Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

NONPARAMETRIC EMPIRICAL BAYES ESTIMATION ON HETEROGENEOUS DATA

By Luella Fu, Gareth James, Wenguang Sun*

University of Southern California

The simultaneous estimation of many parameters η_i, based on a corresponding set of observations x_i, for $i = 1, \ldots, n$, is a key research problem that has received renewed attention in the high-dimensional setting. The classic example involves estimating a vector of normal means μ_i subject to a fixed variance term σ^2. However, many practical situations involve heterogeneous data (x_i, θ_i) where θ_i is a known nuisance parameter e.g. $\theta_i = \sigma_i^2$. 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 Smoothing Tweedie" (NEST) estimator, which efficiently estimates η_i and properly adjusts for heterogeneity via a generalized version of Tweedie's formula. NEST is capable of handling a wider range of settings than previously proposed heterogeneous approaches as it does not make any parametric assumptions on the prior distribution of η_i. The estimation framework is simple but general enough to accommodate any member of the exponential family of distributions; a thorough study of the normal means problem subject to heterogeneous variances is presented to illustrate the proposed framework. Our theoretical results show that NEST is asymptotically optimal, while simulation studies show that it outperforms competing methods, with substantial efficiency gains in many settings. The method is further demonstrated on a data set measuring the performance gap in math scores between socioeconomically advantaged and disadvantaged students in K-12 schools.

1. Introduction. Suppose that we are interested in estimating a population vector of parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ based on a vector of summary statistics $\boldsymbol{X} = (X_1, \cdots, X_n)^T$. The setting where $\eta_i = \mu_i$ and $X_i|\mu_i \sim N(\mu_i, \sigma^2)$ is the most well–known example, but this problem occurs in a wide range of applications. Other common examples include: the Poisson distribution with $\eta_i = \lambda_i$ and the Binomial distribution with $\eta_i = p_i$.

This fundamental problem, largely resolved in the low-dimensional set-

*The research of W. Sun is supported in part by NSF Grants DMS-CAREER-1255406 and DMS-1712983.

Keywords and phrases: Compound decision, empirical Bayes, kernel smoothing, shrinkage estimation, SURE, Tweedie's formula

ting, has recently become a key research question in the high-dimensional setting. For example, in modern large-scale applications it is often of interest to perform selective inference, e.g. first select interesting values of X_i and then carry out inferential analysis for the associated η_i's; see [5, 7, 43, 30], which all discuss approaches for the construction of simultaneous confidence intervals after a selection procedure is applied. 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 [4, 38, 21, 22, 2].

However, the simultaneous inference of thousands of means, or other parameters, is challenging because, as described in [18], 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 [23], has provided an effective bias reduction tool for simultaneous estimation problems. There are several popular classes of methods, including linear shrinkage estimators [23, 19, 6], non–linear thresholding–based estimators motivated by sparse priors [16, 27, 1], and full Bayes or empirical Bayes (EB) estimators with unspecified priors [9, 24, 13]. This article focuses on a class of estimators based on Tweedie's formula, which first appeared in [33]. Tweedie's rule is an elegant shrinkage estimator, for distributions from the exponential family, that has recently received renewed interest [9, 18, 28]. The formula is simple and intuitive since its implementation in the empirical Bayes setting only requires the estimation of the marginal density of X_i. This property is particularly appealing for large–scale estimation problems where nonparametric density estimates can be easily constructed from data. The resultant EB estimator enjoys optimality properties [9] and delivers superior numerical performance [18, 28]. The work of [18] further convincingly demonstrates that Tweedie's formula provides an effective tool for removing selection bias when estimating thousands of parameters simultaneously.

Most of the research in this area has been restricted to models of the form $f(x_i|\eta_i)$ where the distribution of X_i is only a function of η_i. In situations involving a nuisance parameter θ it is generally assumed to be known and identical for all X_i. For example, homoscedastic Gaussian models of the form $X_i|\mu_i, \sigma \overset{ind}{\sim} N(\mu_i, \sigma^2)$ involve the nuisance parameter $\theta = \sigma^2$, but σ^2 is usually treated as known and constant for all X_i. However, in large-scale studies the data are often collected from heterogeneous sources; hence the nuisance parameters often vary over the n observations. Perhaps the most common example, and the setting we concentrate most on, involves heteroscedastic errors, where σ^2 varies over X_i. Microarray data [20, 14],

returns on mutual funds [10], and the state-wide school performance gaps, considered in Section 5.3, are all instances of large-scale data where genes, funds, or schools have heterogeneous variances. Heteroscedastic errors also arise in ANOVA analysis and linear regression [44]. Moreover, in compound Binomial problems, the heterogeneity is reflected by unequal sample sizes across different estimation units. In Section 5.2, we demonstrate that the conventional Tweedie's formula cannot eliminate selection bias for heterogeneous data. Moreover, failing to account for heterogeneity leads to invalid false discovery rate methods [17, 11], unstable multiple testing procedures [41] and inefficient ranking and selection algorithms [22], exacerbating the replicability crisis in large-scale studies [41]. However, relatively few methodologies are available to address this issue.

