i) = \mathbb{E}(\tau_i|y_i, s_i^2, m_i) = \frac{m_i - 3}{(m_i - 1)s_i^2} - \frac{2}{m_i - 1}w_2(y_i, s_i^2; m_i), \quad (2.6)$$

where $w_1(y, s^2; m) := \frac{\partial}{\partial y} \log f_m(y, s^2), \ w_2(y, s^2; m) := \frac{\partial}{\partial s^2} \log f_m(y, s^2)$.

Corollary 1 is significant for several reasons. First, it generalizes Tweedie's formula, which can be recovered from Equation (2.5) by treating τ_i as a known constant and dividing both sides by τ_i. Second, $\hat{\tau}_i^\pi$ in Equation (2.6) has an interesting interpretation. Apart from being the Bayes estimate of τ_i under the squared error loss, $1/\hat{\tau}_i^\pi$ is the Bayes estimate of $1/\tau_i$ under Stein's loss (James and Stein 1961). Third, Corollary 1 can be employed to construct estimators for functions of (ζ_i, τ_i) such as the Sharpe ratio in finance applications (Section 6.2). Finally, Corollary 1 provides the key result for defining our oracle NEST estimator in the next section.

2.3   The oracle NEST estimator

In this section we first present the oracle NEST estimator, and then discuss its theoretical properties.

Definition 2 *Consider hierarchical Model* (2.1) *with* $m_i > 3$. *Then the oracle NEST estimator is*

$$\delta_i^* := \delta^*(y_i, s_i^2, m_i) = \frac{\hat{\zeta}^\pi(y_i, s_i^2, m_i)}{\hat{\tau}^\pi(y_i, s_i^2, m_i)} \quad = \quad y_i + \frac{s_i^2}{m_i}\gamma(y_i, s_i^2, m_i)w_1(y_i, s_i^2, m_i), \qquad (2.7)$$

$$where \ \gamma(y_i, s_i^2, m_i) = \frac{m_i - 1}{m_i - 3 - 2s_i^2 w_2(y_i, s_i^2; m_i)}.$$

The oracle NEST estimator in Equation (2.7) is a ratio of $\hat{\zeta}_i^\pi$ and $\hat{\tau}_i^\pi$, and it involves the shrinkage factor $\gamma_i := \gamma(y_i, s_i^2, m_i)$ that should be applied to s_i^2 when τ_i is unknown. This shrinkage factor is, by construction, positive and depends on $w_{2i} := w_2(y_i, s_i^2; m_i)$ that controls the magnitude of shrinkage that is applied to the sample variance s_i^2. In practical applications, however, the score functions w_{1i} and w_{2i} are unknown. In Section 3, we develop a data-driven NEST estimator with estimated scores.

Equation (2.7) has a striking similarity to Tweedie's formula of Definition 1 in the sense

11

that δ_i^* involves an unbiased estimate y_i of μ_i plus a shrinkage factor. The key difference from Tweedie's formula, however, is that while the shrinkage factor in δ_i^* relies on γ_i, Tweedie's formula uses sample variances s_i^2. Proposition 3 below shows that this difference is important because the estimation risk, under a squared error loss function, of $\boldsymbol{\delta}^* = (\delta_i^*, \ldots, \delta_n^*)$ is uniformly smaller than that of $\boldsymbol{\delta}^{\mathsf{TF}} = (\delta_i^{\mathsf{TF}}, \ldots, \delta_n^{\mathsf{TF}})$.

We impose the following regularity conditions for comparing the estimation risks of $\boldsymbol{\delta}^*$ and δ^{TF} in Proposition 3.

Assumption 1 *The shrinkage factor $\gamma(y, s^2, m)$ satisfies (a)* $\mathbb{E}\left[S^2 \gamma(Y, S^2, m)\big|\mu, \tau\right] \leq 1/\tau$, *(b) $s^2 \gamma(y, s^2, m)$ is non-decreasing in s^2, and (c) $\gamma(y, s^2, m)$ is non-increasing in s^2.*

Assumption 2 *Let $\omega(y, s^2; m) := w_1(y, s^2; m)\dfrac{\partial}{\partial y} w_2(y, s^2; m)$. Then (a) $\omega(y, s^2; m) \leq 0$ and, (b) $\omega(y, s^2; m)$ is non-decreasing in s^2.*

Assumptions 1 and 2 are regularity conditions on the behavior of the shrinkage factor γ and the score functions w_1, w_2. For instance, Assumption 1(a), together with $\mathbb{E}[\gamma(Y, S^2, m)] \leq 1$, guarantees that $Cov[S^2, \gamma(Y, S^2, m)] \leq 0$. Assumptions 1(b) and 1(c) enforce monotonicity, respectively, on the Bayes estimator of the precision (Equation (2.6)) and the shrinkage factor $\gamma(y, s^2, m)$. These conditions are satisfied, for example, when (μ, τ) have a conjugate prior under Model 2.1. Similarly, Assumption 2 holds under conjugate priors and is also true when the prior on τ is discrete with just one mass point. Proposition 1 can be extended as follows.

Proposition 3 *Let $f_m(y, s^2)$ be a log-concave density. Suppose Assumptions 1 – 2 hold. Then under Model (2.1) and for $m_i > 7$, we have,*

$$r(\boldsymbol{\delta}^\pi, \mathcal{G}) < r(\boldsymbol{\delta}^*, \mathcal{G}) < r(\boldsymbol{\delta}^{\mathsf{TF}}, \mathcal{G}) < r(\boldsymbol{Y}, \mathcal{G}).$$

The oracle NEST estimator δ_i^* is, in general, different from the Bayes estimator δ_i^π: δ_i^π has full knowledge of the prior distributions $G_\mu(\cdot|\tau)$ and $H_\tau(\cdot)$, whereas δ_i^* has only the information on the true score functions (w_{1i}, w_{2i}). The departure reflects the intrinsic difficulty in estimating

μ_i (other than the canonical parameter $\boldsymbol{T}$) using an f-modeling approach when the variances are unknown. However, it is worth doing f-modeling for several reasons. First, under the special setting of a conjugate prior on (μ, τ), Corollary 1 in Section 4.2 shows that δ_i^* coincides with δ_i^π. Second, in situations where δ_i^* and δ_i^π do not coincide, our empirical results suggest that the efficiency gain of the data-driven NEST estimator over competing linear shrinkage methods (Xie et al. 2012, Weinstein et al. 2018) and Tweedie's formula (Definition 1), both of which use plug-in sample variances, is substantial across many settings.

2.4 Connection to existing works

There are two main modeling strategies for the empirical Bayes (EB) estimation of normal means, respectively known as the g-modeling and f-modeling strategies in the terminology of Efron (2014). The idea of g-modeling is to first obtain a deconvoluting estimate of G_μ in Equation 2.1, denoted $\hat{G}_\mu$, and then predict μ_i by plugging $\hat{G}_\mu$ into Bayes rule. $\hat{G}_\mu$ can be constructed via the nonparametric maximum likelihood estimate (NPMLE; Kiefer and Wolfowitz 1956, Laird 1978), or be modeled by distributions in a low-dimensional exponential family (Efron 2014). Some notable works along this line include Jiang and Zhang (2009), Koenker and Mizera (2014), Gu and Koenker (2017a), Saha and Guntuboyina (2020), and Soloff et al. (2021). By contrast, the f-modeling strategy directly predicts μ_i based on Tweedie's formula (or its generalized version), which only depends on the marginal density f, sidestepping the need of deconvoluting estimation. Notable works along this line include Brown and Greenshtein (2009) and Efron (2011), both of which use sample variances. NEST adopts the f-modeling strategy and has two advantages over existing f-modeling methods. First, we provide in Proposition 3 precise conditions under which the oracle NEST estimator dominates Tweedie's estimator with sample variances (cf. Definition 1). Second, in contrast with existing f-modeling methods, we develop a convex optimization approach to construct a data-driven rule that is fast, stable, and capable of incorporating various structural constraints. Our numerical results show that the data-driven NEST offers substantial improvement in the estimation risk over other f-modeling methods.

The g-modeling approach via the NPMLE provides an excellent tool for EB estimation of heteroscedastic means. NPMLE is competitive to NEST in most of our numerical studies. However, to the best of our knowledge, the asymptotic properties of the NPMLE are highly nontrivial to establish and often require strong assumptions. For instance, the analysis in Saha and Guntuboyina (2020) only works for a limited class of covariance structures, and the theory on the rate of convergence is applicable only when the degree of heteroscedasticity is "mild"; alternatively the analysis in Soloff et al. (2021) assumes that μ_i are independent of σ_i, which is often violated in practice (Weinstein et al. 2018). In contrast, we establish the asymptotic properties of NEST without assumptions on the degree of heteroscedasticity or independence between μ_i and σ_i.

A key advantage of g-modeling is its capability to deal with a wider range of problems, particularly those in which direct use of the marginal density f itself cannot yield a solution. Meanwhile, the f-modeling approach, which often has a simple and intuitive form (e.g. Tweedie's formula), is attractive when only the information of the marginal distribution is needed for solving the problem of interest. The merit and simplicity of f-modeling are particularly manifested in our theoretical analysis of Section 4.

3   Estimation of shrinkage factors via convex optimization

In this section we discuss the estimation of the shrinkage factors and introduce the data-driven NEST estimator in Definition 3. We begin by introducing some notation. Let $\boldsymbol{x} = (y, s^2)$ be a generic pair of observations from the distribution with marginal density $f_m(\boldsymbol{x})$, which we assume is continuously differentiable with support on $\mathcal{X} \subseteq \mathbb{R} \times \mathbb{R}^+$. Denote the score function

$$\boldsymbol{w}(\boldsymbol{x}; m) = \nabla_{\boldsymbol{x}} \log f_m(\boldsymbol{x}) := \{w_1(\boldsymbol{x}; m), w_2(\boldsymbol{x}; m)\} . \tag{3.8}$$

Next, for $i = 1, \ldots, n$, let $\boldsymbol{z}^i = (\boldsymbol{x}^i, m_i)$ where $\boldsymbol{x}^i$ is an observation from a distribution with density $f_{m_i}(\boldsymbol{x})$.

3.1 Convex optimization

We first describe the methodology and then provide explanations. Let $\mathcal{K}(\boldsymbol{z}^i, \boldsymbol{z}^j) = \exp\{-(\boldsymbol{z}^i - \boldsymbol{z}^j)^T \Omega(\boldsymbol{z}^i - \boldsymbol{z}^j)/2\}$ be the Radial Basis Function (RBF) kernel, with $\Omega^{3\times 3}$ being the inverse of the sample covariance matrix of $(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^n)$. Denote

$$\mathcal{W}_0^{n\times 2} = \left\{\boldsymbol{w}(\boldsymbol{x}^1; m_1), \ldots, \boldsymbol{w}(\boldsymbol{x}^n; m_n)\right\}^T, \text{ where} \tag{3.9}$$

$$\boldsymbol{w}(\boldsymbol{x}^i; m_i) = \nabla_{\boldsymbol{x}} \log f_{m_i}(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{x}_i} := \left\{w_1(\boldsymbol{x}^i; m_i), w_2(\boldsymbol{x}^i; m_i)\right\}.$$

We denote $w_k(\boldsymbol{x}^i; m_i)$ by w_{ki}, $k = 1, 2$, for the remainder of this article.

Let $\nabla_{z_k^j}\mathcal{K}(\boldsymbol{z}^i, \boldsymbol{z}^j)$ be the partial derivative of $\mathcal{K}(\boldsymbol{z}^i, \boldsymbol{z}^j)$ with respect to the kth component of $\boldsymbol{z}^j$. The following matrices are needed in our proposed estimator:

$$\boldsymbol{K}^{n\times n} = [K_{ij}]_{1\leq i\leq n, 1\leq j\leq n}, \quad \boldsymbol{\nabla K}^{n\times 2} = [\nabla K_{ik}]_{1\leq i\leq n, 1\leq k\leq 2},$$

where $K_{ij} = \mathcal{K}(\boldsymbol{z}_i, \boldsymbol{z}_j)$ and $\nabla K_{ik} = \sum_{j=1}^{n} \nabla_{z_k^j}\mathcal{K}(\boldsymbol{z}^i, \boldsymbol{z}^j)$. Next we formally define our proposed Nonparametric Empirical-Bayes Structural Tweedie (NEST) estimator.

Definition 3 *Consider hierarchical Model (2.1) with $m_i > 3$. For a fixed regularization parameter $\lambda > 0$, let $\hat{\mathcal{W}}_n(\lambda) = (\hat{\boldsymbol{w}}_\lambda^1, \ldots, \hat{\boldsymbol{w}}_\lambda^n)^T$, where $\hat{\boldsymbol{w}}_\lambda^i = (\hat{w}_{1,\lambda}^i, \hat{w}_{2,\lambda}^i)$, be the solution to the following quadratic optimization problem:*

$$\min_{\mathcal{W}\in\mathbb{R}^{n\times 2}} \frac{1}{n^2}\text{trace}\left(\mathcal{W}^T\boldsymbol{K}\mathcal{W} + 2\mathcal{W}^T\boldsymbol{\nabla K}\right) + \rho_n(\mathcal{W}; \lambda), \tag{3.10}$$

where $\rho_n(\mathcal{W}; \lambda)$ is a penalty on the elements of $\mathcal{W}$. Then the NEST estimator for μ_i is

$$\delta_i^{\mathsf{ds}}(\lambda) = y_i + \frac{s_i^2}{m_i}\hat{\gamma}_i(\lambda)\hat{w}_{1,\lambda}^i, \tag{3.11}$$

$$\text{where } \hat{\gamma}_i(\lambda) = \frac{m_i - 1}{m_i - 3 - 2s_i^2\hat{w}_{2,\lambda}^i}, \tag{3.12}$$

with the superscript ds denoting "double shrinkage".

Although not immediately obvious, we show in Section 4 that, under the compound estimation setting, minimizing the first term in the objective function (3.10) is asymptotically equivalent to minimizing the kernelized Stein's discrepancy (KSD; Liu et al. 2016, Chwialkowski et al. 2016). Roughly speaking, the KSD measures how far a given $n \times 2$ matrix $\mathcal{W}$ is from the true score matrix $\mathcal{W}_0$. A key property of the KSD is that it is always non-negative and is equal to 0 if and only if $\mathcal{W}$ and $\mathcal{W}_0$ are equal. Hence, solving the convex program (3.10) is equivalent to finding a $\hat{\mathcal{W}}$ that is as close as possible to $\mathcal{W}_0$. Since the oracle NEST estimator in Definition 2 is constructed based on $\mathcal{W}_0$, we can expect that the data-driven NEST estimator based on $\hat{\mathcal{W}}$ would be asymptotically optimal. Theory underpinning this intuition are established in Section 4.

The second term in Equation (3.10), $\rho_n(\mathcal{W}; \lambda)$, is a penalty function that increases as the elements of $\mathcal{W}$ move further away from 0. In this article, we take $\rho_n(\mathcal{W}; \lambda) = (\lambda/n^2) \sum_{i=1}^n \sum_{k=1}^2 \mathcal{W}_{ik}^2 = (\lambda/n^2) \|\mathcal{W}\|_F^2$, where $\mathcal{W}_{ik}$ are the elements of matrix $\mathcal{W}$ and $\|\mathcal{W}\|_F$ is the Frobenius norm of $\mathcal{W}$. A large λ forces the estimated shrinkage factors towards 0, and in the limit the NEST estimate is simply the unbiased estimate y_i. An alternative approach, as pursued in James et al. (2020), is to penalize the lack of smoothness in $\boldsymbol{w}$ as a function of $\boldsymbol{x}$.

A key characteristic of $\delta_i^{\mathsf{ds}}(\lambda)$ in Equation (3.11) is that it exploits the joint structural information available in both Y_i and S_i^2 through $\hat{\mathcal{W}}_n(\lambda)$. Although the loss function only involves the means, we perform shrinkage on both the mean and variance dimensions. Inspecting Equations (3.11) and (3.12), we expect that the improved accuracy achieved by $\hat{\gamma}_i(\lambda)$ will lead to better shrinkage factors for $\delta_i^{\mathsf{ds}}(\lambda)$ and hence additional reduction in the estimation risk. Our numerical results in Sections 5 and 6 reveal that this is indeed true and the proposed NEST estimator dominates other linear shrinkage estimators and Tweedie's formula across many settings. In Section B of the supplementary material we adopt a similar strategy to extend the estimation framework presented in Definition 3 to other distributions in the two-parameter exponential family where the nuisance parameter is known.

We end this section with a discussion of the simpler case of equal sample sizes, i.e. $m_i = m$

for all i. For instance, the leukemia dataset analyzed in Jing et al. (2016) consists of the expression levels of $n = 5{,}000$ genes for $m = 27$ acute lymphoblastic leukemia patients. The heterogeneity across the n units is due to the intrinsic variability instead of the varied number of replicates. When the m_i's are equal, the RBF kernel $\mathcal{K}(\cdot, \cdot)$ needs to be modified to avoid a singular sample covariance matrix. Denote $\mathcal{K}(\boldsymbol{x}^i, \boldsymbol{x}^j) = \exp\{-0.5(\boldsymbol{x}^i - \boldsymbol{x}^j)^T \Omega (\boldsymbol{x}^i - \boldsymbol{x}^j)\}$ the modified RBF kernel with $\Omega^{2 \times 2}$ being the inverse of the sample covariance matrix of $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)$. Correspondingly in Definition 3, $\hat{\mathcal{W}}_n(\lambda)$ are the estimates of the shrinkage factors $\mathcal{W}_0^{n \times 2} = \left\{ \boldsymbol{w}(\boldsymbol{x}^1; m), \ldots, \boldsymbol{w}(\boldsymbol{x}^n; m) \right\}^T$, where $\boldsymbol{w}(\boldsymbol{x}^i; m) = \left\{ w_1(\boldsymbol{x}^i; m), w_2(\boldsymbol{x}^i; m) \right\} := (w_{1i}, w_{2i})$.

3.2 Details around implementation

In this section we discuss details around the implementation of NEST. First note that Equation (3.10) can be solved separately for the two columns of $\mathcal{W}$, which respectively yield the estimates for w_{1i} and w_{2i}. Next, the solution to Equation (3.10), with our penalty $\rho_n(\mathcal{W}; \lambda) = (\lambda/n^2) \|\mathcal{W}\|_F^2$, is available in the closed form of $\hat{\mathcal{W}}_n(\lambda) = -\mathcal{B}(\lambda) \boldsymbol{\nabla} \boldsymbol{K}$, where $\mathcal{B}(\lambda) = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}$. However, in our implementation the closed form solution is replaced by a convex program that directly solves (3.10) with constraint $\mathcal{W} \boldsymbol{a} \preceq \boldsymbol{b}$, where $\boldsymbol{a} = (0, 1)^T$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ with $b_i = \frac{1}{2}(m_i - 3)/s_i^2 - \kappa$ for some $\kappa > 0$. Inspecting Equation (3.12) shows that adding this constraint guarantees that $\hat{\gamma}_i(\lambda) < \infty$. This is desirable in both numerical and theoretical analyses. Similar ideas have been used in the seminal work of Koenker and Mizera (2014).

The practical implementation requires a data-driven scheme for choosing λ. We propose to use a variation of the modified cross validation scheme of Brown et al. (2013), which involves splitting Y_{ij} into two parts: $U_{ij} = Y_{ij} - (1/\alpha)\epsilon_{ij}$ and $V_{ij} = Y_{ij} + \alpha \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \tau_i^{-1})$, and U_{ij} and V_{ij} are used to construct the estimator and to choose the tuning parameter, respectively. However in our setup τ_i is unknown; hence we sample ϵ_{ij}'s from $N(0, \bar{S}^2)$ where $\bar{S}^2 = n^{-1} \sum_{i=1}^n S_i^2$. Let $\bar{V}_i = m_i^{-1} \sum_{j=1}^{m_i} V_{ij}$, $\bar{U}_i = m_i^{-1} \sum_{j=1}^{m_i} U_{ij}$, $\mathcal{U} = \{U_{ij} : 1 \le i \le n, 1 \le j \le m_i\}$ and $\mathcal{V} = \{V_{ij} : 1 \le i \le n, 1 \le j \le m_i\}$. Then

conditional on (μ_i, τ_i), $\bar{U}_i$ and $\bar{V}_i$ are approximately independent with mean $\mathbb{E}(Y_i)$ when m is large. Define $\vartheta_n(\lambda; \mathcal{U}, \mathcal{V}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \bar{V}_i - \delta_i^{\mathsf{ds}}(\bar{U}_i; \mathcal{U}, \lambda) \right\}^2$. The tuning parameter will be chosen as $\hat{\lambda} := \arg\min_{\lambda \in \Lambda} \vartheta_n(\lambda)$. In our numerical studies of Sections 5 and 6, we set $\alpha = 1/2$, $\Lambda = [10^{-2}, 5^2]$. The tuning parameter $\hat{\lambda}$ is obtained from this scheme and then used to estimate μ_i via Equation (3.11).

Remark 1 *$\vartheta_n(\lambda; \mathcal{U}, \mathcal{V})$ provides a practical criterion for choosing λ and works well empirically in our numerical studies. However, ϑ_n is not an unbiased estimate of the true risk. We note in Section 5 that the relative risk of the data-driven NEST is slightly off from 1 when $m = 10$ (e.g. the left panels in Figures 1 & 2). As m increases the covariance between U_{ij} and V_{ij} converges to 0 and the risk of our data-driven estimator is almost identical to that of an oracle (which selects the optimal λ based on the true risk). The development of a SURE criterion for this setting is a challenging topic requiring further research. See also Ignatiadis and Wager (2019) for related discussions.*

We are developing an `R` package, `nest`, to implement the NEST estimator in Definition 3. The `R` code that reproduces the numerical results in Sections 5 and 6 can be downloaded from the link: `https://www.dropbox.com/sh/vh3b48zuq4axo0b/AAD4zTsTqPGLRqzK7CiNW5Iya?dl=0`.

4 Theory

In this section we introduce the Kernelized Stein's Discrepancy (KSD) measure (Liu et al. 2016, Chwialkowski et al. 2016) and discuss its connection to the quadratic program (3.10). While the KSD has been used in various contexts including goodness of fit tests (Liu et al. 2016, Yang et al. 2018), variational inference (Liu and Wang 2016) and Monte Carlo integration (Oates et al. 2017), its connections to compound estimation and empirical Bayes methodology was established only recently (Banerjee et al. 2020). The analysis in this and following sections is geared towards the case $m_i = m$ for $i = 1, \ldots, n$. Under this setting, $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)$ constitute an i.i.d random sample from $f_m(\boldsymbol{x})$. The case of unequal m_i's can be

analyzed in a similar fashion by first assuming that m_i's are a random sample from a distribution with mass function $q(\cdot)$ and are independent of (μ_i, τ_i). Then $\boldsymbol{z} = (\boldsymbol{x}, m)$ has distribution with density $p(\boldsymbol{z}) := q(m) f_m(\boldsymbol{x})$, where $\boldsymbol{z}^i = (\boldsymbol{x}^i, m_i)$, $(\boldsymbol{z}^1, \dots, \boldsymbol{z}^n)$ are realizations of an i.i.d random sample from $p(\boldsymbol{z})$.

4.1 Kernelized Stein's Discrepancy

Suppose $\boldsymbol{X}$ and $\boldsymbol{X}'$ are i.i.d. copies from the marginal distribution of (Y, S^2) that has density f wherein the dependence on m is implicit. Denote $\boldsymbol{w}(\boldsymbol{X})$ and $\boldsymbol{w}(\boldsymbol{X}')$, defined in Equation (3.8), to be the score functions at $\boldsymbol{X}$ and $\boldsymbol{X}'$ respectively. Suppose $\tilde{f}$ is an arbitrary density function on the support of (Y, S^2), for which we similarly define $\tilde{\boldsymbol{w}}(\boldsymbol{X})$. The KSD, formally defined as

$$\mathcal{S}(f, \tilde{f}) = \mathbb{E}_{\boldsymbol{X}, \boldsymbol{X}' \sim f}\left[\left\{\tilde{\boldsymbol{w}}(\boldsymbol{X}) - \boldsymbol{w}(\boldsymbol{X})\right\}^T \mathcal{K}(\boldsymbol{X}, \boldsymbol{X}')\left\{\tilde{\boldsymbol{w}}(\boldsymbol{X}') - \boldsymbol{w}(\boldsymbol{X}')\right\}\right], \qquad (4.13)$$

provides a discrepancy measure between f and $\tilde{f}$ in the sense that $\mathcal{S}(f, \tilde{f})$ tends to increase when there is a bigger disparity between $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ (or equivalently, between f and $\tilde{f}$), and

$$\mathcal{S}(f, \tilde{f}) \geq 0 \text{ and } \mathcal{S}(f, \tilde{f}) = 0 \text{ if and only if } f = \tilde{f}.$$

The direct evaluation of $\mathcal{S}(f, \tilde{f})$ is difficult because $\boldsymbol{w}$ is unknown. Liu et al. (2016) introduced an alternative representation of the KSD that does not directly involve $\boldsymbol{w}$:

$$
\begin{aligned}
\mathcal{S}(f, \tilde{f}) &= \mathbb{E}_f \kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}), \tilde{\boldsymbol{w}}(\boldsymbol{X}')](\boldsymbol{X}, \boldsymbol{X}') \\
&= \mathbb{E}_f\left\{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i), \tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i, \boldsymbol{X}^j)\mathbb{I}(i \neq j)\right\} \\
&= \mathbb{E}_f\left[\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})\right] := \mathbb{M}(\tilde{\mathcal{W}}), \qquad (4.14)
\end{aligned}
$$

where $\{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n\}$ is a random sample from f, $\mathbb{E}_f$ denotes expectation under f and

$$
\begin{aligned}
\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{x}), \tilde{\boldsymbol{w}}(\boldsymbol{x}')](\boldsymbol{x}, \boldsymbol{x}') &= \tilde{\boldsymbol{w}}(\boldsymbol{x})^T \tilde{\boldsymbol{w}}(\boldsymbol{x}')\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') + \tilde{\boldsymbol{w}}(\boldsymbol{x})^T \nabla_{\boldsymbol{x}'}\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') + \nabla_{\boldsymbol{x}}\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}')^T \tilde{\boldsymbol{w}}(\boldsymbol{x}) \\
&+ \operatorname{trace}(\nabla_{\boldsymbol{x}, \boldsymbol{x}'}\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}')) \quad\quad\quad\quad\quad\quad\quad\quad\quad (4.15)
\end{aligned}
$$

is a smooth and symmetric positive definite kernel function associated with the U-statistic $\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})$. The implementation of the NEST estimator in Definition 3 boils down to the estimation of $\mathcal{W}_0$ via the convex program (3.10), which corresponds to minimizing

$$
\hat{\mathbb{M}}_{\lambda(n),n}(\tilde{\mathcal{W}}) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i), \tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i, \boldsymbol{X}^j) + \frac{\lambda(n)}{n^2}\|\tilde{\mathcal{W}}\|_F^2 \quad\quad (4.16)
$$

w.r.t. $\tilde{\mathcal{W}}$. A key observation is that if the empirical criterion $\hat{\mathbb{M}}_{\lambda(n),n}(\tilde{\mathcal{W}})$ is asymptotically equal to the population KSD criterion $\mathbb{M}(\tilde{\mathcal{W}})$, then minimizing $\hat{\mathbb{M}}_{\lambda(n),n}(\tilde{\mathcal{W}})$ with respect to $\tilde{\mathcal{W}}$ is effectively the process of finding an $\tilde{\mathcal{W}}$ that is as close as possible to $\mathcal{W}_0$ in Equation (3.9). This intuitively justifies the NEST estimator in Definition 3. In what follows, we denote $\lambda(n)$ by λ and keep its dependence on n implicit. Next we show that the NEST estimator in Definition 3 is asymptotically close to its oracle counterpart.

4.2 Asymptotic Properties of NEST

This section studies the asymptotic properties of the NEST estimator. We begin by recalling the oracle NEST estimator $\boldsymbol{\delta}^* = (\delta_1^*, \ldots, \delta_n^*)$ for $\boldsymbol{\mu}$, where

$$
\delta_i^* := \delta^*(y_i, s_i^2, m) = \frac{\hat{\zeta}^\pi(y_i, s_i^2, m)}{\hat{\tau}^\pi(y_i, s_i^2, m)} = y_i + \frac{s_i^2}{m}\gamma(y_i, s_i^2; m_i)w_1(y_i, s_i^2, m), \quad\quad (4.17)
$$

and $\hat{\zeta}^\pi$ and $\hat{\tau}^\pi$ are respectively the Bayes estimators of $\tau\mu$ and τ as defined in Equation (2.5). Viewing the proposed NEST estimator $\delta_i^{\mathsf{ds}}(\lambda)$ as a data-driven approximation to δ_i^*, we study the quality of this approximation for large n and fixed m.

We impose the following regularity conditions where f denotes the density function of

the joint marginal distribution of (Y, S^2) and $\mathbb{E}_f$ denotes expectation under f.

Assumption 3 $\mathbb{E}_f|\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i), \tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i, \boldsymbol{X}^j)|^2 < \infty$ *for any* $(i, j) \in \{1, \ldots, n\}$.

Assumption 4 $\int_{\mathcal{X}} g(\boldsymbol{x})^T \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') d\boldsymbol{x} d\boldsymbol{x}' > 0$ *for any* $g : \mathcal{X} \to \mathbb{R}^2$ *s.t.* $0 < \|g\|_2^2 < \infty$.

Assumption 5 *For some* $\epsilon_i \in (0, 1), i = 1, 2, 3$, $\mathbb{E}_G\{\exp(\epsilon_1|\mu|)\} < \infty$, $\mathbb{E}_H\{\exp(\epsilon_2/\tau)\} < \infty$ *and* $\mathbb{E}_H\{\exp(\epsilon_3\tau)\} < \infty$.

Assumption 3 is a standard moment condition on $\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i), \tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i, \boldsymbol{X}^j)$ [see, for example, Section 5.5 in Serfling (2009)] , which is needed for establishing the Central Limit Theorem for the U-statistic $\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})$ in Equation (4.14). Assumption 4 is a condition from Liu et al. (2016), Chwialkowski et al. (2016) for ensuring that $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}')$ is integrally strictly positive definite. This guarantees that the KSD $\mathcal{S}(f, \tilde{f})$ is a valid discrepancy measure in the sense that $\mathcal{S}(f, \tilde{f}) \geq 0$ and $\mathcal{S}(f, \tilde{f}) = 0$ if and only if $f = \tilde{f}$. Assumption 5 represents moment conditions on the prior distributions. Together, they ensure that with high probability $|\mu| \leq \log n$ and $1/\log n \leq \tau \leq \log n$ as $n \to \infty$. This is formalized in Lemma 3 in Section A.6 of the supplementary material. These conditions allow a cleaner statement of our theoretical results. It is likely that Assumption 5, which is mild, can be further relaxed but we do not seek the full generality here.

Theorem 1 below establishes the asymptotic consistency of the sample criterion $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\mathcal{W}})$ around the population criterion $\mathbb{M}(\tilde{\mathcal{W}})$.

Theorem 1 *If* $\lambda(n)/\sqrt{n} \to 0$ *as* $n \to \infty$ *then, under Assumption 3, we have*

$$\left|\hat{\mathbb{M}}_{\lambda,n}(\tilde{\mathcal{W}}) - \mathbb{M}(\tilde{\mathcal{W}})\right| = O_p\big(n^{-1/2}\big).$$

Moreover, along with the fact that $\mathbb{M}(\mathcal{W}_0) = 0$, Theorem 1 justifies $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\mathcal{W}})$ as an appropriate optimization criterion.

Theorem 2 establishes the consistency of the estimated score functions.

21

Theorem 2 *If $\lim_{n \to \infty} c_n n^{-1/2} = 0$ then under the conditions of Theorem 1 and Assumption 4, we have*

$$\lim_{n \to \infty} \mathbb{P}\left\{ \frac{1}{n} \left\| \hat{\mathcal{W}}_n(\lambda) - \mathcal{W}_0 \right\|_F^2 \geq c_n^{-1} \epsilon \right\} = 0, \ \textit{for any } \epsilon > 0.$$

Theorem 3 establishes the optimality theory of $\boldsymbol{\delta}^{\mathsf{ds}}$ by showing that (a) the average squared error between $\boldsymbol{\delta}^{\mathsf{ds}}(\lambda)$ and $\boldsymbol{\delta}^*$ is asymptotically small, and (b) the estimation loss of NEST converges in probability to that of its oracle counterpart as $n \to \infty$.

Theorem 3 *Suppose $\lambda(n)/\sqrt{n} \to 0$ as $n \to \infty$ and Assumptions 3 – 5 hold. If $\lim_{n \to \infty} c_n n^{-1/2} \log^6 n = 0$, then $\frac{c_n}{n} \left\| \boldsymbol{\delta}^{\mathsf{ds}}(\lambda) - \boldsymbol{\delta}^* \right\|_2^2 = o_p(1)$. Furthermore, if additionally $\lim_{n \to \infty} c_n n^{-1/4} \log^6 n = 0$, then $c_n \left| l_n(\boldsymbol{\delta}^{\mathsf{ds}}(\lambda), \boldsymbol{\mu}) - l_n(\boldsymbol{\delta}^*, \boldsymbol{\mu}) \right| = o_p(1)$.*

As mentioned earlier, δ_i^* in Equation (4.17) is not, in general, the Bayes estimator δ_i^π of μ_i. This follows since, dropping subscript i,

$$\delta^*(y, s^2) = \frac{\mathbb{E}(\tau\mu|y, s^2)}{\mathbb{E}(\tau|y, s^2)} = \frac{\mathbb{E}[\tau\mathbb{E}(\mu|y, s^2, \tau)|y, s^2]}{\mathbb{E}(\tau|y, s^2)} \neq \mathbb{E}(\mu|y, s^2) = \delta^\pi(y, s^2)$$

unless μ and τ are conditionally independent given y and s^2, in which case $\delta^*(y, s^2) = \delta^\pi(y, s^2)$. A natural setting where $\mathbb{E}(\mu|y, s^2, \tau)$ is indeed independent of τ is the popular scenario of conjugate priors under which the posterior expectation of μ is a linear combination of the prior expectation of μ and the maximum likelihood estimate y (Diaconis and Ylvisaker 1979).

Corollary 2 *Consider hierarchical Model (2.1) where $G_\mu(\cdot|\tau)$ and $H_\tau(\cdot)$ are, respectively, conjugate prior distributions of $\mu|\tau$ and τ. Under the conditions of Theorem 3, if $\lim_{n \to \infty} c_n n^{-1/2} \log^6 n = 0$, then*

$$\frac{c_n}{n} \left\| \boldsymbol{\delta}^{\mathsf{ds}}(\lambda) - \boldsymbol{\delta}^\pi \right\|_2^2 = o_p(1).$$

Furthermore, under the same conditions, if $\lim_{n \to \infty} c_n n^{-1/4} \log^6 n = 0$, then

$$c_n \left| l_n(\boldsymbol{\delta}^{\mathsf{ds}}(\lambda), \boldsymbol{\mu}) - l_n(\boldsymbol{\delta}^\pi, \boldsymbol{\mu}) \right| = o_p(1).$$

Corollary 2 is a straightforward consequence of Theorem 3 and the fact that under the hierarchical model of Equation (2.1) with conjugate priors, $\boldsymbol{\delta}^* = \boldsymbol{\delta}^\pi$.

5 Numerical Experiments

In this section we assess the performance of the NEST estimation framework for the following two compound estimation problems: estimation of Normal means with unknown variances (Section 5.1) and estimation of Sharpe ratios $\theta_i = m_i \mu_i \sqrt{\tau_i}$ for $i = 1, \ldots, n$ (Section 5.2).

5.1 Compound Estimation of Normal Means - unknown variances

We focus on the hierarchical Model of Equation (2.1) and compare seven approaches for estimating $\boldsymbol{\mu}$ when the variances $\sigma_i = 1/\tau_i$ are assumed to be unknown. These approaches can be categorized into three types: the first consists of the NEST oracle method (NEST orc), which estimates λ by minimizing the true loss, the proposed NEST method and Tweedie's formula (TF) that uses sample variances. For both NEST and TF, λ is chosen using modified cross-validation. The second are linear shrinkage methods: the group linear estimator (Grp linear) of Weinstein et al. (2018); the semi-parametric monotonically constrained SURE estimator that shrinks towards the grand mean (XKB.SG) from Xie et al. (2012); and from the same paper, the parametric SURE estimator that shrinks towards a general data driven location (XKB.M). Finally, the third type is the g-modelling approach of Gu and Koenker (2017a,b). This method is the nearest competitor to NEST as it estimates the joint prior distribution of the mean and variance using nonparametric maximum likelihood estimation (NPMLE) techniques (Kiefer and Wolfowitz 1956, Laird 1978). For the linear shrinkage methods, we use code provided by Weinstein et al. (2018) while for NPMLE we rely on the R package REBayes (Koenker and Gu 2017).

The aforementioned seven approaches are evaluated on six different simulation settings, with the goal of assessing the relative performance of the competing estimators as the het-

erogeneity in the variances σ_i^2 is varied while keeping the sample sizes m_i fixed at m. The six simulation settings can be categorized into four types: a setting where mean and variances are independent; two settings where mean and variance are correlated; a sparse setting; and two settings that represent departures from the Normal data-generating model. For each setting we set $n = 1,000$ and compute the average squared error risk for each competing estimator of $\boldsymbol{\mu}$ across 50 Monte Carlo repetitions. Figures 1 to 6 plot the relative risk which is the ratio of the average squared error risk for any competing estimator to that of NEST orc so that a ratio bigger than 1 represents a poorer risk performance of the competing estimator relative to the NEST oracle method.

The first setting, Figure 1, corresponds to the independent case. Here, for each $i = 1, \ldots, n$, $\mu_i \overset{i.i.d}{\sim} 0.7 \, N(0, .1) + 0.3 \, N(\pm 1, 3)$ and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.5, u)$ where we let u vary across six levels, $\{0.5, 1, 1.5, 2, 2.5, 3\}$. The three plots in Figure 1 show the relative risks as u varies for $m = 10, 15$ and 20 (left to right). We see that for $m = 10$, the competing methods split into two levels of performance. The group with the lowest relative risks consists of NPMLE, TF and NEST while the three linear shrinkage methods exhibit substantially higher relative risks. Moreover, we also see that as heterogeneity increases with increasing u, the gap between the two groups' relative risks increases, indicating that NPMLE, TF and the proposed NEST method are particularly useful for compound estimation of normal means when the variances are unknown and heterogeneous, and the sample size for estimating those variances are themselves small. As m increases, the performance of the three linear shrinkage methods and TF improve which is expected as there are now more replicates per unit of study to construct a relatively reliable estimate of the unknown variances. However, the performance of NEST improves too and particularly at $m = 20$ (Figure 1 right), NPMLE exhibits a slightly higher relative risk than NEST and TF.