The heterogeneity setting can be formulated using a hierarchical approach. For n parallel studies, the data for the ith study is summarized by X_i which is modeled as coming from a member of the exponential family of distributions

$$f_{\theta_i}(x_i|\eta_i) \overset{ind}{=} \exp\{\eta_i x_i - \psi_{\theta_i}(\eta_i)\}\, h_{\theta_i}(x_i), \quad i = 1, \ldots, n, \tag{1.1}$$

where η_i and θ_i are independent and drawn from unspecified priors

$$\eta_i \overset{iid}{\sim} G_\eta(\cdot),\ \theta_i \overset{iid}{\sim} G_\theta(\cdot). \tag{1.2}$$

Many common families of distributions fall into this framework, including the Binomial (with $\eta_i = \log(p_i/(1-p_i))$ and $\theta_i = n_i$), Gamma (with $\eta_i = \alpha_i$ and $\theta_i = \beta_i$), and Beta (with $\eta_i = \alpha_i$ and $\theta_i = \beta_i$). In particular, the heteroscedastic problem is a special case of (1.1) and (1.2) where $\eta_i = \mu_i$, $\theta_i = \sigma_i^2$, and $X_i|\mu_i, \sigma_i^2$ follows a normal distribution. A common goal is then to find the estimator, or make the decision, that minimizes the expected squared error loss. Following tradition, we assume σ_i^2, or in general θ_i, are known (take for example, [32], [9], [45], and [44]) and for implementation, use a consistent estimator, as discussed in [44]. An alternative to estimating the nuisance parameter involves placing an objective prior on θ_i as done in [25], which extends the model in [45] with known variance to the case of unknown variance.

A plausible–seeming solution to the heteroscedastic problem might be to scale each X_i by σ_i so that a homoscedastic method could be applied to $X_i^{sc} = X_i/\sigma_i$, 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– or z–scores. A similar standardization is performed when we compute $\hat{p}_i = X_i/n_i$ in the Binomial setting. However, perhaps

somewhat surprisingly, we demonstrate in both our theoretical and empirical results that this approach disregards important information so is inefficient. More advanced methods have been developed, but all suffer from various limitations. For instance, the methods proposed by [45], [40], [25], [29], and [47] are designed for heteroscedastic data but assume a parametric Gaussian prior, which leads to loss of efficiency when the prior is misspecified.

By contrast, we propose a two–step approach, "Nonparametric Empirical Bayes Smoothing Tweedie" (NEST), which first estimates the marginal distribution of X_i, $f_\theta(x)$, and second predicts η_i using a generalized version of Tweedie's formula. Hence, NEST departs from linear shrinkage and other techniques that focus on specific forms of the prior distribution. A significant challenge in the heterogeneous setting is that $f_\theta(x)$ varies with θ, so we must estimate a two–dimensional function. NEST addresses this issue using a kernel which weights observations by their distance in both the x and θ dimensions. The intuition here is that the density function should change smoothly as a function of θ, so observations with variability close to θ can be used to estimate $f_\theta(x)$. Once we obtain this two–dimensional density function, we simply apply Tweedie's formula to estimate the parameter corresponding to any particular combination of x and θ. Compared to grouping methods [44], which group data according to θ_i and then apply linear estimators within each group, NEST is capable of pooling information from all samples to construct a more efficient estimator.

NEST has four clear advantages. First, it is both easy to understand and compute but nevertheless handles general settings. Second, NEST does not rely on any parametric assumptions on $G_\eta(\eta)$. In fact it makes no explicit assumptions about the prior since it directly estimates the marginal density $f_\theta(x)$ using a nonparametric kernel method. Third, we prove that NEST only requires a few simple assumptions to achieve asymptotic optimality for a broad class of models. Additionally, we develop a Stein-type unbiased risk estimate (SURE) criterion for bandwidth tuning, which explicitly resolves the bias–variance tradeoff issue in compound estimation. Finally, we demonstrate numerically that NEST can provide high levels of estimation accuracy relative to a host of benchmark comparison methods.

The rest of the paper is structured as follows. Section 2 develops a generalization of Tweedie's formula to the multivariate setting and provides the corresponding oracle estimator. In Section 3, we present the NEST decision rule and algorithm, and describe the SURE criterion for choosing bandwidths. Section 4 describes the asymptotic setup and provides the main theorem establishing the asymptotic optimality of NEST. Section 5 concludes with a comparison of methods in several simulations and a data application. The

proofs are given in Section 6. Additional theoretical and numerical results are provided in the Supplementary Material.