The second setting, Figure 2, corresponds to the correlated case. The precisions $\tau_i = 1/\sigma_i^2$ are generated independently from a gamma mixture, with an even chance of drawing $\Gamma(20, rate = 20)$ or $\Gamma(20, rate = u)$ and given τ_i, the means μ_i are independently $N(\pm 0.5/\tau_i, 0.5^2)$. In this setting, the magnitude of the variances increase with u and the
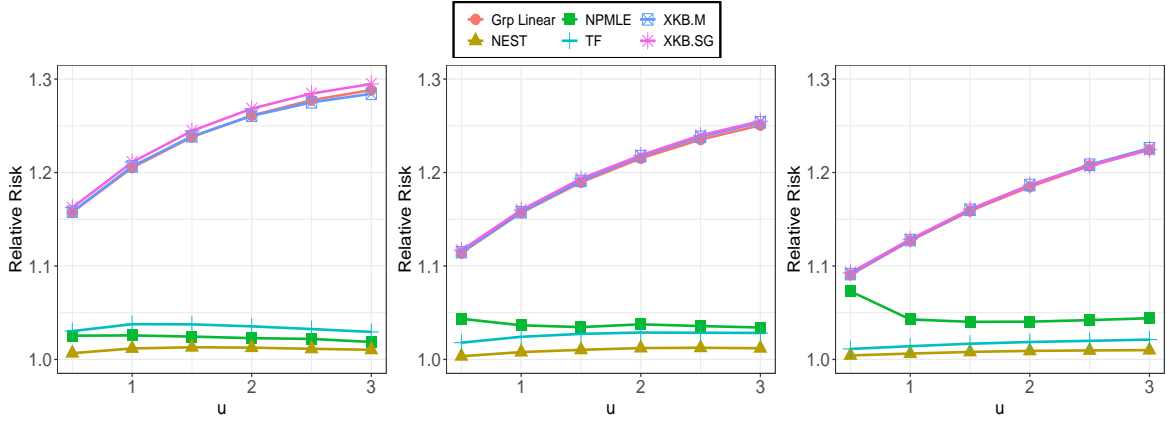
24

Figure 1: Comparison of relative risks when (μ_i, σ_i^2) are independent. Here $\mu_i \overset{i.i.d}{\sim}$ $0.7\ N(0, .1) + 0.3\ N(\pm 1, 3)$ and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.5, u)$. Plots show $m = 10, 15, 20$ left to right.

means grow with the variances. Again, for the left plot $m = 10$, the same groups as in Figure 1 perform well although Grp Linear has a lower relative risk in comparison to the other two linear shrinkage methods considered here while the relative risk of TF is slightly higher than that of NPMLE and NEST. However, the pattern as m grows is more pronounced than before, wherein NEST and TF maintain the lowest risk across $m = 15, 20$.

Figure 2: Comparison of relative risks for correlated (μ_i, τ_i). Here $\tau_i \overset{i.i.d}{\sim} 0.5\Gamma(20, \texttt{rate} = 20) + 0.5\Gamma(20, \texttt{rate} = u)$ and $\mu_i | \tau_i \overset{ind.}{\sim} N(\pm 0.5/\tau_i, 0.5^2)$. Plots show $m = 10, 15, 20$ left to right.

In the third setting, Figure 3, (μ_i, τ_i) continue to be correlated and have a conjugate prior distribution. The precisions τ_i are drawn from $\Gamma(20, rate = u)$ and conditional on

τ_i, μ_i are independently $N(0, 0.5/\tau_i)$. Under this data generating scheme, the posterior mean of μ_i is $my_i/(m+2)$ which is independent of u. This is the reason that the relative risks of the competing estimators in Figure 3 do not vary with the heterogeneity in the variances. Compared to the first two settings, we see that the linear shrinkage estimators have a relatively better performance. This is expected because in this setting the posterior mean of μ_i is indeed a linear function of the sample mean y_i. For $m = 10$ and 15, we notice that the relative risk of NEST is marginally better than the competing estimators while at $m = 20$ Grp Linear and NEST have similar risk performance.

Figure 3: Comparison of relative risks when (μ_i, τ_i) have conjugate priors. Here $\mu_i | \tau_i \overset{ind.}{\sim} N(0, 0.5/\tau_i)$ and $\tau_i \overset{i.i.d}{\sim} \Gamma(20, \texttt{rate} = u)$. Plots show $m = 10, 15, 20$ left to right.

The fourth setting, Figure 4, corresponds to the sparse case. The precisions τ_i are drawn from a gamma mixture, with an even chance of drawing $\Gamma(20, rate = 20)$ or $\Gamma(20, rate = u)$, but μ_i are only 30% likely to come from $N(\pm 0.5/\tau_i, 1)$ and 70% likely from a point mass at 0. We see similar patterns to those in Figures 1 and 2 at $m = 10$. For $m = 15$ and 20, we notice that the relative risks of the linear shrinkage methods are now higher than their levels at $m = 10$. This is not surprising for in this setting, while the risk performance of all methods have improved with larger sample sizes, NPMLE, TF and NEST exhibit a bigger improvement in risk than those of Grp Linear, XKB.M and XKB.SG. Moreover in this setting, NPMLE has a marginally better risk performance than NEST across $m = 10, 15, 20$.
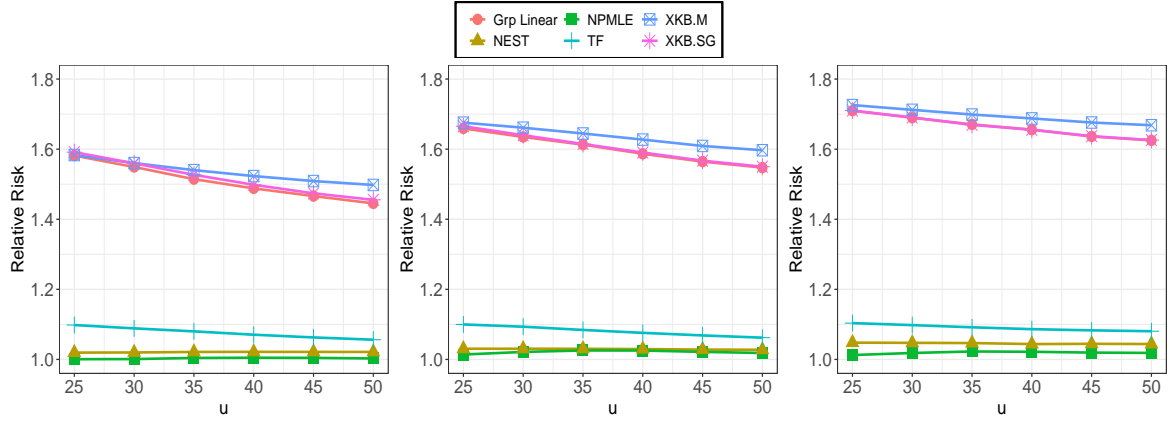
Figure 4: Comparison of relative risks when $\boldsymbol{\mu}$ is sparse. Here $\tau_i \overset{i.i.d}{\sim} 0.5\ \Gamma(20, \texttt{rate} = 20) + 0.5\ \Gamma(20, \texttt{rate} = u)$ and $\mu_i|\tau_i \overset{ind.}{\sim} 0.7\delta_{(0)} + 0.3\ N(\pm 0.5/\tau_i, 1)$. Plots show $m = 10, 15, 20$ left to right.

The fifth and sixth settings, Figures 5 and 6, correspond to the setting where the data $Y_{ij}|(\mu_i, \sigma_i^2)$ are not normally distributed. In Figure 5, $Y_{ij}|(\mu_i, \sigma_i^2)$ are generated independently from a uniform distribution between $\mu_i \pm \sqrt{3}\sigma_i$, σ_i^2 are sampled independently from $N(u, 1)$ which is truncated below at 0.1, and $\mu_i|\sigma_i \overset{ind.}{\sim} 0.8N(\sigma_i^2/4, 0.25) + 0.2N(\sigma_i^2, 1)$. For Figure
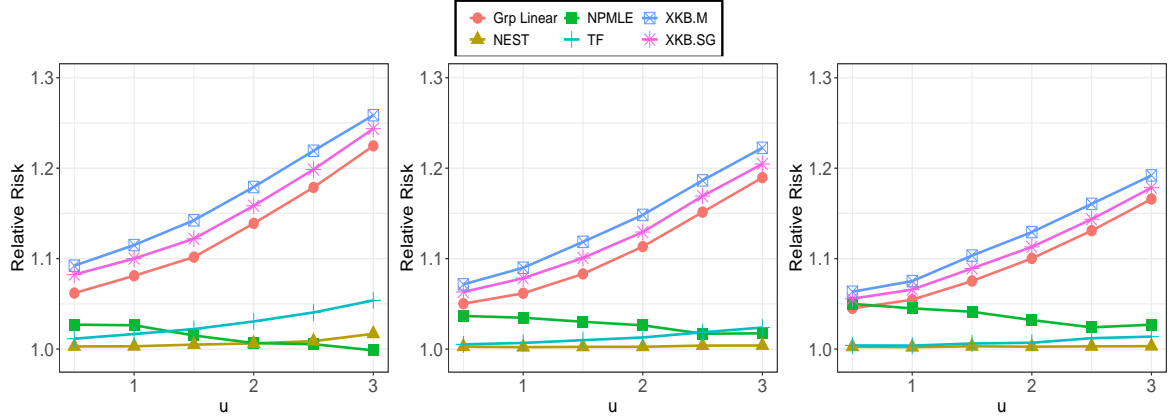

Figure 5: Comparison of relative risk for non-normal data. Here $Y_{ij}|\mu_i, \sigma_i \overset{i.i.d}{\sim} U(\mu_i - \sqrt{3}\sigma_i, \mu_i + \sqrt{3}\sigma_i)$, σ_i^2 are sampled independently from $N(u, 1)$ truncated below at 0.1 and $\mu_i|\sigma_i^2 \overset{ind.}{\sim} 0.8\ N(0.25\sigma_i^2, 0.25) + 0.2\ N(\sigma_i^2, 1)$. Plots show $m = 10, 15, 20$ left to right.

6, $Y_{ij}|(\mu_i, \sigma_i^2)$ are generated with additional non-normal noise $N(\mu_i, \sigma_i^2) + Lap(0, 1)$ and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.1, u)$ with $\mu_i|\sigma_i \overset{ind.}{\sim} 0.8N(\sigma_i^2/2, 0.5) + 0.2N(2\sigma_i^2, 2)$. Across both these settings the proposed NEST method demonstrates robustness to departures from the Normal model.

Proposition 7 in Barp et al. (2019) guarantees that, in general, the influence function of minimum KSD estimators, such as the NEST estimator, is bounded under data corruption and the behavior of the NEST estimator in Settings 5 and 6 is potentially an example of such a robustness property.
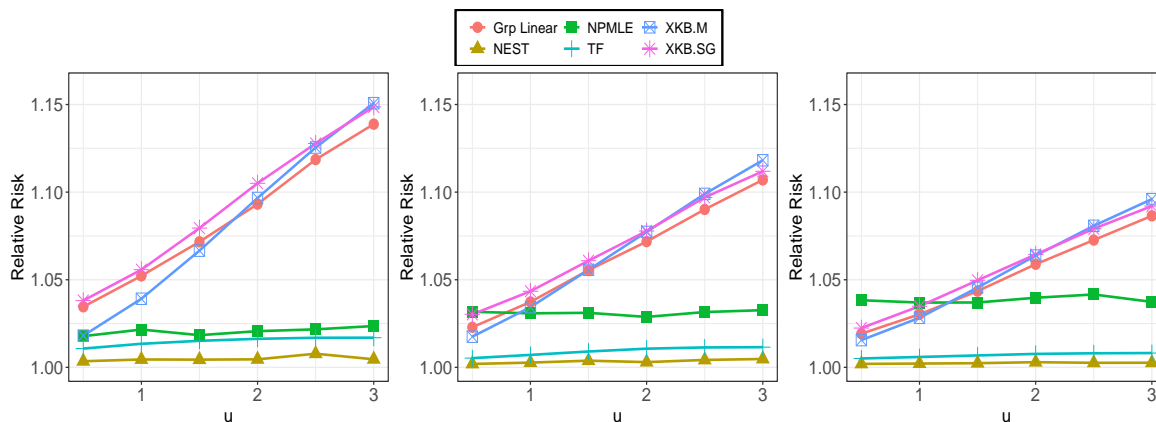

Figure 6: Comparison of relative risk for non-normal data. Here $Y_{ij}|\mu_i,\sigma_i \overset{i.i.d}{\sim} N(\mu_i,\sigma_i^2) + Lap(0,1)$, $\sigma_i^2 \overset{i.i.d}{\sim} U(0.1,u)$ and $\mu_i|\sigma_i^2 \overset{ind.}{\sim} 0.8 \; N(0.5\sigma_i^2,0.5) + 0.2 \; N(2\sigma_i^2,2)$. Plots show $m = 10, 15, 20$ left to right.

Overall, the results of the preceding six simulation settings reveal that when the variances are unknown, the NEST estimation framework enjoys a relatively better risk performance than the linear shrinkage methods and Tweedie's formula that rely on sample variances. For larger sample sizes, we observe that NEST is marginally better than NPMLE. In Section C of the supplementary material we consider additional simulation experiments wherein we evaluate the risk performance of these competing estimators for (i) unequal sample sizes m_i (Section C.1), (ii) compound estimation for location mixture of Gaussians (Section C.2) and, (iii) compound estimation in other two-parameter exponential families when the nuisance parameter is known (Section C.3).

5.2 Compound Estimation of Ratios

In this section we demonstrate the use of the NEST estimation framework for compound estimation of the n ratios $\theta_i = \sqrt{m_i}\mu_i/\sigma_i$ which represent a popular financial metric for as-

sessing mutual fund performance (see Section 6.2 for a related real data application involving compound estimation of mutual fund Sharpe ratios.).
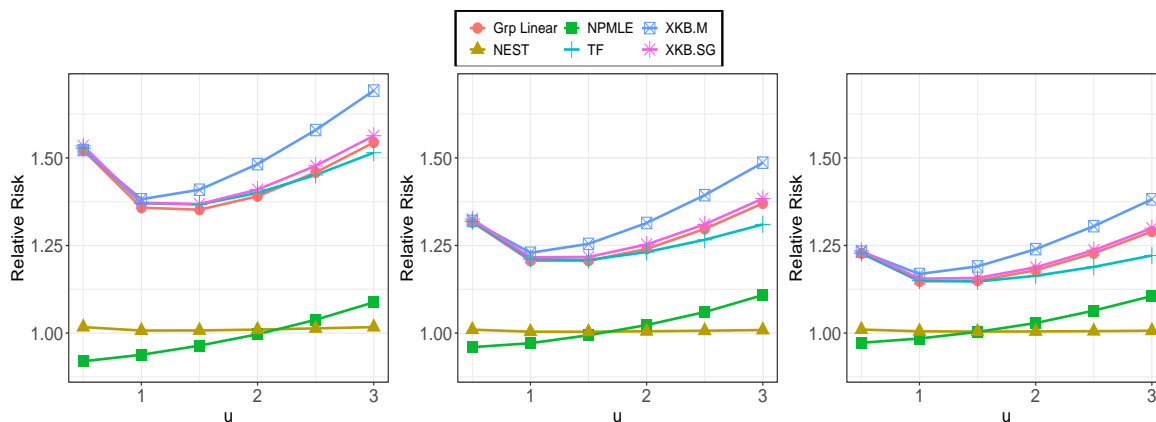

Figure 7: Comparison of relative risk for estimating $\boldsymbol{\theta}$. Here $\sigma_i^2 \overset{i.i.d}{\sim} U(0.1, u)$ and $\mu_i | \sigma_i \overset{ind.}{\sim} N(\pm 0.5\sigma_i^2, 0.5^2)$. Plots show $m = 10, 15, 20$ left to right. 1396 x 498

We evaluate the performance of the same seven methods with n fixed at 1,000 and $m_i = m$ for $i = 1, \ldots, n$. The data Y_{ij} are generated independently from $N(\mu_i, \sigma_i^2)$ and the variances σ_i^2 are simulated as in Figure 6 while the means are independently drawn from a mixture model with half chance $N(-\sigma_i^2/2, 0.5^2)$ and the other half $N(\sigma_i^2/2, 0.5^2)$. Figure 7 shows the relative risk performance of the competing estimators of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. We continue to see that NEST has a lower relative risk than the linear shrinkage methods and TF that use sample variances, while NPMLE dominates NEST for small values of u. As u increases the heterogeneity in the data grows and we see that the relative risks of the linear shrinkage methods and Tweedie's Formula across all m first decrease and then increase. The shift in the behavior of these estimators is related to the observation that as u increases, the centers of the mixture model that generates μ_i, are on average, further away from one another. This makes estimating the numerator of the ratio easier for all methods up until a point. As heterogeneity increases further, the risks of these methods that use the sample standard deviation in the denominator of θ_i are relatively worse than the risk of NEST and NPMLE.

6 Real Data Analyses

6.1 Baseball Data

We analyze the monthly data on the number of "at bats" and "hits" for all U.S Major League baseball players over the regular seasons from 2002 until 2011. In this analysis we focus on both pitchers and non-pitchers using an approach similar to that of Gu and Koenker (2017a). The data are available from the R package REBayes and have been aggregated into half seasons to produce an unbalanced panel. It includes observations on 932 players who have at least ten at bats in any half season and appear in no fewer than five half-seasons (note that there are a total of 20 half-seasons that a player can appear in).

Following Brown (2008), let the transformed batting average Y_{ij} for player $i (= 1, \ldots, n)$ at time $j (= 1, \ldots, m_i)$ be denoted by $Y_{ij} = \arcsin\left(\sqrt{\dfrac{H_{ij} + 0.25}{N_{ij} + 0.5}}\right)$ where H_{ij} denotes the number of "hits" and N_{ij} denotes the number of "at bats" at time j for player i. We assume that $Y_{ij} \sim N(\mu_i, v_{ij}^2/\tau_i)$ where $\mu_i = \arcsin(\sqrt{p_i})$, p_i being player i's batting success probability, and $v_{ij}^2 = 1/(4N_{ij})$. Here $1/\tau_i$ are player specific scale parameters as described in Gu and Koenker (2017a). Under this setup, the sufficient statistics are

$$
\begin{aligned}
\hat{\mu}_i &= \left(\sum_{j=1}^{m_i} 1/v_{ij}^2\right)^{-1} \sum_{j=1}^{m_i} Y_{ij}/v_{ij}^2 \sim N(\mu_i, v_i^2/\tau_i) \text{ with } v_i^2 = \left(4\sum_{j=1}^{m_i} N_{ij}\right)^{-1}, \\
S_i^2 &= \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (Y_{ij} - \hat{\mu}_i)^2/v_{ij}^2 \text{ with } (m_i - 1)S_i^2\tau_i \sim \mathcal{X}_{m_i-1}^2.
\end{aligned}
$$

In this analysis, the goal is to use the 2002-2011 data to predict the batting averages of the players in 2012. Players are divided into three categories: all, non-pitchers, and pitchers. We consider the following seven estimators of μ_i: two non-parametric maximum likelihood based estimators, denoted NPMLE-Indep and NPMLE-Dep which assume, respectively, independent and dependent priors on $(\mu_i, 1/\tau_i)$, the sufficient statistics $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ of $\boldsymbol{\mu}$, the grand mean across all players in the 2011 season $\bar{Y}^{2011} = n^{-1}\sum_{i=1}^{n} Y_{i,2011}$, the proposed NEST estimator and its oracle counterpart (NEST orc), and the naive estimator

Table 1: Performance of the competing estimators relative to the performance of the naive estimator $\boldsymbol{Y}_{2011}$. Here R-$TSE(\boldsymbol{\delta}) = TSE(\boldsymbol{\delta})/TSE(\boldsymbol{Y}_{2011})$ with similar definitions for R-NSE and R-TSEp. The smallest two relative errors are bolded

		NPMLE-Indep	NPMLE-Dep	$\hat{\boldsymbol{\mu}}$	$\bar{Y}^{2011}$	NEST	NEST orc
All							
n for estimation: 932	R-TSE	0.463	0.469	0.352	1.958	**0.349**	**0.348**
n for prediction: 370	R-NSE	**0.668**	0.674	0.676	1.495	0.670	**0.669**
	R-TSEp	0.582	0.591	0.501	1.783	**0.496**	**0.495**
Nonpitchers							
n for estimation: 792	R-TSE	0.535	0.546	**0.503**	0.551	**0.528**	**0.528**
n for prediction: 325	R-NSE	**0.656**	**0.661**	0.679	0.973	0.672	0.672
	R-TSEp	0.651	0.663	**0.619**	0.682	**0.644**	**0.644**
Pitchers							
n for estimation: 140	R-TSE	0.659	0.655	0.629	0.804	**0.628**	**0.624**
n for prediction: 45	R-NSE	0.659	0.659	**0.649**	0.769	0.652	**0.642**
	R-TSEp	**0.124**	**0.117**	0.133	0.337	0.144	0.128

that uses 2011 batting averages $\boldsymbol{Y}_{2011} = (Y_{1,2011}, \ldots, Y_{n,2011})$. To assess how well these methods predict 2012 batting averages $\boldsymbol{Y}_{2012} = (Y_{1,2012}, \ldots, Y_{n,2012})$, we consider three criteria for evaluating any estimate δ_i of μ_i: total squared error from Brown (2008) and defined as $TSE(\boldsymbol{\delta}) = \sum_{i=1}^{n}\left\{(Y_{i,2012} - \delta_i)^2 - (4N_{i,2012})^{-1}\right\}$, normalized squared error from Gu and Koenker (2017a) and defined as $NSE(\boldsymbol{\delta}) = \sum_{i=1}^{n}\left\{4N_{i,2012}(Y_{i,2012} - \delta_i)^2\right\}$, and total squared error on a probability scale from Jiang et al. (2010) which is defined as $TSEp(\hat{\boldsymbol{p}}) = \sum_{i=1}^{n}\left\{(p_{i,2012} - \hat{p}_i)^2 - p_{i,2012}(1 - p_{i,2012})(4N_{i,2012})^{-1}\right\}$. Here $\hat{p}_i = \sin^2(\delta_i)$ and $p_{i,2012} = \sin^2(Y_{i,2012})$.