2. Tweedie's Formula. In Section 2.1, we review Tweedie's formula for univariate models, while Section 2.2 generalizes the formula to multivariate models; a special case of this general result gives the oracle estimator for the nuisance parameter setting.

2.1. *Tweedie's formula.* [18] demonstrated that for the hierarchical model (1.1) and (1.2), with $\theta_i = \theta$, the estimator minimizing the expected squared error loss is given by $\boldsymbol{\delta}^{TF} = (\delta_i^{TF} : 1 \le i \le n)$, where

$$\delta_i^{TF} = \mathbb{E}(\eta_i | x_i) = l_{f,\theta}^{(1)}(x_i) - l_{h,\theta}^{(1)}(x_i), \tag{2.1}$$

where $l_{f,\theta}(x) = \log f_\theta(x)$ and $l_{h,\theta}(x) = \log h_\theta(x)$. In general $h_\theta(x)$ can be directly computed for any particular member of the exponential family while $f_\theta(x)$ can be estimated based on the observed data.

In particular, when $f_{\sigma_i}(x_i|\mu_i)$ is a $N(\mu_i, \sigma_i^2)$ density with, $\sigma_i^2 = \sigma^2$, then

$$\delta_i^{TF} = \mathbb{E}(\mu_i | x_i) = x_i + \sigma^2 \frac{d}{dx} \log f_\sigma(x_i) = x_i + \sigma^2 \frac{f_\sigma^{(1)}(x_i)}{f_\sigma(x_i)} \tag{2.2}$$

as first appeared in [33], where $f_\sigma(x) = \int f_\sigma(x|\mu) dG_\mu(\mu)$ is the marginal density of X_i, and $f_\sigma^{(1)}(x) = \frac{d}{dx} f_\sigma(x)$ its derivative. An important property of the formula is that it only requires the estimation of the marginal distribution of X_i in order to compute the estimator, which is particularly appealing in large-scale studies where one observes thousands of X_i, making it possible to obtain an accurate estimate of $f_\sigma(x)$.

2.2. *Multivariate Tweedie's formula.* Suppose we observe $\boldsymbol{x} = (x_1, , \ldots, x_n)^T$ with conditional distribution $f_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\eta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots \theta_n)^T$ is a vector of known nuisance parameters and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ are parameters of interest. We assume that $\boldsymbol{\eta}$ comes from an unknown prior distribution $\boldsymbol{\eta} \sim G_{\boldsymbol{\eta}}(\boldsymbol{\eta})$. Furthermore, we assume that, conditional on $\boldsymbol{\theta}$, $\boldsymbol{x}$ comes from an exponential family of the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\eta}) = \exp\left\{\boldsymbol{\eta}^T \boldsymbol{x} - \psi_{\boldsymbol{\theta}}(\boldsymbol{\eta})\right\} h_{\boldsymbol{\theta}}(\boldsymbol{x}). \tag{2.3}$$

The Gaussian distribution,

$$f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\}, \tag{2.4}$$

is the most common case of (2.3), where $\boldsymbol{\eta} = \Sigma^{-1}\boldsymbol{\mu}, \boldsymbol{\theta} = \Sigma, \psi_{\boldsymbol{\theta}}(\boldsymbol{\eta}) = \frac{1}{2}\boldsymbol{\eta}^T\Sigma\boldsymbol{\eta}$, and $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right\}$.

We demonstrate that Tweedie's formula can be extended to the multivariate setting. Let $\boldsymbol{\delta} = \{\delta_1(\boldsymbol{x}), \cdots, \delta_n(\boldsymbol{x})\}^T$ be an estimator for $\boldsymbol{\eta}$. Denote $l_n(\boldsymbol{\delta}, \boldsymbol{\eta}) = n^{-1}\|\boldsymbol{\delta} - \boldsymbol{\eta}\|_2^2$ the squared error loss of estimating $\boldsymbol{\eta}$ using $\boldsymbol{\delta}$. The compound Bayes risk of $\boldsymbol{\delta}$ under squared error loss is

$$r(\boldsymbol{\delta}, G_{\boldsymbol{\eta}}) = \int\int l_n(\boldsymbol{\delta}, \boldsymbol{\eta}) f_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\eta}) d\boldsymbol{x} dG_{\boldsymbol{\eta}}(\boldsymbol{\eta}). \tag{2.5}$$

The next theorem derives the optimal estimator under risk (2.5).

THEOREM 1. ***(The multivariate Tweedie's formula)***. *Under Model 2.3, the optimal estimator that minimizes (2.5) is*

$$\boldsymbol{\delta}^{\pi}_{\boldsymbol{\eta}}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{\eta}|\boldsymbol{x}, \boldsymbol{\theta}) = l^{(1)}_{f,\boldsymbol{\theta}}(\boldsymbol{x}) - l^{(1)}_{h,\boldsymbol{\theta}}(\boldsymbol{x}) \tag{2.6}$$

where $l^{(1)}_{f,\boldsymbol{\theta}}(\boldsymbol{x}) = \left\{\frac{d}{dx_1} l_{f,\boldsymbol{\theta}}(\boldsymbol{x}), \ldots, \frac{d}{dx_n} l_{f,\boldsymbol{\theta}}(\boldsymbol{x})\right\}^T, l_{f,\boldsymbol{\theta}}(\boldsymbol{x}) = \log f_{\boldsymbol{\theta}}(\boldsymbol{x}), f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int f_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\eta}) dG(\boldsymbol{\eta})$, *and* $l_{h,\boldsymbol{\theta}}(\boldsymbol{x}) = \log h_{\boldsymbol{\theta}}(\boldsymbol{x})$. *In particular, when* $\boldsymbol{x}|\boldsymbol{\mu}$ *follows the Gaussian distribution (2.4) then* $l^{(1)}_{h,\boldsymbol{\theta}}(\boldsymbol{x}) = -\Sigma^{-1}\boldsymbol{x}$ *and*

$$\boldsymbol{\delta}^{\pi}_{\boldsymbol{\mu}}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{\mu}|\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x} + \Sigma l^{(1)}_{f,\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{x} + \Sigma\frac{\boldsymbol{f}^{(1)}_{\Sigma}(\boldsymbol{x})}{f_{\Sigma}(\boldsymbol{x})}, \tag{2.7}$$

where $f_{\Sigma}(\boldsymbol{x}) = \int f_{\Sigma}(\boldsymbol{x}|\boldsymbol{\mu}) dG_{\boldsymbol{\mu}}(\boldsymbol{\mu})$ *is the marginal density of* $\boldsymbol{X}$.

REMARK 1. *For the Gaussian case, if* $\boldsymbol{\delta}^{\pi}$ *is the minimizer of the risk (2.5), then* $\Sigma\boldsymbol{\delta}^{\pi}$ *will be the minimizer of*

$$r(\boldsymbol{\delta}, G_{\boldsymbol{\mu}}) = \int\int l_n(\boldsymbol{\delta}, \boldsymbol{\mu}) f_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\mu}) d\boldsymbol{x} dG_{\boldsymbol{\mu}}(\boldsymbol{\mu}) \tag{2.8}$$

because $r(\boldsymbol{\delta}, G_{\boldsymbol{\eta}}) = \Sigma^{-2} r(\Sigma\boldsymbol{\delta}, G_{\boldsymbol{\mu}})$. *Hence we end up with the same estimator for* $\boldsymbol{\mu}$ *using either (2.6) or (2.7), both of which will be denoted* $\boldsymbol{\delta}^{\pi}$ *hereinafter.*

Consider the special case where $\eta_1, \ldots, \eta_n$ are independent and also $x_1, \ldots, x_n$ are mutually independent so $f_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\eta}) = \prod_i \exp\{\eta_i x_i - \psi_{\theta_i}(\eta_i)\} h_{\theta_i}(x_i)$. For the Gaussian case this holds when $\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$ is a diagonal matrix, and the elements μ_i in $\boldsymbol{\mu}$ are independent and follow a common prior distribution $G_{\mu}(\cdot)$. Then the next corollary gives the result for the nusiance parameter setting with which we are concerned.

COROLLARY 1. *Under Models 1.1, & 1.2, the optimal estimator is* $\boldsymbol{\delta}^{\pi} = (\delta_1^{\pi}, \cdots, \delta_n^{\pi})^T$, *where*

$$\delta_i^{\pi} = \mathbb{E}(\eta_i | X_i = x_i, \theta_i) = l^{(1)}_{f,\theta_i}(x_i) - l^{(1)}_{h,\theta_i}(x_i), \tag{2.9}$$

where $l_{f,\theta_i}(x_i) = \log f_{\theta_i}(x_i)$ *and* $l_{h,\theta_i}(x_i) = \log h_{\theta_i}(x_i)$. *In particular, for the Gaussian case,* $l^{(1)}_{h,\theta_i}(x_i) = -x_i/\sigma_i^2$ *and*

$$\delta_i^{\pi} = x_i + \sigma_i^2 l^{(1)}_{f,\sigma_i}(x_i) = x_i + \sigma_i^2 \frac{f^{(1)}_{\sigma_i}(x_i)}{f_{\sigma_i}(x_i)}. \tag{2.10}$$

The proof of the corollary is straightforward and omitted. The estimator (2.10) envelopes previous work. If we assume homoscedastic errors $\sigma_i = \sigma$, then $f_{\sigma_i}(x_i) = f_{\sigma}(x_i)$ and (2.10) reduces to the univariate Tweedie's formula (2.2). Appendix A provides analogous calculations to (2.10) for the Binomial, Negative Binomial, Gamma and Beta distributions.