In Table 1, we report the performance of the competing estimators relative to the performance of the naive estimator $\boldsymbol{Y}_{2011}$ wherein R-$TSE(\boldsymbol{\delta}) = TSE(\boldsymbol{\delta})/TSE(\boldsymbol{Y}_{2011})$ with similar definitions for R-NSE and R-TSE_p. Thus, a smaller value of R-TSE, R-NSE or R-TSE_p indicates a relatively better prediction error. Across "All" and "Nonpitchers", NEST exhibits the best relative risk for two of the three performance metrics. It is interesting to note that the sufficient statistics $\hat{\boldsymbol{\mu}}$ are quite competitive in this example while NPMLE with independent priors dominate the one with dependent priors across "All" and "Nonpitchers". The compound estimation problem for "Pitchers" is an example of a setting where n is relatively small and NEST demonstrates a better risk performance than NPMLE for total squared error and normalized squared error losses.

6.2 Mutual Fund Sharpe Ratios

In this section we analyze a dataset on $n_1 = 5{,}000$ monthly mutual fund returns spanning 12 months from January 2014 to December 2014. This data are sourced from the Wharton research data services (Wharton School 1993). The goal in this analysis is to use Sharpe ratios constructed using the data on the first $m_1 = 6$ months, January 2014 - June 2014, to predict the corresponding Sharpe ratios for the next 6 months. Formally, let Y_{ij} denote the excess return of fund $i(= 1, \ldots, n_1)$ in month $j(= 1, \ldots, m_1)$ over the return on the 3 month treasury yield. Denote $\bar{Y}_i = m_1^{-1} \sum_{j=1}^{m_1} Y_{ij}$, $S_i^2 = (m_1 - 1)^{-1} \sum_{j=1}^{m_1} (Y_{ij} - \bar{Y}_i)^2$ and $\delta_i^{\mathsf{naive}} = \bar{Y}_i / \sqrt{S_i^2}$ to be, respectively, the sample mean, the sample variance and the observed Sharpe ratio of the monthly excess returns. Of the 5,000 funds available during these first 6 months, there are $n_2 = 4{,}958$ funds that appear in the next 6 months, July 2014 - December 2014, and have at least 3 months of returns available during this period. For our prediction, we consider these n_2 funds to assess the performance of various estimators for predicting $\theta_i = \mu_i / \sigma_i$ where μ_i and σ_i are the sample mean and sample standard deviation of the excess returns of the n_2 funds during the next 6 months.

We consider the following estimators of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$: NEST, NEST orc, Tweedie's formula (TF), Grp Linear and XKB.SG from Section 5.1, as well as NPMLE-Indep, NPMLE-Dep from Section 6.1. Additionally, we consider the SURE estimator XKB.G from Xie et al. (2012) that, unlike XKB.SG, imposes a Normal prior on μ and shrinks towards the grand mean. Note that for predicting θ_i, TF, Grp Linear, XKB.SG and XKB.G rely on the sample variances S_i^2. To evaluate the performance of these estimators for predicting $\boldsymbol{\theta}$, we consider the following three criteria with $m_{i,2} \in [3, 6]$: Total Squared Error : $TSE(\boldsymbol{\delta}) = \sum_{i=1}^{n_2} (\theta_i - \delta_i)^2$; weighted Squared Error : $WSE(\boldsymbol{\delta}) = \sum_{i=1}^{n_2} m_{i,2} (\theta_i - \delta_i)^2$; and weighted Absolute Error : $WAE(\boldsymbol{\delta}) = \sum_{i=1}^{n_2} m_{i,2} |1 - \delta_i / \theta_i|$. In Table 2, we present the performance of the competing estimators relative to the performance of the naive estimator $\boldsymbol{\delta}^{\mathsf{naive}} = (\delta_i^{\mathsf{naive}} : 1 \leq i \leq n)$ so that a smaller value of R-TSE, R-WSE or R-WAE indicates a relatively better prediction error. NPMLE-Indep and NPMLE-Dep revealed convergence issues on this data and so we do not include them in Table 2. Along all performance measures, NEST has the smallest

Table 2: Performance of the competing estimators relative to the performance of the naive estimator δ^{naive}. Here R-$TSE(\delta) = TSE(\delta)/TSE(\delta^{\mathsf{naive}})$ with similar definitions for R-WSE and R-WAE. The smallest two relative errors are bolded

	n_1	n_2	Grp Linear	XKB.G	XKB.SG	TF	NEST	NEST orc
R - TSE	5000	4958	0.896	0.931	0.922	0.997	**0.687**	**0.687**
R - WSE	5000	4958	0.893	0.930	0.920	0.997	**0.685**	**0.685**
R - WAE	5000	4958	1.067	0.974	0.983	1.002	**0.785**	**0.783**

relative risk among all competing estimators considered in this example. With respect to the weighted Absolute Error, Grp Linear appears to be doing relatively worse in predicting the small Sharpe ratios and exhibits an R-WAE bigger than 1. When compared against the three linear shrinkage methods considered here and Tweedie's formula, NEST demonstrates an overall value in joint shrinkage estimation of the means μ_i and the variances σ_i^2 for predicting θ.

References

Abramovich, F., Y. Benjamini, D. L. Donoho, and I. M. Johnstone (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist. 34*, 584–653.

Banerjee, T., Q. Liu, G. Mukherjee, and W. Sun (2020). A general framework for empirical bayes estimation in discrete linear exponential family. *arXiv preprint arXiv:1910.08997 (accepted in Journal of Machine Learning Research).*

Barp, A., F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey (2019). Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pp. 12964–12976.

Basu, P., T. T. Cai, K. Das, and W. Sun (2017). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association 0*(ja), 0–0.

Benjamini, Y. and Y. Hochberg (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics 24*, 407–418.

Benjamini, Y. and D. Yekutieli (2011). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association 100*(469), 71–81.

Berger, J. O. (1976, January). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist. 4*(1), 223–226.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013, 04). Valid post-selection inference. *Ann. Statist. 41*(2), 802–837.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics*, 113–152.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics 37*, 1685–1704.

Brown, L. D., E. Greenshtein, and Y. Ritov (2013). The poisson compound decision problem revisited. *Journal of the American Statistical Association 108*(502), 741–749.

Brown, S. J., W. Goetzmann, R. G. Ibbotson, and S. A. Ross (1992). Survivorship bias in performance studies. *The Review of Financial Studies 5*(4), 553–580.

Cai, J., X. Han, Y. Ritov, and L. Zhao (2021). Nonparametric empirical bayes estimation and testing for sparse and heteroscedastic signals. *arXiv preprint arXiv:2106.08881*.

Cai, T. T. and W. Sun (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc. 104*, 1467–1481.

Castillo, I. and A. van der Vaart (2012, 08). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist. 40*(4), 2069–2101.

Chiaretti, S., X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa (2004, 4). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood 103*(7), 2771–2778.

Chwialkowski, K., H. Strathmann, and A. Gretton (2016). A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings.

Diaconis, P. and D. Ylvisaker (1979). Conjugate priors for exponential families. *The Annals of statistics*, 269–281.

Donoho, D. L. and J. M. Jonhstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425.

Dyson, F. (1926, 07). A Method for Correcting Series of Parallax Observations. *Monthly Notices of the Royal Astronomical Society 86*(9), 686–706.

Eddington, A. S. (1940, 03). The Correction of Statistics for Accidental Error. *Monthly Notices of the Royal Astronomical Society 100*(5), 354–361.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci. 23*, 1–22.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. (2014). Two modeling strategies for empirical bayes estimation. *Statistical science 29*(2), 285–301.

Efron, B. and C. N. Morris (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association 70*(350), 311–319.

Erickson, S. and C. Sabatti (2005). Empirical Bayes estimation of a sparse vector of gene expression change. *Statistical applications in genetics and molecular biology 4*(1), 1132.

Gu, J. and R. Koenker (2017a). Empirical bayesball remixed: Empirical bayes methods for longitudinal data. *Journal of Applied Econometrics 32*(3), 575–599.

Gu, J. and R. Koenker (2017b). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics 35*(1), 1–16.

He, L., S. K. Sarkar, and Z. Zhao (2015). Capturing the severity of type ii errors in high-dimensional multiple testing. *Journal of Multivariate Analysis 142*, 106 – 116.

Henderson, N. C. and M. A. Newton (2016). Making the cut: improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*, 1467–9868.

Ignatiadis, N. and S. Wager (2019). Covariate-powered empirical bayes estimation. In *Advances in Neural Information Processing Systems*, pp. 9620–9632.

James, G. M., P. Radchenko, and B. Rava (2020). Irrational exuberance: Correcting bias in probability estimates. *Journal of the American Statistical Association*, 1–14.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif., pp. 361–379. University of California Press.

Jiang, W. and C.-H. Zhang (2009, 08). General maximum likelihood empirical bayes estimation of normal means. *Ann. Statist. 37*(4), 1647–1684.

Jiang, W., C.-H. Zhang, et al. (2010). Empirical bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pp. 263–273. Institute of Mathematical Statistics.

Jing, B.-Y., Z. Li, G. Pan, and W. Zhou (2016). On sure-type double shrinkage estimation. *Journal of the American Statistical Association 111*(516), 1696–1704.

Johnstone, I. M. and B. W. Silverman (2004). Needles and straw in haystacks: empirical Bayes estimates to possibly sparse sequences. *Annals of Statistics 32*(4), 1594–1649.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Koenker, R. and J. Gu (2017). Rebayes: Empirical bayes mixture methods in r. *Journal of Statistical Software 82*(8), 1–26.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association 109*(506), 674–685.

Kou, S. C. and J. J. Yang (2017). *Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models*, Chapter 25, pp. 249–284. Cham: Springer International Publishing.

Kwon, Y. and Z. Zhao (2018). On f-modelling based empirical bayes estimation of variances. *arXiv preprint arXiv:1806.06377*.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association 73*(364), 805–811.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016, 06). Exact post-selection inference, with application to the lasso. *Ann. Statist. 44*(3), 907–927.

Liu, Q., J. D. Lee, and M. I. Jordan (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Liu, Q. and D. Wang (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pp. 2378–2386.

Oates, C. J., M. Girolami, and N. Chopin (2017). Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(3), 695–718.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. on Math. Statistic. and Prob. 1*, 157–163.

Saha, S. and A. Guntuboyina (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics 48*(2), 738–762.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Volume 162. John Wiley & Sons.

Soloff, J. A., A. Guntuboyina, and B. Sen (2021). Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *arXiv preprint arXiv:2109.03466*.

Sun, W. and A. C. McLain (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association 107*(498), 673–687.

Tan, Z. (2015). Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli 21*, 574–603.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences 98*(9), 5116–5121.

Weinstein, A., W. Fithian, and Y. Benjamini (2013). Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association 108*(501), 165–176.

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2018). Group-linear empirical bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association 0*(0), 1–13.

Wharton School (1993). Wharton research data services. *https://wrds-web.wharton.upenn.edu/wrds/*.

Xie, X., S. Kou, and L. D. Brown (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association 107*(500), 1465–1479.

Yang, J., Q. Liu, V. Rao, and J. Neville (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pp. 5561–5570.

Zhang, X. and A. Bhattacharya (2017). Empirical Bayes, sure, and sparse normal mean models. Preprint.

Supplementary Material for "Nonparametric Empirical Bayes Estimation On Heterogeneous Data"

In Section A we collect the proofs of the theoretical results in the paper. In Section B we discuss extensions of our methodology to several well known members in the two-parameter exponential family when the nuisance parameter is known. In Section C, we provide additional numerical experiments for the following cases: unequal sample sizes m_i (Section C.1), compound estimation of Normal means with known variances (Section C.2) and compound estimation for Gamma and Weibull mixtures (Section C.3).

A   Proofs

A.1   Proof of Proposition 1

Recall from Definition 1 that

$$\delta_i^{\mathsf{TF}} := \delta^{\mathsf{TF}}(y_i, s_i^2, m_i) = y_i + \frac{s_i^2}{m_i} w_1(y_i, s_i^2; m_i),$$

where $w_1(y, s^2; m) := \frac{\partial}{\partial y} \log f_m(y, s^2)$. Since the n study units are independent, we will focus on unit i.

We first note that under the squared error loss, δ_i^π is the Bayes estimate of μ_i in Model (2.1) and δ_i^{TF} is an approximation to δ_i^π. So, the Bayes risk of δ_i^π is strictly less than the Bayes risk of δ_i^{TF}. This establishes the inequality in the left hand side of Proposition 1. To prove the inequality in the right hand side of Proposition 1, we proceed as follows.

Denote $w_{1i} := w_1(y_i, s_i^2; m_i)$. We have,

$$
\begin{aligned}
\mathbb{E}(\mu_i - \delta_i^{\mathsf{TF}})^2 &= \frac{1}{m_i}\mathbb{E}(1/\tau_i) + \frac{1}{m_i^2}\mathbb{E}(S_i^2 w_{1i})^2 + \frac{2}{m_i}\mathbb{E}(Y_i - \mu_i)S_i^2 w_{1i} \\
&= \mathbb{E}(\mu_i - Y_i)^2 + \frac{1}{m_i^2}\mathbb{E}(S_i^2 w_{1i})^2 + \frac{2}{m_i^2}\mathbb{E}\frac{S_i^2}{\tau_i}w_{1i}',
\end{aligned}
\tag{A.1}
$$

where $w_{1i}' := w_1'(y_i, s_i^2; m_i)$ and $w_1'(y, s^2; m) = \dfrac{\partial}{\partial y}w_1(y, s^2; m)$. The equality in Equation (A.1) follows from integration by parts and the fact that $Y_i|\mu_i, \tau_i \sim N(\mu_i, 1/(m_i\tau_i))$. Consider the term $T_1 := \dfrac{1}{m_i^2}\mathbb{E}(S_i^2 w_{1i})^2 + \dfrac{2}{m_i^2}\mathbb{E}\left(\dfrac{S_i^2}{\tau_i}w_{1i}'\right)$ and note that

$$
\begin{aligned}
T_1 &= \frac{1}{m_i^2}\mathbb{E}\left(S_i^2 w_{1i}\right)^2 + \frac{1}{m_i^2}\mathbb{E}\left[(S_i^2)^2 w_{1i}'\right] + \frac{2}{m_i^2}\mathbb{E}\left(\frac{S_i^2}{\tau_i}w_{1i}'\right) - \frac{1}{m_i^2}\mathbb{E}\left[(S_i^2)^2 w_{1i}'\right] \\
&= \frac{1}{m_i^2}\mathbb{E}\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right]w_{1i}'.
\end{aligned}
\tag{A.2}
$$

The equality in Equation (A.2) follows because, dropping subscript i,

$$
\mathbb{E}\left[(S^2)^2(w_1^2 + w_1')\right] = \mathbb{E}\left[(S^2)^2 \frac{f_{m,(1)}''(Y, S^2)}{f_m(Y, S^2)}\right] = 0,
$$

where $f_{m,(1)}''(y, s^2)$ is the second order partial derivative of $f_m(y, s^2)$ with respect to y. Now, we can re-write Equation (A.2) as

$$
T_1 = \frac{1}{m_i^2}\mathbb{E}_{\mu_i, \tau_i}\mathbb{E}_{Y_i, S_i^2|\mu_i, \tau_i}\left\{\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right]w_{1i}'\right\} = \frac{1}{m_i^2}\mathbb{E}_{\mu_i, \tau_i}\left(T_2\right) + \frac{1}{m_i^2}\mathbb{E}_{\mu_i, \tau_i}\left(T_3\right),
\tag{A.3}
$$

where $\mathbb{E}_{\mu, \tau}$ is the expectation with respect to the joint distribution of (μ, τ), $\mathbb{E}_{Y, S^2|\mu, \tau}$ is the expectation with respect to the joint distribution of (Y, S^2) conditional on (μ, τ) and

$$
\begin{aligned}
T_2 &= \mathbb{E}_{Y_i, S_i^2|\mu_i, \tau_i}\left\{\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right]w_{1i}'\Big| S_i^2 < \frac{2}{\tau_i}\right\}\mathbb{P}\left(S_i^2 < \frac{2}{\tau_i}\Big|\tau_i\right), \\
T_3 &= \mathbb{E}_{Y_i, S_i^2|\mu_i, \tau_i}\left\{\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right]w_{1i}'\Big| S_i^2 > \frac{2}{\tau_i}\right\}\mathbb{P}\left(S_i^2 > \frac{2}{\tau_i}\Big|\tau_i\right).
\end{aligned}
$$

Denote $p := \mathbb{P}(S_i^2 > 2/\tau_i \mid \tau_i)$ and let c_i be the partial derivative of $w_1(y, s^2; m)$ with respect to y and evaluated at $(y_i, 2/\tau_i, m_i)$. Now, using equations (A.2) and (A.3) in Equation (A.1),

2

we get

$$\mathbb{E}(\mu_i - \delta_i^{\mathsf{TF}})^2 = \mathbb{E}(\mu_i - Y_i)^2 + \frac{1}{m_i^2}\mathbb{E}_{\mu_i,\tau_i}(T_2 + T_3). \tag{A.4}$$

We will show that $T_2 + T_3 < 0$ which will be enough to prove the statement of Proposition 1 using Equation (A.4).