REMARK 2. Notice that Corollary 1 also demonstrates why the standardization approach is sub-optimal. Applying Tweedie directly to the standardized data X_i^{sc} and then multiplying by σ_i gives an estimate of

$$\delta_i^{sc} = \sigma_i \left(x_i^{sc} + \frac{f^{(1)}_{x^{sc}}(x_i^{sc})}{f_{x^{sc}}(x_i^{sc})} \right) = x_i + \sigma_i^2 \frac{f^{(1)}_{x}(x_i)}{f_{x}(x_i)} \tag{2.11}$$

where $f_x(x) = \int f_x(x|\sigma) dG_{\sigma}(\sigma)$ and $f_{x^{sc}}(x^{sc}) = \int \sigma^{-1} f_x(\frac{x}{\sigma}|\sigma) dG_{\sigma}(\sigma)$. The last equality in (2.11) can be established by verifying that $\frac{f^{(1)}_{x^{sc}}(x_i^{sc})}{f_{x^{sc}}(x_i^{sc})} = \sigma_i \frac{f^{(1)}_{x}(x_i)}{f_{x}(x_i)}$. Equations (2.10) and (2.11) have similar forms but the key distinction is that (2.11) computes the density averaged over all σ while the oracle estimator (2.10) uses the conditional distribution at $\sigma = \sigma_i$.

The Bayes rule (2.10) is an oracle estimator, which cannot be implemented directly because the densities $f_{\sigma_i}(x)$ are typically unknown in practice. The next section introduces an empirical Bayes (EB) approach that can tackle the implementation issue. The EB approach also provides a powerful framework for studying the risk properties of the proposed estimator.

3. NEST for the Heteroscedastic Normal Means Problem. This section describes our proposed NEST approach. We primarily concentrate on the Gaussian heteroscedastic case to illustrate the proposed estimation framework. Methodological extensions to other distributions can be done by

following similar strategies; related formulae are provided in Appendix A. In Section 3.1, we consider an empirical Bayes framework for estimating the oracle rule and discuss new challenges with nuisance parameters. Section 3.2 presents our SURE criterion for selecting the tuning parameters.

3.1. *Weighted kernel density estimation.* In order to implement (2.10) we must first form estimates for $f_{\sigma_i}(x_i)$ and $f^{(1)}_{\sigma_i}(x_i)$. We propose a weighted kernel density estimator. Let $\boldsymbol{h} = (h_x, h_\sigma)$ be tuning parameters (bandwidths). Define

$$\hat{f}_{\sigma,\boldsymbol{h}}(x) := \sum_{j=1}^{n} w_j \phi_{h_{xj}}(x - x_j), \tag{3.1}$$

where $w_j \equiv w_j(\sigma, h_\sigma) = \phi_{h_\sigma}(\sigma - \sigma_j)/\{\sum_{j=1}^n \phi_{h_\sigma}(\sigma - \sigma_j)\}$ is the weight that determines the contribution from (x_j, σ_j), $h_{xj} = h_x\sigma_j$ is a bandwidth that varies across j, and $\phi_h(z) = \frac{1}{\sqrt{2\pi}h} \exp\left\{-\frac{z^2}{2h^2}\right\}$ is a Gaussian kernel. The weights w_j have been standardized to ensure that $\hat{f}_{\sigma,\boldsymbol{h}}(x)$ itself is a proper density function. When the errors are homoscedastic, $w_j = 1/n$ for all j and (3.1) becomes the usual kernel density estimator.

Next we give explanations for the proposed estimator. First, estimating $f_{\sigma_i}(x_i)$ directly is difficult as we only have one pair of observations (x_i, σ_i) for each density function. To exploit the fact that $f_\sigma(x)$ changes smoothly as a function of σ, we propose using weights, which are determined by a kernel function, to borrow strength from observations with variability close to σ_i, while placing little weight on points where σ_i and σ_j are far apart. Second, we set $h_{xj} = h_x\sigma_j$, which provides a varying bandwidth to adjust for the heteroscedasticity in the data. Specifically, the bandwidth of the kernel placed on the point X_j is proportional to σ_j; hence data points observed with higher variations are associated with flatter kernels. Our numerical results show that the varying bandwidth provides substantial efficiency gain over fixed bandwidths. A related idea has been used in the variable kernel method (e.g. [36], pp. 21), which employs bandwidths that are proportional to the sparsity of the data points in a region. Finally, a plethora of kernel functions may be used to construct our estimator. We have chosen the Gaussian kernel $\phi_h(\cdot)$ to facilitate our theoretical analysis. Another advantage of using the Gaussian kernel is that it leads to good numerical performance. In contrast, as observed by [9], kernels with heavy tails typically introduce a significant amount of bias in the corresponding EB estimates. Compared to the choice of kernel, the selection of tuning parameters $\boldsymbol{h}$ is a more important issue; a detailed discussion is given in Section 3.2.

We follow the standard method in the literature (e.g. [42]) to obtain the estimate of the derivative $\hat{f}^{(1)}_{\sigma,\boldsymbol{h}}(x)$:

$$\hat{f}^{(1)}_{\sigma,\boldsymbol{h}}(x) := \frac{d}{dx}\hat{f}_{\sigma,\boldsymbol{h}}(x) = \sum_{j=1}^{n} w_j \phi_{h_{xj}}(x - x_j)\left(\frac{x_j - x}{h^2_{xj}}\right). \tag{3.2}$$

Our methodological development is focused on a class of estimation procedures of the form $\boldsymbol{\delta}_{\boldsymbol{h}} \equiv \boldsymbol{\delta}_{\boldsymbol{h}}(\boldsymbol{X}) = (\delta_{\boldsymbol{h},i} : 1 \le i \le n)^T$, where

$$\delta_{\boldsymbol{h},i} = x_i + \sigma_i^2 \frac{\hat{f}^{(1)}_{\sigma_i,\boldsymbol{h}}(x_i)}{\hat{f}_{\sigma_i,\boldsymbol{h}}(x_i)}. \tag{3.3}$$

3.2. *Selecting the tuning parameters.* Implementing (3.3) requires selecting the tuning parameters $\boldsymbol{h} = (h_x, h_\sigma)$. Ideally, these parameters should be chosen to minimize the true risk, which is unknown in practice. We propose using Stein's unbiased risk estimate (SURE; [37]) as a criterion for tuning $\boldsymbol{h}$. Our SURE method requires a training dataset to estimate the densities, as well as an independent dataset to evaluate the true risk. Next we describe a cross–validation approach for constructing the estimator and calculating the corresponding SURE function, which is further used for selecting the tuning parameters.

Let $\mathcal{X} = \{1, \cdots, n\}$ denote the index set of all observations. We first divide the data into K equal or nearly equal subsets so that for each fold $k = 1, \ldots, K$, there is a holdout subset $\mathcal{X}_k$ taken from the full dataset $\mathcal{X}$. Let $\mathcal{X}_k^C = \mathcal{X} \setminus \mathcal{X}_k$. For each $i \in \mathcal{X}_k$, we first use all $x_j \in \mathcal{X}_k^C$ to estimate the density (and corresponding first and second derivatives), and then evaluate the risk using the following SURE function:

$$S_i(\boldsymbol{h}) := S(\boldsymbol{h}; x_i, \sigma_i^2) = \sigma_i^2 + {\sigma_i}^4 \left[\frac{2\hat{f}_{\sigma_i,\boldsymbol{h}}(x_i)\hat{f}^{(2)}_{\sigma_i,\boldsymbol{h}}(x_i) - \left\{\hat{f}^{(1)}_{\sigma_i,\boldsymbol{h}}(x_i)\right\}^2}{\left\{\hat{f}_{\sigma_i,\boldsymbol{h}}(x_i)\right\}^2}\right], \tag{3.4}$$

where the second derivative is computed as

$$\hat{f}^{(2)}_{\sigma,\boldsymbol{h}}(x) = \sum_{j \in \mathcal{X}_k^C} w_j \frac{\phi_{h_{xj}}(x - x_j)}{h^2_{xj}} \left\{\left(\frac{x - x_j}{h_{xj}}\right)^2 - 1\right\}.$$

In the above formula, the dependence of $S_i(\boldsymbol{h})$ upon the training data has been suppressed for notational simplicity. The compound SURE function can be obtained by combining all individual SURE functions defined by

(3.4): $S(\boldsymbol{h}) = \sum_{i=1}^{n} S_i(\boldsymbol{h})$. The tuning parameters are chosen to minimize $S(\boldsymbol{h})$: $\hat{\boldsymbol{h}} = \arg\min_{\boldsymbol{h}} S(\boldsymbol{h})$. Substituting $\hat{\boldsymbol{h}}$ in place of $\boldsymbol{h}$ in (3.3) provides the proposed "Nonparametric Empirical Bayes Smoothing Tweedie" (NEST) estimator, denoted $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \cdots, \hat{\delta}_n)$.