We first state a few results that are straightforward consequences of Model (2.1) and the assumptions of Proposition 1. We have,

1. Under Model 2.1, $(m_i - 1)S_i^2\tau_i \sim \chi^2_{m_i-1}$.

2. Additionally,
$$\mathbb{E}_{Y_i,S_i^2|\mu_i,\tau_i}\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right] = \frac{m_i - 3}{(m_i - 1)\tau_i^2} > 0,$$

   since $m_i > 3$ in the statement of Proposition 1.

3. Since $f_m(y, s^2)$ is a log-concave density, $w'_{1i} \leq 0$ and so $T_2 \leq 0$ while $T_3 \geq 0$.

Assume, without loss of generality, $w'_{1i} < 0$. Since $\mathbb{E}_{Y_i,S_i^2|\mu_i,\tau_i}[2S_i^2/\tau_i - (S_i^2)^2] > 0$, we have

$$(1-p)\mathbb{E}_{Y_i,S_i^2|\mu_i,\tau_i}\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\Big|S_i^2 < \frac{2}{\tau_i}\right] > -p\mathbb{E}_{Y_i,S_i^2|\mu_i,\tau_i}\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\Big|S_i^2 > \frac{2}{\tau_i}\right] > 0. \tag{A.5}$$

Furthermore, as $w'_{1i} < 0$ and w_{1i} is a non-decreasing function of s_i^2,

$$T_2 \leq (1-p)\mathbb{E}_{Y_i,S_i^2|\mu_i,\tau_i}\left\{\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right]c_i\Big|S_i^2 < \frac{2}{\tau_i}\right\}, \tag{A.6}$$

Note that $c_i \leq 0$, and as defined earlier, it is the partial derivative of $w_1(y, s^2; m)$ with respect to y and evaluated at $(y_i, 2/\tau_i, m_i)$. Therefore, using equations (A.5) and (A.6),

$$T_2 \leq (1-p)\mathbb{E}_{Y_i,S_i^2|\mu_i,\tau_i}\left\{\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right]c_i\Big|S_i^2 < \frac{2}{\tau_i}\right\} < -p\mathbb{E}_{Y_i,S_i^2|\mu_i,\tau_i}\left\{\left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2\right]c_i\Big|S_i^2 > \frac{2}{\tau_i}\right\} < 0. \tag{A.7}$$

3

Now, we consider the term T_3. Recall that

$$T_3 = p\mathbb{E}_{Y_i, S_i^2 | \mu_i, \tau_i} \left\{ \left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2 \right] w_{1i}' \Big| S_i^2 > \frac{2}{\tau_i} \right\} > 0.$$

Since $w_{1i} < 0$ and w_{1i} is a non-decreasing function of s_i^2,

$$0 < T_3 \leq -p\mathbb{E}_{Y_i, S_i^2 | \mu_i, \tau_i} \left\{ \left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2 \right] |c_i| \Big| S_i^2 > \frac{2}{\tau_i} \right\}. \tag{A.8}$$

So, using equations (A.7) and (A.8)

$$T_2 + T_3 < -p\mathbb{E}_{Y_i, S_i^2 | \mu_i, \tau_i} \left\{ \left[\frac{2S_i^2}{\tau_i} - (S_i^2)^2 \right] \left(c_i + |c_i| \right) \Big| S_i^2 > \frac{2}{\tau_i} \right\} = 0,$$

Hence the desired result follows from the display above and Equation (A.4).

A.2   Proof of Proposition 2

Proposition 2 follows by recalling that under the hierarchical model of equation (2.1),

$$f_{m_i}(y_i, s_i^2 | \mu_i, \tau_i) \propto \exp\left\{ -\frac{\tau_i}{2} \left[m_i y_i^2 + (m_i - 1)s_i^2 \right] + m_i \tau_i \mu_i y_i - \frac{m_i}{2} \tau_i \mu_i^2 + \frac{m_i - 3}{2} \log s_i^2 \right\}.$$

Therefore from Bayes theorem,

$$f_{m_i}(\mu_i, \tau_i | y_i, s_i^2) = \frac{f_{m_i}(y_i, s_i^2 | \mu_i, \tau_i)}{f_{m_i}(y_i, s_i^2)} g(\mu_i | \tau_i) h(\tau_i) \propto \exp\left\{ \boldsymbol{\eta}_i^T \boldsymbol{T}(\mu_i, \tau_i) - A(\boldsymbol{\eta}_i) \right\} g(\mu_i | \tau_i) h(\tau_i).$$

Here $\boldsymbol{\eta}_i = (m_i y_i, -m_i y_i^2 - (m_i - 1)s_i^2) := (\eta_{1i}, \eta_{2i})$, $\boldsymbol{T}(\mu_i, \tau_i) = (\tau_i \mu_i, \tau_i/2)$ and

$$
\begin{aligned}
A(\boldsymbol{\eta}_i) &= -0.5(m_i - 3)\log\gamma(\eta_{1i}, \eta_{2i}) + \log f_{m_i}\left\{ m_i^{-1}\eta_{1i}, \gamma(\eta_{1i}, \eta_{2i}) \right\}, \\
\gamma(\eta_{1i}, \eta_{2i}) &= \frac{-\eta_{2i} - m_i^{-1}\eta_{1i}^2}{m_i - 1},
\end{aligned}
$$

with $f_m(y, s^2) = \int \int f_m(y, s^2 | \mu, \tau) g_\mu(\mu | \tau) h_\tau(\tau) \mathrm{d}\mu \mathrm{d}\tau$ being the marginal density function of (Y, S^2).

Corollary 1 is a consequence of the properties of exponential family of distributions and follows from proposition 2 under the squared error loss. From proposition 2 and dropping subscript i, we have,

$$\hat{\tau}^\pi := \hat{\tau}^\pi(y, s^2, m) = \mathbb{E}(\tau | y, s^2, m) = 2\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_2} = \frac{m-3}{(m-1)s^2} - \frac{2}{m-1}w_2(y, s^2; m).$$

Furthermore, with $\zeta = \tau\mu$,

$$\begin{aligned} \hat{\zeta}^\pi := \hat{\zeta}^\pi(y, s^2, m) &= \mathbb{E}(\zeta | y, s^2, m) = \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_1} = \frac{(m-3)y}{(m-1)s^2} + m^{-1}w_1(y, s^2; m) - 2yw_2(y, s^2; m) \\ &= y\mathbb{E}(\tau | y, s^2, m) + m^{-1}w_1(y, s^2; m). \end{aligned}$$

A.3 Proof of Proposition 3

We will first collect a few notations that will be used throughout the proof. Denote $w_{1i} := w_1(y_i, s_i^2; m_i)$, $w_{2i} := w_2(y_i, s_i^2; m_i)$ and $\gamma_i := \gamma(y_i, s_i^2, m_i)$. Let $w_1'(y, s^2; m) = \frac{\partial}{\partial y}w_1(y, s^2; m)$ and denote $w_{1i}' := w_1'(y_i, s_i^2; m_i)$. Similarly, $\nu(y, s^2; m) = \frac{1}{f_m(y, s^2)}\frac{\partial^2}{\partial y^2}f_m(y, s^2)$ and denote $\nu_i := \nu(y_i, s_i^2; m_i)$. Finally, let $\gamma'(y, s^2, m) = \frac{\partial}{\partial y}\gamma(y, s^2, m)$ and denote $\gamma_i' := \gamma'(y_i, s_i^2, m_i)$.

The proof of Proposition 3 will use the following two lemmata.

Lemma 1 *Suppose $f_m(y, s^2)$ is a log concave density and Assumption 1(c) holds. Then under Model (2.1) and $m_i > 7$, we have,*

$$\mathbb{E}\left\{\left[\frac{2S_i^2}{\tau_i}\left(1 - \gamma_i\right) - \left(S_i^2\right)^2\left(1 - \gamma_i^2\right)\right]w_{1i}'\right\} > 0.$$

Lemma 2 *Under Assumptions 1, 2 and Model (2.1), we have,*

$$\frac{1}{m_i^2}\mathbb{E}\left[S_i^2\left(S_i^2\gamma_i^2\nu_i + 2\frac{w_{1i}\gamma_i'}{\tau_i}\right)\right] \leq 0.$$

Lemmata 1 and 2 are proved in Sections A.3.1 and A.3.2 respectively. We now prove Propo-

sition 3. Recall from Definition 2 that the oracle NEST estimator for μ_i is

$$\delta_i^* = y_i + \frac{s_i^2}{m_i}\gamma(y_i, s_i^2, m_i)w_1(y_i, s_i^2, m_i),$$

where

$$\gamma(y_i, s_i^2, m_i) = \frac{m_i - 1}{m_i - 3 - 2s_i^2 w_2(y_i, s_i^2; m_i)}.$$

Since the n study units are independent, we will focus on unit i.

Under the squared error loss, δ_i^π is the Bayes estimate of μ_i in Model (2.1) and δ_i^* is an approximation to δ_i^π. So, the Bayes risk of δ_i^π is strictly less than the Bayes risk of δ_i^*. This establishes the strict inequality on the left hand side of Proposition 3. Together with Proposition 1, we only need to show $r(\boldsymbol{\delta}^*, \mathcal{G}) < r(\boldsymbol{\delta}^{\mathsf{TF}}, \mathcal{G})$. First note that

$$
\begin{aligned}
r(\boldsymbol{\delta}^{\mathsf{TF}}, \mathcal{G}) - r(\boldsymbol{\delta}^*, \mathcal{G}) &= \frac{1}{m_i^2}\mathbb{E}\Big[(S_i^2)^2\Big(1 - \gamma_i^2\Big)w_{1i}^2\Big] + \frac{2}{m_i}\mathbb{E}\Big[(Y_i - \mu_i)S_i^2\Big(1 - \gamma_i\Big)w_{1i}\Big] \\
&= \frac{1}{m_i^2}\mathbb{E}\Big[(S_i^2)^2\Big(1 - \gamma_i^2\Big)w_{1i}^2\Big] + \frac{2}{m_i^2}\mathbb{E}\Big\{\frac{S_i^2}{\tau_i}\Big[(1 - \gamma_i)w_{1i}' - w_{1i}\gamma_i'\Big]\Big\}. \quad \text{(A.9)}
\end{aligned}
$$

The equality in equation (A.9) follows from integration by parts and the fact that $Y_i \sim N\left(\mu_i, \frac{1}{m_i\tau_i}\right)$. We can re-write Equation (A.9) as,

$$
\begin{aligned}
\frac{1}{m_i^2}\mathbb{E}\Big[(S_i^2)^2\Big(1 - \gamma_i^2\Big)w_{1i}^2\Big] &+ \frac{1}{m_i^2}\mathbb{E}\Big[(S_i^2)^2\Big(1 - \gamma_i^2\Big)w_{1i}'\Big] - \frac{2}{m_i^2}\mathbb{E}\Big(\frac{S_i^2}{\tau_i}w_{1i}\gamma_i'\Big) \\
&+ \frac{1}{m_i^2}\mathbb{E}\Big\{\Big[\frac{2S_i^2}{\tau_i}\Big(1 - \gamma_i\Big) - \Big(S_i^2\Big)^2\Big(1 - \gamma_i^2\Big)\Big]w_{1i}'\Big\}.
\end{aligned}
\quad \text{(A.10)}
$$

From Lemma 1, the last term in Equation (A.10) is positive. Let us consider the first three terms in Equation (A.10) and denote them by,

$$T := \frac{1}{m_i^2}\mathbb{E}\Big[(S_i^2)^2\Big(1 - \gamma_i^2\Big)w_{1i}^2\Big] + \frac{1}{m_i^2}\mathbb{E}\Big[(S_i^2)^2\Big(1 - \gamma_i^2\Big)w_{1i}'\Big] - \frac{2}{m_i^2}\mathbb{E}\Big(\frac{S_i^2}{\tau_i}w_{1i}\gamma_i'\Big).$$

As shown in the proof of Proposition 1, $\mathbb{E}[(S^2)^2(w_1^2 + w_1')] = 0$. The above display involving

6

the term T can be written as

$$T = -\frac{1}{m_i^2}\mathbb{E}\Big[(S_i^2)^2\gamma_i^2\nu_i\Big] - \frac{2}{m_i^2}\mathbb{E}\Big(\frac{S_i^2}{\tau_i}w_{1i}\gamma_i'\Big). \tag{A.11}$$

From Lemma 2, the term T in Equation (A.11) is non-negative. This establishes the inequality on the right hand side of $r(\boldsymbol{\delta}^*, \mathcal{G})$ in Proposition 3 and completes the proof.

A.3.1 Proof of Lemma 1

Denote

$$Z_i := \frac{2S_i^2}{\tau_i}(1 - \gamma_i) - (S_i^2)^2(1 - \gamma_i^2) = \frac{Q_i}{\tau_i},$$

where $Q_i = 2S_i^2(1 - \gamma_i) - (S_i^2)^2(1 - \gamma_i^2)\tau_i$. Since $f_m(y, s^2)$ is a log-concave density, $w_{1i}' \leq 0$. Suppose, without loss of generality, $w_{1i}' < 0$. We will show that $Q_i < 0$ which will be sufficient to prove $Z_i w_{1i}' > 0$ and hence the statement of Lemma 1.

Dropping subscript i,

$$\mathbb{E}(Q) = \mathbb{E}_{Y,S^2}\Big[2S^2(1 - \gamma) - (S^2)^2(1 - \gamma^2)\hat{\tau}^\pi\Big] = \mathbb{E}_{Y,S^2}[R(Y, S^2)],$$

where $\hat{\tau}^\pi = \mathbb{E}(\tau|y, s^2, m)$ from Equation (2.6) and $\mathbb{E}_{Y,S^2}$ is expectation with respect to the joint marginal distribution of (Y, S^2). Suppose, if possible, $Q_i \geq 0$ for all $(\tau_i, y_i, s_i^2) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$. We will show that $\mathbb{E}(Q) < 0$ which will present a contradiction to $Q \geq 0$ for all $(\tau, y, s^2) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$. Fix a $y \in \mathbb{R}$ and consider the following cases.

Case 1 – Suppose $0 < \gamma(y, s^2, m) \leq 1/(2c)$ where $c \geq 1$ is a constant. Then, we have $s^2\hat{\tau}^\pi \geq 2c$ and consequently $s^2(1 + \gamma)\hat{\tau}^\pi > 2$. So $R(y, s^2) < 0$. Now, from Assumption 1(c), γ is a continuous and non-increasing function of s^2. Therefore, there exists $c_1(y) \in \mathbb{R}^+$, depending on y, such that $s^2 > c_1(y)$ whenever $0 < \gamma(y, s^2, m) \leq 1/(2c)$. Thus, $R(y, s^2) < 0$ for $s^2 > c_1(y)$.

Case 2 – Next, suppose $1/2 < \gamma(y, s^2, m) \leq (m - 1)/(2m - 6)$. Then $(2m - 6)/(m - 1) \leq s^2\hat{\tau}^\pi < 2$ and $s^2(1 + \gamma)\hat{\tau}^\pi \geq 3(m - 3)/(m - 1)$. Since $m > 7$, $3(m - 3)/(m - 1) > 2$ and

7

so $s^2(1+\gamma)\hat{\tau}^\pi > 2$. Thus $R(y, s^2) < 0$ and using Assumption 1(c), $R(y, s^2) < 0$ for $c_2(y) < s^2 \le c_1(y)$, where $c_2(y)$ is such that $s^2 > c_2(y)$ whenever $\gamma(y, s^2, m) \le (m-1)/(2m-6)$.

The remaining four cases proceed in a similar manner as follows:

Case 3 – Suppose $(m-1)/(2m-6) < \gamma(y, s^2, m) \le (2m-6)/(3m-11)$. Then $(3m-11)/(2m-6) \le s^2\hat{\tau}^\pi < (2m-6)/(m-1)$. Since $(2m-6)/(3m-11) < 1$ from $m > 7$, we have $s^2(1+\gamma)\hat{\tau}^\pi > 2$ and so $R(y, s^2) < 0$ for $c_3(y) < s^2 \le c_2(y)$. Similarly, we can show that $R(y, s^2) < 0$ for $c_4(y) < s^2 \le c_3(y)$ where $s^2 \in (c_4(y), c_3(y)]$ whenever $(2m-6)/(3m-11) < \gamma(y, s^2, m) \le 1$.

Case 4 – Now suppose, $1 < \gamma(y, s^2, m) \le (2m-5)/(2m-6)$. Then $(2m-6)/(2m-5) \le s^2\hat{\tau}^\pi < 1$ and $s^2(1+\gamma)\hat{\tau}^\pi \in [(4m-12)/(2m-5), (4m-11)/(2m-6))$. Note that here $\gamma > 1$ as opposed to $\gamma \le 1$ in the earlier cases. Therefore, if $s^2(1+\gamma)\hat{\tau}^\pi < 2$ then $R(y, s^2) < 0$ for $c_5(y) < s^2 \le c_4(y)$. The upper limit of the interval for $s^2(1+\gamma)\hat{\tau}^\pi$ is $(4m-11)/(2m-6)$ which, for $m > 7$, is approximately 2 and converges to 2 for a moderately large m.

Case 5 – Similarly, if γ is in the intervals $((2m-5)/(2m-6), (m-2)/(m-3)]$, $((m-2)/(m-3), (m-1)/(m-3)]$ and $((m-1)/(m-3), (m+1)/(m-3)]$ then we have $\gamma \ge 1$ and $s^2(1+\gamma)\hat{\tau}^\pi < 2$. Thus, on each of the corresponding intervals for s^2, $R(y, s^2) < 0$.

Case 6 – Denote $r_1 = 1$ and $r_t = 2r_{t-1} + 3$ for $t = 2, 3, \ldots$. Suppose $(m + r_{t-1})/(m-3) < \gamma(y, s^2, m) \le (m + r_t)/(m-3)$. Then for each of these intervals indexed by $t \ge 2$, we have $\gamma \ge 1$, $s^2(1+\gamma)\hat{\tau}^\pi < 2$ and so $R(y, s^2) < 0$ on the corresponding intervals for s^2.

So from these six cases, $R(y, s^2) < 0$ for all $s^2 > 0$ and consequently $\mathbb{E}_{Y,S^2}[R(Y, S^2)] < 0$. Therefore, $\mathbb{E}(Q) < 0$ which contradicts that $Q \ge 0$ for all $(\tau, y, s^2) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$. Now suppose that for some $\Omega \subset \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$, $Q \ge 0$ whenever $(\tau, y, s^2) \in \Omega$. However, Assumption 1(c) and the aforementioned six cases imply that $\mathbb{E}(Q|\Omega) < 0$. Thus, $Q < 0$ for all $(\tau, y, s^2) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$. So, we have $Zw_1' > 0$ and this completes the proof of Lemma 1.

A.3.2 Proof of Lemma 2

Dropping subscript i, denote,

$$T := -\frac{1}{m^2}\mathbb{E}\Big[(S^2\gamma(Y,S^2,m))^2\nu(Y,S^2,m)\Big] - \frac{2}{m^2}\mathbb{E}\Big[\frac{S^2}{\tau}w_1(Y,S^2,m)\gamma^{'}(Y,S^2,m)\Big].$$

We will show that $T \geq 0$.