REMARK 3. The SURE criterion is different from the proposals developed in the density estimation literature for bandwidth selection. Since the emphasis in the kernel smoothing literature is to pick a bandwidth to produce a good estimate of the density, this bandwidth may not be the same as that which produces the best decision rule to estimate $\boldsymbol{\mu}$. In fact, the theoretical analyses in [9] (and their Remark 5) suggest that the bandwidth for the compound estimation problem should converge to 0 "just faster" than $(\log n)^{-1}$. By contrast, the optimal choice of bandwidth is equal to $h_x \sim n^{-1/5}$ for a continuously twice differentiable density (e.g. [42]). Our numerical results show that the SURE criterion leads to much improved performance compared to using the standard bandwidth in the density estimation literature.

Finally we prove that (3.4) provides an unbiased estimate of the risk. Let X_* be a generic random variable obeying Models 1.1 & 1.2, from which we also have an independent sample of training data. Denote μ_* the unknown mean and σ_*^2 the variance associated with X_*. The corresponding estimator in the class (3.3) is denoted $\delta^*_{\boldsymbol{h}}$. Again, the dependence of $\delta^*_{\boldsymbol{h}}$ on the training data is suppressed. For a fixed μ_*, the risk associated with $\delta^*_{\boldsymbol{h}}$ is $R(\delta^*_{\boldsymbol{h}}, \mu_*) = E_{X_*|\mu_*}(\delta^*_{\boldsymbol{h}} - \mu_*)^2$, where the expectation is taken with respect to X_* given μ_* and the training data. The corresponding Bayes risk is given by $r(\delta^*_{\boldsymbol{h}}, G) = \int R(\delta^*_{\boldsymbol{h}}, \mu_*) dG_\mu(\mu_*)$. The next proposition justifies the SURE function (3.4).

PROPOSITION 1. *Consider the heteroscedastic Models 1.1 & 1.2. Then we have* $R(\delta^*_{\boldsymbol{h}}, \mu_*) = \mathbb{E}_{X_*|\mu_*}\{S(\boldsymbol{h}; X_*, \sigma_*^2)\}$ *and* $r(\delta^*_{\boldsymbol{h}}, G) = \mathbb{E}_{X_*,\mu_*}\{S(\boldsymbol{h}; X_*, \sigma_*^2)\}$, *where the last expectation is taken with respect to the joint distribution of* (X_*, μ_*) *for a fixed training dataset.*

4. Asymptotic Properties of NEST. This section studies the risk behavior of the proposed NEST estimator and establishes its asymptotic optimality. Section 4.1 describes the basic setup. The main theorem is presented in Section 4.2, where we also explain the main steps and provide some intuitions behind the proofs. Finally, in Section 4.3 we quantify the degree of bias that occurs in selecting extreme values and show that NEST asymptotically corrects this bias.

4.1. *Asymptotic setup.* Consider the hierarchical Models 1.1 and 1.2. We are interested in estimating $\boldsymbol{\mu}$ based on the observed $(\boldsymbol{X}, \boldsymbol{\sigma}^2)$; this is referred to as a *compound decision* problem ([32, 33]) as the performances of the n coordinate–wise decisions will be combined and evaluated together. Let $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)^T$ be a general decision rule. We call $\boldsymbol{\delta}$ a *simple* rule if for all i, δ_i depends only on (x_i, σ_i), and $\boldsymbol{\delta}$ a *compound* rule if δ_i depends also on (x_j, σ_j), $j \neq i$ ([32]). Let $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Then the (compound) Bayes risk is

$$r(\boldsymbol{\delta}, G) = \mathbb{E}_{\boldsymbol{D}} \left\{ l_n(\boldsymbol{\delta}, \boldsymbol{\mu}) \right\}, \tag{4.1}$$

where G is used to denote the unspecified joint prior on $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$.

Consider the oracle setting where the marginal density $f_\sigma(x)$ is known. We have shown that the oracle rule that minimizes the Bayes risk is $\boldsymbol{\delta}^\pi$ [cf. Equation 2.10], which is a simple rule, a useful fact that can be exploited to simplify our analysis. Concretely, let (X, μ, σ) be a generic triple of random variables from the hierarchical Model 1.1 and 1.2, and δ^π the (scalar) oracle rule for estimating μ based on (X, σ). It follows that the risk of the compound estimation problem (e.g. using $\boldsymbol{\delta}^\pi$ to estimate $\boldsymbol{\mu}$) reduces to the risk of a single estimation problem (e.g. using δ^π for estimating μ):

$$r(\boldsymbol{\delta}^\pi, G) = r(\delta^\pi, G) \coloneqq \mathbb{E}_{X,\mu,\sigma} \left\{ (\delta^\pi - \mu)^2 \right\}. \tag{4.2}$$