Let $\gamma := \gamma(y,s^2,m), \nu := \nu(y,s^2,m), w_1 := w_1(y,s^2,m)$ and $\gamma^{'} := \gamma^{'}(y,s^2,m)$. First note that using standard integration by parts, we have

$$-\frac{1}{m^2}\mathbb{E}\Big[(S^2\gamma)^2\nu\Big] = \frac{2}{m^2}\mathbb{E}\Big[(S^2)^2\gamma w_1\gamma^{'}\Big].$$

So, we can write

$$T = \frac{2}{m^2}\mathbb{E}\Big\{S^2w_1\gamma^{'}\Big[S^2\gamma - \frac{1}{\tau}\Big]\Big\}.$$

Furthermore, from Definition 2,

$$\gamma^{'} = \frac{2}{m-1}\gamma^2s^2w_2^{'},$$

where $w_2^{'} := \frac{\partial}{\partial y}w_2(y,s^2,m)$. So, we have

$$T = \frac{4}{m^2(m-1)}\mathbb{E}\Big\{(S^2\gamma)^2w_1w_2^{'}\Big[S^2\gamma - \frac{1}{\tau}\Big]\Big\}.$$

From Assumption 1(b), $s^2\gamma$ is a continuous and non-decreasing function of s^2. So, there exists a c_τ, depending on τ, such that $s^2\gamma \leq 1/\tau$ whenever $s^2 \leq c_\tau$ and $s^2\gamma > 1/\tau$ whenever $s^2 > c_\tau$. Moreover, since $\mathbb{E}_{Y,S^2|\mu,\tau}(S^2\gamma) \leq 1/\tau$ from Assumption 1(a), we have

$$\Big[1 - \mathbb{P}\Big(S^2 > c_\tau \mid \tau\Big)\Big]\mathbb{E}_{Y,S^2|\mu,\tau}\Big[S^2\gamma - \frac{1}{\tau}\Big|S^2 \leq c_\tau\Big]$$
$$\leq -\mathbb{P}\Big(S^2 > c_\tau \mid \tau\Big)\mathbb{E}_{Y,S^2|\mu,\tau}\Big[S^2\gamma - \frac{1}{\tau}\Big|S^2 > c_\tau\Big] \leq 0. \tag{A.12}$$

9

Let $p \coloneqq \mathbb{P}(S^2 > c_\tau \mid \tau)$ and denote,

$$
\begin{aligned}
T_1 &\coloneqq (1-p)\mathbb{E}_{Y,S^2|\mu,\tau}\Big\{\Big[S^2\gamma - \frac{1}{\tau}\Big](S^2\gamma)^2 w_1 w_2' \Big| S^2 \le c_\tau\Big\}, \text{ and}\\
T_2 &\coloneqq p\mathbb{E}_{Y,S^2|\mu,\tau}\Big\{\Big[S^2\gamma - \frac{1}{\tau}\Big](S^2\gamma)^2 w_1 w_2' \Big| S^2 > c_\tau\Big\}.
\end{aligned}
$$

Now, from Assumption 2(a), $w_1 w_2' \le 0$ and so $T_1 \ge 0$, $T_2 \le 0$. Moreover, from Assumption 2(b) $w_1 w_2'$ is continuous and non-decreasing in s^2. Therefore, from Equation (A.12), $T_1+T_2 \ge 0$. This completes the proof of Lemma 2.

A.4   Proof of Theorem 1

Consider the sample criterion $\hat{\mathbb{M}}_{\lambda,n}(\tilde{\mathcal{W}})$ and population criterion $\mathbb{M}(\tilde{\mathcal{W}})$. We have

$$
\begin{aligned}
\Big|\hat{\mathbb{M}}_{\lambda,n}(\tilde{\mathcal{W}}) - \mathbb{M}(\tilde{\mathcal{W}})\Big| &\le \Big|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i),\tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i,\boldsymbol{X}^j)\mathbb{I}(i \ne j) - \mathbb{M}(\tilde{\mathcal{W}})\Big| + \\
&\quad \Big|\frac{1}{n^2}\sum_{i=1}^{n}\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i),\tilde{\boldsymbol{w}}(\boldsymbol{X}^i)](\boldsymbol{X}^i,\boldsymbol{X}^i)\Big| + \frac{\lambda(n)}{n^2}\|\tilde{\mathcal{W}}\|_F^2 \qquad \text{(A.13)}\\
&\coloneqq I_1 + I_2 + I_3.
\end{aligned}
$$

Define $\bar{\mathbb{M}}_n(\tilde{\mathcal{W}}) = [n(n-1)]^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i),\tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i,\boldsymbol{X}^j)\mathbb{I}(i \ne j)$. Then

$$
I_1 \le |\bar{\mathbb{M}}_n(\tilde{\mathcal{W}}) - \mathbb{M}(\tilde{\mathcal{W}})| + n^{-1}|\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})|.
$$

By assumption 3, $n^{-1}|\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})|$ is $O_p(n^{-1})$. Now, note that $\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})$ is an unbiased estimator of $\mathbb{M}(\tilde{\mathcal{W}})$ and is a U-statistic with a symmetric kernel function $\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i),\tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i,\boldsymbol{X}^j)$. From assumption 3, $\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i),\tilde{\boldsymbol{w}}(\boldsymbol{X}^j)](\boldsymbol{X}^i,\boldsymbol{X}^j)$ has finite second moments. Moreover, from theorem 4.1 of Liu et al. (2016), $\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})$ is a non-degenerate U-statistic whenever $f \ne \tilde{f}$. Thus, from the CLT for U-statistics (Serfling (2009) section 5.5), $|\bar{\mathbb{M}}_n(\tilde{\mathcal{W}}) - \mathbb{M}(\tilde{\mathcal{W}})|$ is $O_p(n^{-1/2})$.

For the terms I_2 and I_3 in equation (A.13), we have

$$I_2 + I_3 = \{1 + \lambda(n)\}\Big|\frac{1}{n^2}\sum_{i=1}^{n}\kappa[\tilde{\boldsymbol{w}}(\boldsymbol{X}^i), \tilde{\boldsymbol{w}}(\boldsymbol{X}^i)](\boldsymbol{X}^i, \boldsymbol{X}^i)\Big|.$$

From Assumption 3, $I_2 + I_3$ is $O_p(\{1 + \lambda(n)\}/n)$. Theorem 1 is proved by combining these results and using the condition $\lambda(n)n^{-1/2} \to 0$.

A.5 Proof of Theorem 2

Denote $\lambda(n)$ by λ and keep its dependence on n implicit for notational ease. According to Proposition 3.3 of Liu et al. (2016), Assumption 4 implies that $\mathbb{M}(\tilde{\mathcal{W}})$ is a valid discrepancy measure in the sense that $\mathbb{M}(\tilde{\mathcal{W}}) > 0$ if and only if $f \neq \tilde{f}$. It follows that given a fixed $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\mathbb{P}\Big\{\frac{c_n}{n}\big\|\hat{\mathcal{W}}_n(\lambda) - \mathcal{W}_0\big\|_F^2 \geq \epsilon_0\Big\} \leq \mathbb{P}\big[c_n\big\{\mathbb{M}(\hat{\mathcal{W}}_n(\lambda)) - \mathbb{M}(\mathcal{W}_0)\big\} \geq \delta\big].$$

But the right hand side in the display above is upper bounded by the sum of three terms: $\mathbb{P}[c_n\{\mathbb{M}(\hat{\mathcal{W}}_n(\lambda)) - \hat{\mathbb{M}}_{\lambda,n}(\hat{\mathcal{W}}_n(\lambda))\} \geq \delta/3]$, $\mathbb{P}[c_n\{\hat{\mathbb{M}}_{\lambda,n}(\hat{\mathcal{W}}_n(\lambda)) - \hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0)\} \geq \delta/3]$ and $\mathbb{P}[c_n\{\hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0) - \mathbb{M}(\mathcal{W}_0)\} \geq \delta/3]$. From Theorem 1 the first term goes to zero as $n \to \infty$ while the second term is zero since $\hat{\mathbb{M}}_{\lambda,n}(\hat{\mathcal{W}}_n(\lambda)) \leq \hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0)$.

Consider the third term $\mathbb{P}[c_n\{\hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0) - \mathbb{M}(\mathcal{W}_0)\} \geq \delta/3]$. Here $\mathbb{M}(\mathcal{W}_0) = 0$ and we can write

$$\begin{aligned}\big|\hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0)\big| &\leq \big|\hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0) - \tilde{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0)\big| + \big|\tilde{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0)\big| \\ &\coloneqq T_1 + T_2,\end{aligned}$$

where $\tilde{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0) = \bar{\mathbb{M}}_n(\mathcal{W}_0) + \lambda n^{-2}\|\mathcal{W}_0\|_F^2$ and

$$\bar{\mathbb{M}}_n(\mathcal{W}_0) = [n(n-1)]^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa[\boldsymbol{w}(\boldsymbol{X}^i), \boldsymbol{w}(\boldsymbol{X}^j)](\boldsymbol{X}^i, \boldsymbol{X}^j)\mathbb{I}(i \neq j).$$

11

So, we have $T_1 \leq \frac{1}{n}\left|\bar{\mathbb{M}}_n(\mathcal{W}_0)\right| + \frac{1}{n^2}\left|\sum_{i=1}^n \kappa[\boldsymbol{w}(\boldsymbol{X}^i), \boldsymbol{w}(\boldsymbol{X}^i)](\boldsymbol{X}^i, \boldsymbol{X}^i)\right|$, and $T_2 \leq \left|\bar{\mathbb{M}}_n(\mathcal{W}_0)\right| + \lambda n^{-2}\|\mathcal{W}_0\|_F^2$. Therefore,

$$\left|\hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0)\right| \leq T_1 + T_2 \leq (1+n^{-1})\left|\bar{\mathbb{M}}_n(\mathcal{W}_0)\right| + \frac{1+\lambda}{n^2}\left|\sum_{i=1}^n \kappa[\boldsymbol{w}(\boldsymbol{X}^i), \boldsymbol{w}(\boldsymbol{X}^i)](\boldsymbol{X}^i, \boldsymbol{X}^i)\right| \quad (A.14)$$

From Theorem 4.1 of Liu et al. (2016), $\bar{\mathbb{M}}_n(\tilde{\mathcal{W}})$ is a degenerate U-statistic when $\tilde{\mathcal{W}} = \mathcal{W}_0$. From Assumption 3 and the CLT for U-statistics [cf. Section 5.5.2 of Serfling (2009)], $\bar{\mathbb{M}}_n(\mathcal{W}_0)$ is $O_p(n^{-1})$. Also, from Assumption 3, the second term in equation (A.14) above is $O_p((1 + \lambda)/n)$. Finally, using these results in equation (A.14) along with the condition that $\lambda n^{-1/2} \to 0$ as $n \to \infty$, we conclude that the third term

$$\mathbb{P}[c_n\{\hat{\mathbb{M}}_{\lambda,n}(\mathcal{W}_0) - \mathbb{M}(\mathcal{W}_0)\} \geq \delta/3] \to 0$$

as $n \to \infty$. The desired result thus follows.

A.6 Proof of Theorem 3 and Corollary 2

We first state two lemmata that are needed for proving Theorem 3. Denote $c_0, c_1, \ldots$ some generic positive constants which may vary in different statements.

Lemma 3 *If Assumption 5 holds, then with probability tending to 1, $|\mu| \leq C_1 \log n$ and $C_2/\log n \leq \tau \leq C_3 \log n$ for some positive constants C_1, C_2 and C_3.*

Lemma 4 *Consider Model (2.1). Suppose Assumption 5 holds. Then with probability tending to 1,*
$$|w_{1,i}| \leq c_0(\log n)^2 \ and \ \mathbb{E}(\tau_i|y_i, s_i^2) \geq \frac{c_1}{\log n}.$$

Lemmata 3 and 4 are proved in Sections A.6.1 and A.6.2, respectively. We now prove Theorem 3.

To establish the first part of Theorem 3, note that

$$\frac{1}{n}\|\boldsymbol{\delta}^{\mathsf{ds}}(\lambda) - \boldsymbol{\delta}^*\|_2^2 \quad = \quad \frac{1}{nm^2}\sum_{i=1}^{n}\left|\frac{w_{1,i}}{\hat{\tau}_i^\pi} - \frac{\hat{w}_{1,\lambda}^i}{\tau_i^{\mathsf{ds}}(\lambda)}\right|^2$$

$$\leq \quad \frac{2}{nm^2}\sum_{i=1}^{n}\frac{1}{[\tau_i^{\mathsf{ds}}(\lambda)]^2}\left|w_{1,i} - \hat{w}_{1,\lambda}^i\right|^2 + \frac{2}{nm^2}\sum_{i=1}^{n}\left|w_{1,i}\right|^2\left|\frac{1}{\tau_i^{\mathsf{ds}}(\lambda)} - \frac{1}{\hat{\tau}_i^\pi}\right|^2$$

$$:= \quad T_1 + T_2.$$

Consider the first term T_1. From the discussion in Section 3.2, there is a positive constant c_0 such that $\tau_i^{\mathsf{ds}}(\lambda) > c_0 > 0$ for all $i = 1, \ldots, n$. It follows that for some constant $c_1 > 0$ depending on the fixed m,

$$T_1 \quad \leq \quad \frac{c_1}{n}\left\|\boldsymbol{w}_1 - \hat{\boldsymbol{w}}_{1,\lambda}\right\|_2^2 \leq \frac{c_1}{n}\left\|\mathcal{W}_0 - \hat{\mathcal{W}}_n(\lambda)\right\|_F^2, \tag{A.15}$$

where $\hat{\boldsymbol{w}}_{1,\lambda} = (\hat{w}_{1,\lambda}^1, \ldots, \hat{w}_{1,\lambda}^n)$ and $\boldsymbol{w}_1 = (w_{1,1}, \ldots, w_{1,n})$. From Theorem 2 the last term on the right hand side of the inequality in equation (A.15) is $O_p(n^{-1/2})$.

Next consider the second term T_2. We have

$$T_2 \leq \frac{c_2}{n}\sum_{i=1}^{n}\left|\frac{w_{1,i}}{\hat{\tau}_i^\pi}\right|^2|w_{2,i} - \hat{w}_{2,\lambda}^i|^2 \tag{A.16}$$

We will use Lemma 4 to bound the terms $|w_{1,i}/\hat{\tau}_i^\pi|$ in equation (A.16). First note that Model (2.1), Assumption 5 and Lemma 3 imply that with probability tending to 1, $|Y_i| \leq c_0 \log n$ and

$$(m-1)n^{-1} \leq (m-1)S_i^2\tau_i \leq (m-1) + 2\sqrt{(m-1)\log n} + 2\log n$$

[cf. Lemma 1 of Laurent and Massart (2000)]. So conditional on these events, we have, from Lemma 4, $|w_{1,i}/\hat{\tau}_i^\pi| \leq c_3 \log^3 n$. Thus,

$$T_2 \quad \leq \quad \frac{c_4 \log^6 n}{n}\sum_{i=1}^{n}\left|w_{2,i} - \hat{w}_{2,\lambda}^i\right|^2,$$

which is $O_p(\log^6 n/\sqrt{n})$ from Theorem 2. Thus $n^{-1}\|\boldsymbol{\delta}^{\mathsf{ds}}(\lambda) - \boldsymbol{\delta}^*\|_2^2$ is $O_p(\log^6 n/\sqrt{n})$.

13

Now we will prove the second part of Theorem 3. Observe that $|l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^*) - l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^{\mathsf{ds}}(\lambda))|$ equals

$$\frac{1}{n}\Big|\big\|\boldsymbol{\mu} - \boldsymbol{\delta}^*\big\|_2 - \big\|\boldsymbol{\mu} - \boldsymbol{\delta}^{\mathsf{ds}}(\lambda)\big\|_2\Big| \ \Big|\big\|\boldsymbol{\mu} - \boldsymbol{\delta}^*\big\|_2 + \big\|\boldsymbol{\mu} - \boldsymbol{\delta}^{\mathsf{ds}}(\lambda)\big\|_2\Big|$$

and Triangle inequality implies

$$\frac{1}{\sqrt{n}}\Big|\big\|\boldsymbol{\mu} - \boldsymbol{\delta}^*\big\|_2 - \big\|\boldsymbol{\mu} - \boldsymbol{\delta}^{\mathsf{ds}}(\lambda)\big\|_2\Big| \ \leq \ \frac{1}{\sqrt{n}}\big\|\boldsymbol{\delta}^{\mathsf{ds}}(\lambda) - \boldsymbol{\delta}^*\big\|_2. \tag{A.17}$$

The quantity on the right hand side of the inequality in equation (A.17) is $O_p(\log^3 n/n^{1/4})$ from the first part of theorem 3. Thus, it follows from equation (A.17) that

$$\sqrt{l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^{\mathsf{ds}}(\lambda))} \leq \sqrt{l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^*)} + O_p(\log^3 n/n^{1/4}), \text{ and}$$

$$\big|l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^*) - l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^{\mathsf{ds}}(\lambda))\big| \leq 4\sqrt{l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^*)}\ \big|\sqrt{l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^*)} - \sqrt{l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^{\mathsf{ds}}(\lambda))}\big|\{1 + o_p(1)\}. \tag{A.18}$$

Now Assumption 5, together with Lemmata 3 and 4 imply $l_n(\boldsymbol{\mu}, \boldsymbol{\delta}^*)$ is $O_p(\log^6 n)$. Thus, from equation (A.17) and the first part of Theorem 3, we have the desired result.

Corollary 2 is a consequence of Theorem 2 of Diaconis and Ylvisaker (1979). When applied to the hierarchical model of equation (2.1) with conjugate priors it establishes that the posterior expectation of μ_i is a linear combination of the prior mean and y_i. The weights in this linear combination are proportional to m and the prior sample size m_0. For instance, if we consider the following normal conjugate model,

$$Y_{ij} \mid \mu_i, \tau_i \overset{i.i.d}{\sim} N(\mu_i, 1/\tau_i), \quad \mu_i \mid \tau_i \overset{ind}{\sim} N(\mu_0, 1/\tau_i), \quad \tau_i \overset{i.i.d}{\sim} \Gamma(\alpha, \beta), \tag{A.19}$$

then under model (A.19) standard calculations give $\delta_i^\pi = (\mu_0 + my_i)/(m+1)$ where μ_0 is the prior mean and $m_0 = 1$ is the prior sample size. Moreover, under model (A.19) we also have

$$\log f(y_i, s_i^2) = c_0 + \frac{m-3}{2}\log s_i^2 - (\alpha + m/2)\log\Big\{\beta^{-1} + 0.5(m-1)s_i^2 + 0.5\frac{m}{m+1}(y_i - \mu_0)^2\Big\},$$

where c_0 is a constant independent of (y_i, s_i^2). From the above display,

$$
\begin{aligned}
w_1(y_i, s_i^2) &= -\frac{\alpha + m/2}{\beta^{-1} + 0.5(m-1)s_i^2 + 0.5m/(m+1)(y_i - \mu_0)^2} \\
w_2(y_i, s_i^2) &= \frac{(m-3)}{2s_i^2} - \frac{0.5(\alpha + m/2)(m-1)}{\beta^{-1} + 0.5(m-1)s_i^2 + 0.5m/(m+1)(y_i - \mu_0)^2}.
\end{aligned}
$$

Substituting these expressions for $w_1(y_i, s_i^2), w_2(y_i, s_i^2)$ in equation (4.17) give $\delta_i^* = \delta_i^\pi$. This suffices to prove the statement of Corollary 2.

A.6.1 Proof of Lemma 3

The proof of Lemma 3 follows directly from Assumption 5 and Markov's inequality. For example, fix a $\nu > 0$ and note that, for $r = \epsilon_2^{-\nu} > 1$,

$$
\mathbb{P}\left(\tau \leq \frac{\epsilon_2^{1+\nu}}{\log n}\right) \leq \frac{\mathbb{E}_H\left\{\exp(\epsilon_2/\tau)\right\}}{n^r}.
$$

A.6.2 Proof of Lemma 4

Recall that $f(y_i, s_i^2) = \int_{\mathbb{R}^+} \int_{\mathbb{R}} f_1(y_i|\mu, \tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau$, where $g(\cdot|\tau)$ and $h(\cdot)$ are, respectively, the density functions associated with the distribution functions $G_\mu(\cdot|\tau)$ and $H_\tau(\cdot)$ in equation (2.1), f_1 is the density of a Gaussian random variable with mean μ and variance $1/(m\tau)$ and f_2 is the density of S^2 where $(m-1)S^2\tau \sim \mathcal{X}_{m-1}^2$. We will denote the partial derivative of $f(y, s^2)$ with respect to y by $f'_{(1)}(y, s^2)$. From Model (2.1), $|f'_{(1)}(y_i, s_i^2)| \leq T_1 + T_2$, where

$$
\begin{aligned}
T_1 &= m|Y_i| \int_{\mathbb{R}^+} \int_{\mathbb{R}} \tau f_1(y_i|\mu, \tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau \qquad \text{(A.20)} \\
T_2 &= m \int_{\mathbb{R}^+} \int_{\mathbb{R}} |\mu| \tau f_1(y_i|\mu, \tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau.
\end{aligned}
$$

Consider the term T_1 in equation (A.20) above. Fix $\nu_3 > 0$ such that $\epsilon_3^{-\nu_3} > 4m$ in Lemma 3. With $C_3 = \epsilon_3^{-1-\nu_3}$ we have,

$$
\begin{aligned}
T_1 &\leq m|Y_i|f(y_i, s_i^2)C_3 \log n + m|Y_i| \int_{\mathbb{R}^+} \int_{\mathbb{R}} \tau \mathbb{I}(\tau \geq C_3 \log n) f_1(y_i|\mu,\tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau \\
&\leq m|Y_i|f(y_i, s_i^2)C_3 \log n + mc_0|Y_i|\mathbb{E}_H\left\{\tau^{5/2}\mathbb{I}(\tau \geq C_3 \log n)\right\} \qquad\qquad\qquad (\text{A.21}) \\
&\leq m|Y_i|f(y_i, s_i^2)C_3 \log n + mc_1|Y_i|\left\{\mathbb{P}\left(\tau \geq C_3 \log n\right)\right\}^{1/2} \qquad\qquad\qquad\quad (\text{A.22}) \\
&\leq m|Y_i|f(y_i, s_i^2)C_3 \log n + mc_2|Y_i|n^{-2m} \qquad\qquad\qquad\qquad\qquad\qquad\quad (\text{A.23})
\end{aligned}
$$

In equation (A.21) we have used the fact that $f_1(y_i|\mu,\tau) \leq \sqrt{\tau/(2\pi)}$ and $f_2(s_i^2|\tau) \leq c_2\tau$ while for equations (A.22), (A.23) we use the Cauchy Schwartz inequality, Assumption 3 and Lemma 3.

Now, consider the term T_2 in equation (A.20). With $C_3 = \epsilon_3^{-1-\nu_3}$ and $C_1 = \epsilon_1^{-1-\nu_1}$, where $\nu_1 > 0$ is such that $\epsilon_1^{-\nu_1} > 4m$ in Lemma 3, T_2 term can be expressed as the sum of the following four integrals:

$$
\begin{aligned}
I_1 &= m \int_{\mathbb{R}^+} \int_{\mathbb{R}} |\mu|\mathbb{I}(|\mu| \leq C_1 \log n)\tau \mathbb{I}(\tau \leq C_3 \log n) f_1(y_i|\mu,\tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau \quad (\text{A.24}) \\
I_2 &= m \int_{\mathbb{R}^+} \int_{\mathbb{R}} |\mu|\mathbb{I}(|\mu| \geq C_1 \log n)\tau \mathbb{I}(\tau \leq C_3 \log n) f_1(y_i|\mu,\tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau \quad (\text{A.25}) \\
I_3 &= m \int_{\mathbb{R}^+} \int_{\mathbb{R}} |\mu|\mathbb{I}(|\mu| \leq C_1 \log n)\tau \mathbb{I}(\tau \geq C_3 \log n) f_1(y_i|\mu,\tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau \\
I_4 &= m \int_{\mathbb{R}^+} \int_{\mathbb{R}} |\mu|\mathbb{I}(|\mu| \geq C_1 \log n)\tau \mathbb{I}(\tau \geq C_3 \log n) f_1(y_i|\mu,\tau) f_2(s_i^2|\tau) g(\mu|\tau) h(\tau) \mathrm{d}\mu \mathrm{d}\tau
\end{aligned}
$$

We will bound I_1 and I_2 and the other two integrals can be bounded using similar arguments. For I_1 in equation (A.24), note that $I_1 \leq c_0 \log^2 n f(y_i, s_i^2)$ and for I_2 in equation (A.25),

$$
\begin{aligned}
I_2 &\leq c_1(\log^{5/2} n)\left\{\mathbb{P}\left(|\mu| \geq C_1 \log n\right)\right\}^{1/2} \\
&\leq c_2(\log^{5/2} n)n^{-2m} \qquad\qquad\qquad\qquad\qquad (\text{A.26})
\end{aligned}
$$

In equation (A.26) we have used $f_1(y_i|\mu,\tau) \leq \sqrt{\tau/(2\pi)}$, $f_2(s_i^2|\tau) \leq c_2\tau$, the Cauchy Schwartz inequality, Assumption 3 and Lemma 3. Similarly, we can show that $I_3 \leq c_0 \log n/n^{2m}$ and

16

$I_4 \leq c_1/n^{4m}$. Finally, putting the upper bounds for I_i, $i = 1, \cdots, 4$, together, we get,

$$T_2 \leq c_0 \log^2 n f(y_i, s_i^2) + c_1 \frac{\log^{5/2} n}{n^{2m}} \tag{A.27}$$

From equations (A.23) and (A.27) we have

$$|w_1(Y_i, S_i^2)| \leq c_0 |Y_i| \log n + c_1 \log^2 n + c_2 \frac{|Y_i| + \log^{5/2} n}{f(y_i, s_i^2) n^{2m}}.$$

Moreover, noting that Model (2.1) and Lemma 3 imply that $|Y_i| \leq c_3 \log n$ with high probability, we have, conditional on this event,

$$|w_1(Y_i, S_i^2)| \leq c_4 \log^2 n + c_5 \frac{\log n + \log^{5/2} n}{f(y_i, s_i^2) n^{2m}}. \tag{A.28}$$

We will now analyze the behavior of $f(y_i, s_i^2)$ that appears in the display above. Gaussian concentration implies that with high probability $\{(Y_i - \mu_i)^2 m \tau_i\} \leq 2 \log n$ and so, conditional on this event,

$$f_1(y_i | \mu, \tau) \geq c_0 \frac{\sqrt{\tau}}{n}. \tag{A.29}$$

Moreover, using the Chi-square concentration in Lemma 1 of Laurent and Massart (2000), $(m-1)S_i^2 \tau_i \leq (m-1) + 2\sqrt{(m-1)\log n} + 2 \log n$ and $S_i^2 \tau_i \geq n^{-1}$ with high probability. It follows that

$$f_2(s_i^2 | \tau) \geq c_1 \tau \frac{a_n}{n^{(m+1)/2} \log n}, \tag{A.30}$$

conditional on this event, where $a_n = \exp\{-\sqrt{(m-1)\log n}\}$. Using equations (A.29) and (A.30), we have

$$f(y_i, s_i^2) \geq c_2 \frac{a_n}{n^{(m+3)/2} \log n} \int_{\mathbb{R}+} \tau^{3/2} h(\tau) d\tau. \tag{A.31}$$

Now, use Assumption 5 and Lemma 3 on the quantity $\int_{\mathbb{R}+} \tau^{3/2} h(\tau) d\tau$ in equation (A.31) to conclude that

$$\int_{\mathbb{R}+} \tau^{3/2} h(\tau) d\tau \geq \frac{c_3}{\log^{3/2} n} \mathbb{P}\left(\tau \geq C_2 / \log n\right). \tag{A.32}$$

So, with equations (A.32), (A.31) and Lemma 3, we have with high probability,

$$f(y_i, s_i^2) \geq c_4 \frac{a_n}{n^{(m+3)/2} \log^{5/2} n}. \tag{A.33}$$

The first statement of Lemma 4 thus follows from equations (A.33) and (A.28).

We will now prove the second statement of Lemma 4. Fix $\nu_2 > 0$ such that $\epsilon_2^{-\nu_2} > 2m$ in Lemma 3 and let $C_2 = \epsilon_2^{1+\nu_2}$. First note that from Assumption 5, $\mathbb{P}(\tau \leq C_2/\log n) \leq c_0/n^{2m}$. Now, Markov's inequality implies,

$$\mathbb{E}(\tau|y_i, s_i^2) \geq \frac{C_2}{\log n}\Big\{1 - \mathbb{P}(\tau \leq C_2/\log n|y_i, s_i^2)\Big\}.$$

Moreover,

$$\mathbb{P}(\tau \leq C_2/\log n|y_i, s_i^2) = \{f(y_i, s_i^2)\}^{-1} \int_0^{C_2/\log n} h(\tau)\Big\{\int_{\mathbb{R}} g(\mu|\tau)f_1(y_i|\mu,\tau)f_2(s_i^2|\tau)\mathrm{d}\mu\Big\}\mathrm{d}\tau.$$

Now $f_1(y_i|\mu,\tau) \leq \sqrt{\tau/(2\pi)}$ and $f_2(s_i^2|\tau) \leq c_1\tau$ where $c_1 > 0$ is a constant. So for some positive constant c_2,

$$\begin{aligned}
\mathbb{P}(\tau \leq C_2/\log n|y_i, s_i^2) &\leq \frac{c_2}{f(y_i, s_i^2)} \int_0^{C_2/\log n} \tau^{3/2} h(\tau)\Big\{\int_{\mathbb{R}} g(\mu|\tau)\mathrm{d}\mu\Big\}\mathrm{d}\tau \\
&\leq \frac{c_3}{f(y_i, s_i^2)\log^{3/2} n} \int_0^{C_2/\log n} h(\tau)\mathrm{d}\tau = \frac{c_3}{f(y_i, s_i^2)\log^{3/2} n}\mathbb{P}(\tau \leq C_2/\log n).
\end{aligned}$$

Thus, from the above display, Assumption 5 and Lemma 3,

$$\mathbb{E}(\tau|y_i, s_i^2) \geq \frac{C_2}{\log n}\Big\{1 - c_3 \frac{n^{-2m}}{f(y_i, s_i^2)\log^{3/2} n}\Big\}.$$

Finally, equation (A.33) and the above display prove the second statement of Lemma 4.

B   Extensions

This section considers the extension of our methodology to several well known members in the two-parameter exponential family. We will focus on several examples where the nuisance parameter is known. Our proposed estimation framework is motivated by the double shrinkage idea, but the approach nonetheless handles the case with known nuisance parameters. We discuss four examples, in each of which we derive the Bayes estimator of the natural parameter similar to that in Corollary 1. The Bayes estimator in these examples relies on the unknown score function (of the marginal density of the sufficient statistic), which can be efficiently estimated using the ideas in Section 3. The numerical performance of the new estimators considered in this section is investigated in Section C.3.

Example B.1 (Location mixture of Gaussians) *Consider the following hierarchical model*

$$Y_i \mid \mu_i, \tau_i \overset{ind.}{\sim} N(\mu_i, 1/\tau_i), \quad \mu_i \overset{i.i.d}{\sim} G_\mu(\cdot), \quad for \ i = 1, \dots, n, \tag{B.34}$$

where τ_i are known and $G_\mu(\cdot)$ is an unspecified prior. Equation (B.34) represents the heteroscedastic normal means problem with known variances $1/\tau_i$ [see for example Weinstein et al. (2018)]. In this setting, the sufficient statistic for μ_i is Y_i and the Bayes estimator of μ_i is given by

$$\mu_i^\pi := \mathbb{E}(\mu_i | y_i, \tau_i) = y_i + \frac{1}{\tau_i} \frac{\partial}{\partial y_i} \log f(y_i | \tau_i),$$

where $f(\cdot | \tau_i)$ is the pdf of the distribution of Y_i given τ_i marginalizing out μ_i. From Section 3 and with $m_i = 1$, $\boldsymbol{x}_i = (y_i, \tau_i)$, the NEST estimate of μ_i is given by $\delta_i^{\mathsf{nest}}(\lambda) = y_i + \frac{1}{\tau_i} \hat{w}_{1,\lambda}^i$.

Example B.2 (Scale mixture of Gamma distributions) *Consider the following model*

$$Y_{ij} \mid \alpha_i, 1/\beta_i \overset{i.i.d}{\sim} \Gamma(\alpha_i, 1/\beta_i), \quad 1/\beta_i \overset{i.i.d}{\sim} G(\cdot), \tag{B.35}$$

where the shape parameters α_i are known and $G(\cdot)$ is an unspecified prior distribution on scale parameters $1/\beta_i$. Here $T_i = \sum_{j=1}^m Y_{ij}$ is a sufficient statistic and $T_i | \alpha_i, \beta_i \overset{ind.}{\sim} \Gamma(m\alpha_i, 1/\beta_i)$.

The posterior distribution of $1/\beta_i$ belongs to a one-parameter exponential family with density

$$f(1/\beta_i|T_i,\alpha_i) \propto \exp\left\{-T_i\beta_i + (m\alpha_i - 1)\log T_i - \log f(T_i|\alpha_i)\right\}, \qquad \text{(B.36)}$$

where $f(\cdot|\alpha_i)$ is the pdf of the distribution of T_i given α_i (marginalizing out $1/\beta_i$). From Equation (B.36), the Bayes estimator of β_i is given by

$$\beta_i^\pi := \mathbb{E}(\beta_i|T_i,\alpha_i) = \frac{m\alpha_i - 1}{T_i} - \frac{\partial}{\partial T_i}\log f(T_i|\alpha_i).$$

With $\boldsymbol{x}_i = (T_i,\alpha_i)$, the NEST estimate of β_i is given by $\delta_i^{\mathsf{nest}}(\lambda) = \dfrac{m\alpha_i - 1}{T_i} - \hat{w}_{1,\lambda}^i$.

Example B.3 (Shape mixture of Gamma distributions) *We consider the following model:*

$$Y_{ij} \mid \alpha_i, 1/\beta_i \overset{i.i.d}{\sim} \Gamma(\alpha_i, 1/\beta_i), \quad \alpha_i \overset{i.i.d}{\sim} G(\cdot), \qquad \text{(B.37)}$$

where the scale parameters $1/\beta_i$ are known and $G(\cdot)$ is an unspecified prior distribution on the shape parameters α_i. Let $Y_i = \sum_{j=1}^m Y_{ij}$. Then $Y_i|\alpha_i,\beta_i \overset{ind.}{\sim} \Gamma(m\alpha_i, 1/\beta_i)$ and $T_i = \log Y_i$ is a sufficient statistic. Moreover, the posterior distribution of α_i belongs to a one-parameter exponential family with density

$$f(\alpha_i|T_i, 1/\beta_i) \propto \exp\left\{(m\alpha_i)T_i - \beta_i\exp\left(T_i\right) - \log f(T_i|1/\beta_i)\right\}, \qquad \text{(B.38)}$$

where $f(\cdot|1/\beta_i)$ is the density of the distribution of T_i given $1/\beta_i$ marginalizing out α_i. From Equation (B.38), the Bayes estimator of α_i is given by

$$\alpha_i^\pi := \mathbb{E}(\alpha_i|T_i, 1/\beta_i) = \frac{\beta_i\exp\left(T_i\right)}{m} + \frac{1}{m}\frac{\partial}{\partial T_i}\log f(T_i|1/\beta_i).$$

With $\boldsymbol{x}_i = (T_i, 1/\beta_i)$, the NEST estimate of α_i is $\delta_i^{\mathsf{nest}}(\lambda) = \dfrac{\beta_i\exp\left(T_i\right)}{m} + \dfrac{1}{m}\hat{w}_{1,\lambda}^i$.

Example B.4 (Scale mixture of Weibulls) *We consider the following model:*

$$Y_{ij} \mid k_i, \beta_i \overset{i.i.d}{\sim} Weibull(k_i, \beta_i), \quad \beta_i \overset{i.i.d}{\sim} G(\cdot). \tag{B.39}$$

We have $f(y \mid k, \beta) = \beta k y^{k-1} \exp(-\beta y^k)$. *In Equation* (B.39) *the shape parameters* k_i *are known,* $G(\cdot)$ *is an unspecified prior distribution on the scale parameters* β_i, $T_i = \sum_{j=1}^{m} \{Y_{ij}\}^{k_i}$ *is a sufficient statistic, and* $T_i|k_i, 1/\beta_i \overset{ind.}{\sim} \Gamma(m, 1/\beta_i)$. *From Example 2, the Bayes estimator of* β_i *is*

$$\beta_i^{\pi} := \mathbb{E}(\beta_i|T_i, k_i) = \frac{m-1}{T_i} + \frac{\partial}{\partial T_i} \log f(T_i|k_i).$$

With $\boldsymbol{x}_i = (T_i, k_i)$, *the NEST estimate of* β_i *is given by* $\delta_i^{\mathsf{nest}}(\lambda) = \frac{m-1}{T_i} - \hat{w}_{1,\lambda}^i$.

The preceding examples present a setting with known nuisance parameter. When both parameters are unknown, extensions of our estimation framework to an arbitrary member of the two-parameter exponential family is difficult. The main reason is that in the Gaussian case the sufficient statistics are independent and their marginal distributions are known. However, for other distributions such as the Gamma and Beta, the joint distribution of the two sufficient statistics is generally unknown. This impedes a full generalization of our approach. We anticipate that an iterative scheme that conducts shrinkage estimation on the primary and nuisance coordinates in turn may be developed by combining the ideas in Examples 2 and 3 above. We do not pursue those extensions in this article.

C   Additional Numerical Experiments

C.1   Numerical Experiments with unequal sample sizes m_i

In this section, we present the risk performance of the seven competing approaches of sections 5.1 and 5.2 when the sample sizes m_i are different across the $n = 1000$ units of study. We continue to use the simulation settings of sections 5.1 and 5.2, and only change how $\boldsymbol{m} = (m_1, \ldots, m_n)$ are generated in these settings. Figures 8 to 10 present the relative
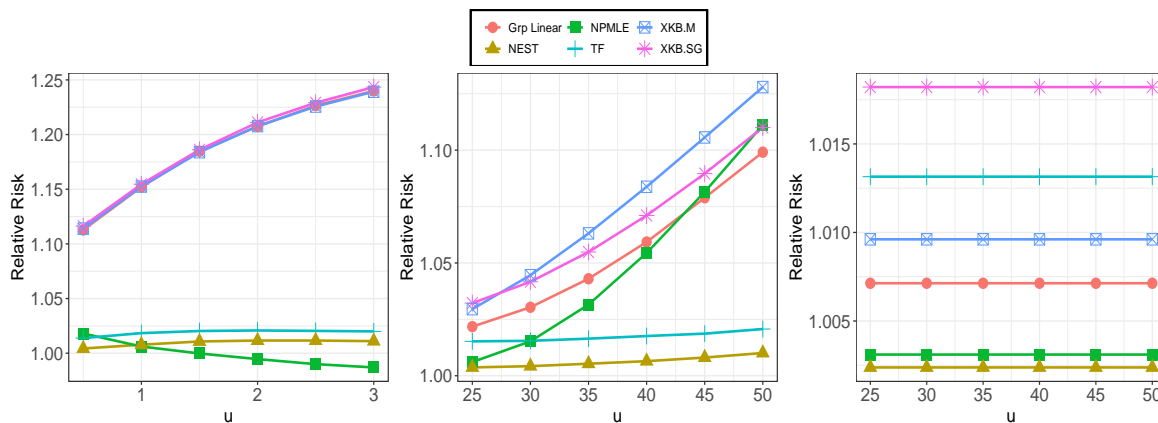
Figure 8: Left to Right: Simulation settings 1, 2 and 3. Here $\boldsymbol{m}$ is a fixed vector of size n with elements sampled randomly from $(10, 11, \ldots, 20)$ with replacement.

risks of the competing estimators across the seven simulations settings of sections 5.1 and 5.2. Here $\boldsymbol{m}$ is generated as a fixed vector of size n with elements sampled randomly from $(10, 11, \ldots, 20)$ with replacement. For setting 1, figure 8 left, we note that NPMLE dominates NEST as u increases while both NEST and TF exhibit a better risk performance than the three linear shrinkage estimators. Under setting 2, figure 8 center, NEST dominates all other competing estimators considered here while for setting 3 (figure 8 right), which represents the conjugate case, the NEST and NPMLE have the best risk performance. Settings 5 and 6 (figure 9 center and right) from section 5.1 represent scenarios wherein the data $Y_{ij}|(\mu_i, \sigma_i^2)$ are not normally distributed. Under both these settings, NEST provides an overall better risk performance than the linear shrinkage estimators as the heterogeneity in the data increases with u while NPMLE dominates NEST in setting 6 (figure 9 right). Setting 7 (figure 10) is the ratio estimation scenario considered in section 5.2 and we see that NEST has a relatively better risk performance than the linear shrinkage estimators and Tweedie's formula while NPMLE dominates NEST for relatively smaller values of u.

C.2   Compound Estimation of Normal Means - known variances

We focus on the hierarchical Model of equation (2.1) with known variances and compare the following six approaches for estimating $\boldsymbol{\mu}$: the NEST oracle method, which estimates λ
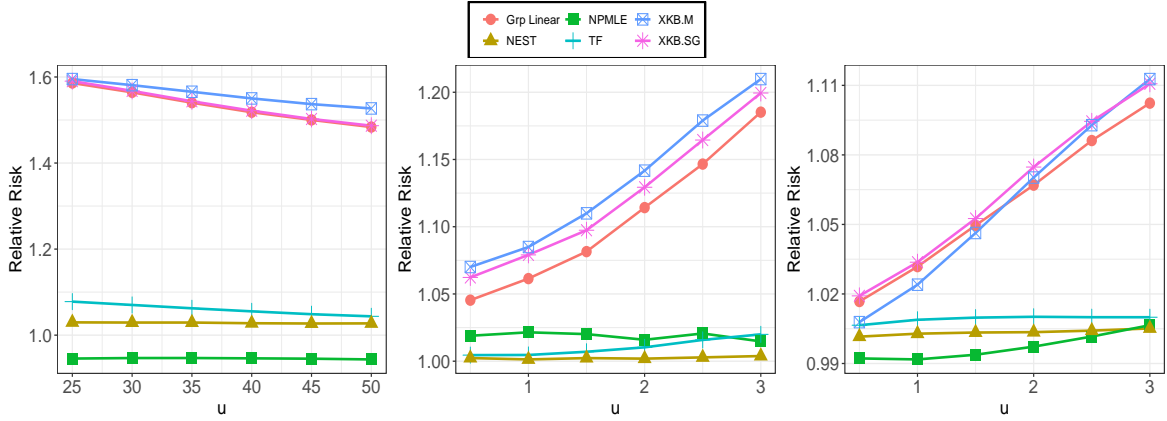
Figure 9: Left to Right: Simulation settings 4, 5 and 6 (from section 5.2). In each setting $\boldsymbol{m}$ is a fixed vector of size n with elements sampled randomly from $(10, 11, \ldots, 20)$ with replacement.
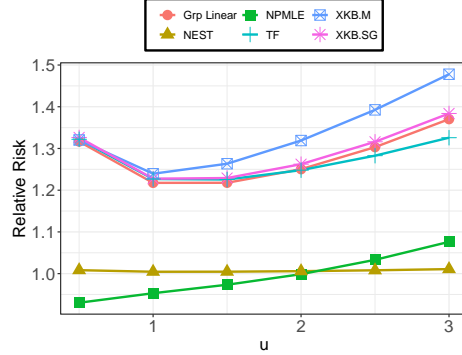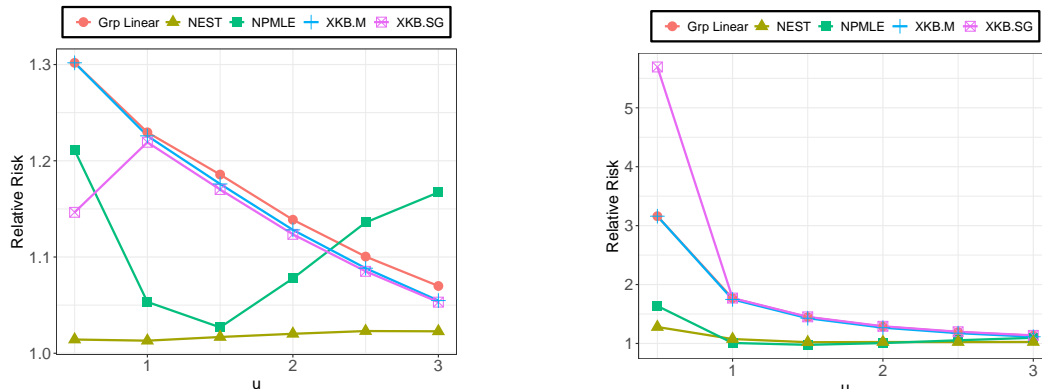

Figure 10: Simulation setting 7 from section 5.2. Here $\boldsymbol{m}$ is a fixed vector of size n with elements sampled randomly from $(10, 11, \ldots, 20)$ with replacement.

by minimizing the true loss (NEST orc), and the proposed NEST method, where λ is chosen using modified cross-validation, the g-modelling approach of Gu and Koenker (2017a,b) (NPMLE), group linear estimator (Grp linear) of Weinstein et al. (2018), the semi-parametric monotonically constrained SURE estimator that shrinks towards the grand mean (XKB.SG) from Xie et al. (2012) and from the same paper, the parametric SURE estimator that shrinks towards a general data driven location (XKB.M).

The aforementioned six approaches are evaluated on four different simulation settings, with the goal of assessing the relative performance of the competing estimators as the heterogeneity in the variances σ_i^2 is varied while keeping n fixed at 1,000. For each setting
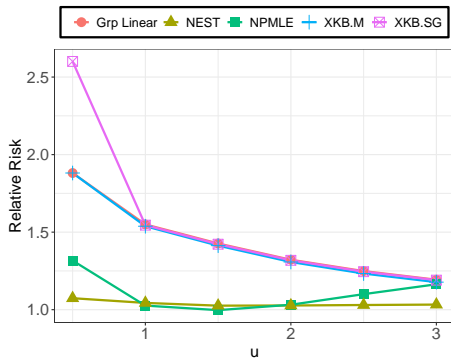
we compute the average squared error risk for each competing estimator of $\boldsymbol{\mu}$ across 50 Monte Carlo repetitions. Figures 11 and 12 plot the relative risk which is the ratio of the average squared error risk for any competing estimator to that of NEST orc so that a ratio bigger than 1 represents a poorer risk performance of the competing estimator relative to the NEST oracle method. The first setting, panel (a) of Figure 11, corresponds to the
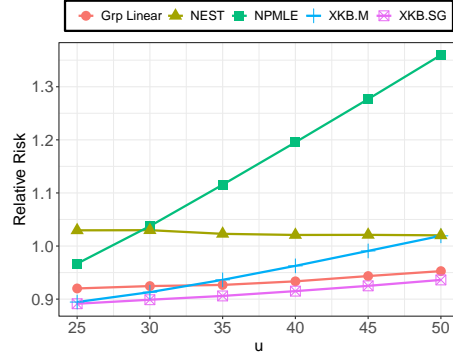

(a) Setting 1: $\mu_i \overset{i.i.d}{\sim} 0.7\ N(0,.1) + 0.3\ N(\pm 1, 3)$ and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.5, u)$.

(b) Setting 2: μ_i can take values 0 or 3 with equal probability and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.5, u)$.

Figure 11: Comparison of relative risks under simulation settings 1 and 2.

scenario where the mean and the variances are independent. Here, for each $i = 1, \ldots, n$, $\mu_i \overset{i.i.d}{\sim} 0.7\ N(0,.1) + 0.3\ N(\pm 1, 3)$ and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.5, u)$ where we let u vary across six levels, $\{0.5, 1, 1.5, 2, 2.5, 3\}$. This is the same setting 1 of Figure 1 in Section 5.1 but with known variances. From panel (a) of Figure 11, we see that as heterogeneity increases with increasing u, the gap between the linear shrinkage methods and NEST decreases. NPMLE, on the other hand, demonstrates a higher relative risk than NEST except at $u = 1.5$. The second setting, panel (b) of Figure 11, is another example of the scenario where the mean and the variances are independent. Here, for each $i = 1, \ldots, n$, μ_i can take the values 0 or 3 with equal probability and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.5, u)$ with $u \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. Under this setting, we expect NPMLE to perform substantially better as the true prior distribution on the means is discrete. Moreover, from panel (b) of Figure 11 we note that both NEST and NPMLE dominate the linear shrinkage methods. The third setting, panel (a) of Figure 12, corresponds to the sparse case. The variances σ_i^2 are drawn from the same uniform distribution as in

(a) Setting 3: $\mu_i \overset{i.i.d}{\sim} 0.7\delta_{(0)} + 0.3N(3,1)$ and $\sigma_i^2 \overset{i.i.d}{\sim} U(0.5, u)$.

(b) Setting 4: $\tau_i \overset{ind.}{\sim} 0.5\Gamma(20, \texttt{rate} = 20) + 0.5\Gamma(20, \texttt{rate} = u)$ and $\mu_i | \tau_i \overset{ind.}{\sim} N(\pm 0.5/\tau_i, 0.5^2)$.

Figure 12: Comparison of relative risks under simulation settings 3 and 4.

Figure 11, but μ_i are only 30% likely to come from $N(3,1)$ and 70% likely from a point mass at 0. We continue to see a similar pattern in the relative risks of the competing estimators wherein both NPMLE and NEST are better than the linear shrinkage estimators considered here. The fourth setting, panel (b) of Figure 12, corresponds to the correlated case of Figure 2 in Section 5.1 with known variances. The precisions $\tau_i = 1/\sigma_i^2$ are generated independently from a mixture distribution that has an even chance of drawing from either $\Gamma(20, rate = 20)$ or $\Gamma(20, rate = u)$ and given τ_i, the means μ_i are independently $N(\pm 0.5/\tau_i, 0.5^2)$. In this setting, the magnitude of the variances increase with u and the means grow with the variances. Panel (b) of Figure 12 reveals that XKB.SG and Grp Linear dominate all other competing estimators although XKB.M is marginally better than Grp Linear for relatively smaller values of u. This is in sharp contrast with the performance of Grp Linear and XKB.SG under the same setting but with unknown variances (Figure 2 in Section 5.1). The distribution of the variances in this setting is bimodal and carry important structural information regarding the means. When these variances are known, Grp Linear is able to exploit the additional information in the variances to conduct a superior estimation of the means. However, when these variances are unknown, Grp Linear, relying on sample variances, has a substantially higher relative risk than NEST as the heterogeneity increases (Figure 2 in Section 5.1).

C.3   Gamma and Weibull mixtures

In this section we provide numerical evidence that demonstrate the performance of the NEST estimator for compound estimation of the scale parameter of Gamma and Weibull mixtures in examples B.2 and B.4 of Section B. To choose λ in these settings, we use two fold cross validation by randomly splitting the observed data $\boldsymbol{Y}_i = (Y_{ij} : 1 \leq j \leq m)$ for unit i into two approximately equal parts $(\boldsymbol{U}_i, \boldsymbol{V}_i)$ such that $\boldsymbol{Y}_i = (\boldsymbol{U}_i, \boldsymbol{V}_i)$.

Scale mixture of Gamma distributions - Consider the setting of example B.2 in Section B where for $j = 1, \ldots, m$ and $i = 1, \ldots, n$,

$$Y_{ij} \mid \alpha_i, 1/\beta_i \overset{i.i.d}{\sim} \Gamma(\alpha_i, 1/\beta_i), \quad 1/\beta_i \overset{i.i.d}{\sim} G(\cdot).$$

Here the shape parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ are known and $G(\cdot)$ is an unspecified prior distribution on the scale parameters $1/\beta_i$. We fix $m = 30$ and consider two scenarios wherein, for scenario 1, we let $\beta_i \overset{i.i.d}{\sim} \Gamma(7, 1)$ and $\boldsymbol{\alpha}$ to be a fixed vector of size n with elements sampled randomly from $(1, 3, 5)$ with replacement. For scenario 2, $\beta_i \overset{i.i.d}{\sim} \mathcal{X}_5^2$ and $\boldsymbol{\alpha}$ is a fixed vector of size n with elements sampled randomly from $(1, 2, 3)$ with replacement. We consider four competing estimators of $\boldsymbol{\beta}$ across the two scenarios as n varies. The competing estimators include the NEST estimator from example 2 with λ obtained from two fold cross validation, the `NPMLE` based estimator of $\boldsymbol{\beta}$ from Koenker and Gu (2017), the maximum likelihood estimator (`MLE`) of β_i which is $m\alpha_i / \sum_{j=1}^{m} Y_{ij}$ and `NEST orc` which represents the NEST estimator of example 2 with oracle λ. The average squared error risk of these estimators is calculated over 100 Monte-Carlo repetitions and figure 13 presents the ratio of the average risks wherein the denominator is the average risk of `NEST orc`. For scenario 1 (figure 13 left), the NEST estimator is almost as good as the NPMLE estimator of $\boldsymbol{\beta}$ for large n. Scenario 2, on the other hand, represents a challenging setting where β_i are relatively smaller in comparison to those in scenario 1. In this setting, the NEST estimator has a lower relative risk than the NPMLE across all n.
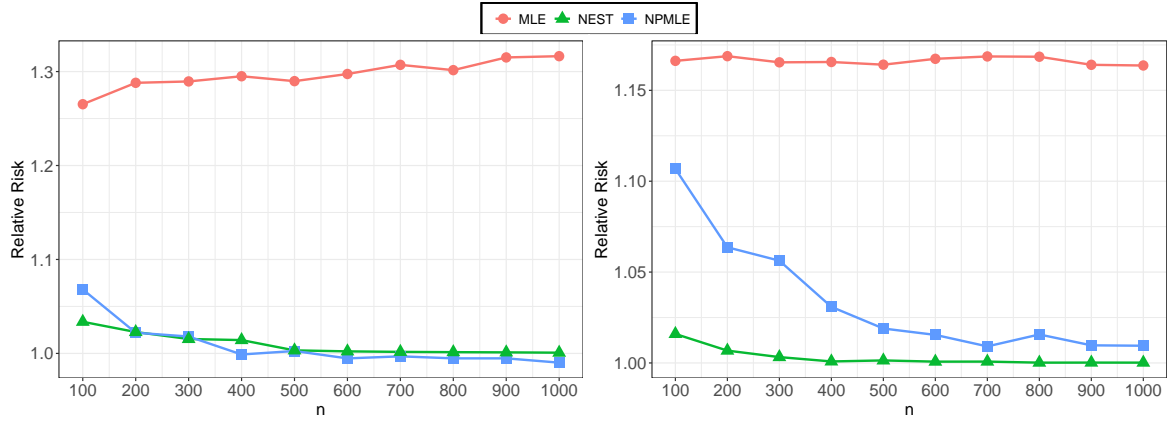
Figure 13: Comparison of relative risk for estimating $\boldsymbol{\beta}$ under a scale mixture of Gamma distributions. Left: Scenario 1 - $\beta_i \overset{i.i.d}{\sim} \Gamma(7,1)$ and $\boldsymbol{\alpha}$ to be a fixed vector of size n with elements sampled randomly from $(1,3,5)$ with replacement. Right: Scenario 2 - $\beta_i \overset{i.i.d}{\sim} \mathcal{X}_5^2$ and $\boldsymbol{\alpha}$ is a fixed vector of size n with elements sampled randomly from $(1,2,3)$ with replacement. Here m is fixed at 30.

Scale mixture of Weibull distributions - Consider the setting of example 4 in Section B where for $j = 1, \ldots, m$ and $i = 1, \ldots, n$,

$$Y_{ij} \mid k_i, \beta_i \overset{i.i.d}{\sim} Weibull(k_i, \beta_i), \quad \beta_i \overset{i.i.d}{\sim} G(\cdot).$$

Here the shape parameters $\boldsymbol{k} = (k_1, \ldots, k_n)$ are known and $G(\cdot)$ is an unspecified prior distribution on the scale parameters β_i. We fix $m = 30$ and consider two scenarios wherein, for scenario 1, we let $\beta_i = |Z_i|$ where $Z_i \overset{i.i.d}{\sim} N(1, 0.1^2)$ and for scenario 2, $\beta_i \overset{i.i.d}{\sim} (1/3)\delta_{(0.75)} + (1/3)\delta_{(1)} + (1/3)\delta_{(1.25)}$. In each case, $\boldsymbol{\alpha}$ is a fixed vector of size n with elements sampled randomly from $(1, 2, 3)$ with replacement. Analogous to figure 13, figure 14 presents the relative risks of the competing estimators of $\boldsymbol{\beta}$ across the two scenarios as n varies. The competing estimators include the NEST estimator from example 4 with λ obtained from two fold cross validation, the NPMLE based estimator of $\boldsymbol{\beta}$, the maximum likelihood estimator (MLE) of β_i which is $m / \sum_{j=1}^{m} Y_{ij}^{k_i}$ and NEST orc which represents the NEST estimator of example 4 with oracle λ. We note that for both the scenarios and particularly for scenario 2, the NEST estimator has a lower relative risk than the NPMLE for large n.
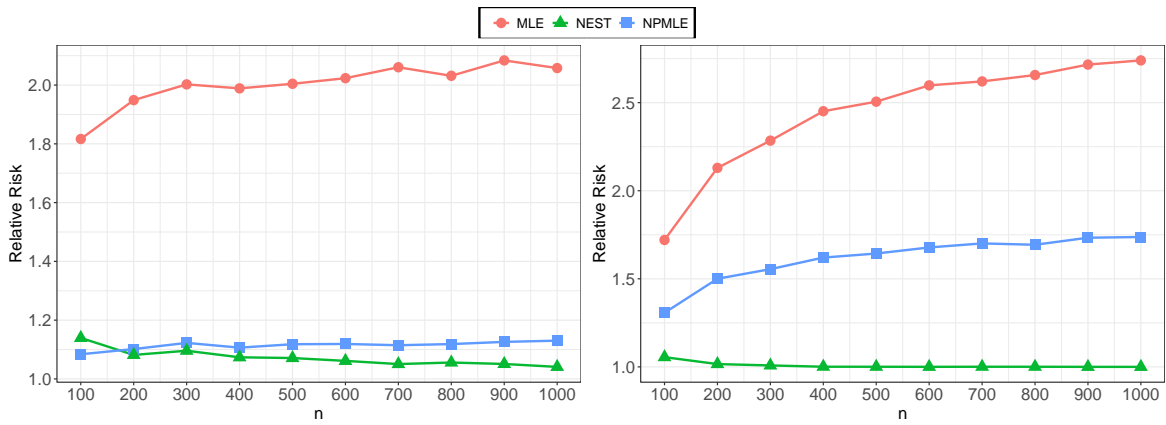
Figure 14: Comparison of relative risk for estimating $\boldsymbol{\beta}$ under a scale mixture of Weibull distributions. Left: Scenario 1 - $\beta_i = |Z_i|$ where $Z_i \overset{i.i.d}{\sim} N(1, 0.1^2)$. Right: Scenario 2 - $\beta_i \overset{i.i.d}{\sim} (1/3)\delta_{(0.75)} + (1/3)\delta_{(1)} + (1/3)\delta_{(1.25)}$. Here $\boldsymbol{\alpha}$ is a fixed vector of size n with elements sampled randomly from $(1, 2, 3)$ with replacement and m is fixed at 30.