The oracle risk (4.2) characterizes the optimal performance of all decision rules. Moreover, it is easy to analyze as it can be explicitly written as the following integral

$$\mathbb{E}_{X,\mu,\sigma} \left\{ (\delta^\pi - \mu)^2 \right\} = \int \int \int (\delta^\pi - \mu)^2 \phi_\sigma(x - \mu) dx dG_\mu(\mu) dG_\sigma(\sigma). \tag{4.3}$$

We focus on the setting where the oracle risk is bounded below by a constant: $r(\delta^\pi, G) \geq C > 0$. Following [34], we call a decision rule $\boldsymbol{\delta}$ *asymptotically optimal* if

$$\lim_{n \to \infty} r(\boldsymbol{\delta}, G) = r(\delta^\pi, G). \tag{4.4}$$

The major goal of our theoretical analysis is to show that the NEST estimator $\hat{\boldsymbol{\delta}}$ is asymptotically optimal in the sense of (4.4). The NEST estimator is difficult to analyze because it is a compound rule, i.e. the decision for μ_i depends on all of the elements of $\boldsymbol{X}$ and $\boldsymbol{\sigma}^2$. The compound risk of the form (4.1) cannot be explicitly written as simple integrals as done in (4.3) because each X_i is used twice: for both constructing the estimator and evaluating the risk. We discuss strategies to overcome this difficulty in the next section.

4.2. *Asymptotic optimality of NEST.* We first describe the assumptions that are needed in our theory.

ASSUMPTION 1. *All means are bounded above by a sequence, which grows to infinity at a rate slower than any polynomial of* n*, i.e.* $\forall i$, $|\mu_i| \leq C_n$*, where* $C_n = o(n^\epsilon)$ *for every* $\epsilon > 0$.

This is considered to be a mild assumption as it allows one to take $C_n = O\{(\log n)^k\}$ for any constant k. The particular case of $k = 1/2$ corresponds to interesting scenarios in a range of large–scale inference problems such as signal detection ([15]), sparse estimation ([1]), and false discovery rate analysis ([31, 12]). For example, in the signal detection problem considered by [15], choosing $\mu_n = \sqrt{2r \log n}$, where $0 < r < 1$, makes the global testing problem neither trivial nor too difficult.

Under Assumption 1, it is natural to consider a truncated version of NEST which returns the component-wise $\max(\hat{\delta}_i, K \log n)$ for some large K. The truncation tends to slightly improve the numerical performance. For notational simplicity, the truncated estimator is also denoted by $\hat{\delta}$ and subsequently used in our proof. The modification is proposed mainly for theoretical considerations; we justify in Section B.1.1 that the truncation will always reduce the MSE.

ASSUMPTION 2. *The variances are uniformly bounded, i.e. there exist* σ_l^2 *and* σ_u^2 *such that* $\sigma_l^2 \leq \sigma_i^2 \leq \sigma_u^2$ *for all* i.

This is a reasonable assumption for most real life cases. When σ_i^2 measures the variability of a summary statistic for an inference unit, the assumption is fulfilled when the sample size for obtaining the summary statistic is neither too large nor too small.

Next we state our main theorem, which formally establishes the asymptotic optimality of the proposed NEST estimator.

THEOREM 2. *Consider Models 1.1 and 1.2. Let* $h_x \sim n^{-\eta_x}$ *and* $h_\sigma \sim n^{-\eta_s}$*, where* η_x *and* η_s *are small constants such that* $0 < \eta_x + \eta_s < 1$*. Then under Assumptions 1-2, the NEST estimator* $\hat{\boldsymbol{\delta}}$ *is asymptotically optimal in the sense of* (4.4).

In the remainder of this section, we explain the main ideas and steps in the proof of the theorem. The challenge in analyzing NEST lies in the dependence of $\hat{\delta}_i$ upon all of the elements of $\boldsymbol{X}$ and $\boldsymbol{\sigma^2}$; hence the asymptotic analysis, which involves expectations over the joint distributions of

$(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, is difficult to handle. To overcome the difficulty, we divide the task by proving three propositions. The first proposition involves the study of the risk from applying NEST to a new pair of observations (X, σ^2) obeying Models 1.1 and 1.2:

$$\hat{\delta} = X + \sigma^2 \frac{\hat{f}^{(1)}_{\sigma,\hat{\boldsymbol{h}}}(X)}{\hat{f}_{\sigma,\hat{\boldsymbol{h}}}(X)},$$

where $\hat{f}_{\sigma,\hat{\boldsymbol{h}}}$ and $\hat{f}^{(1)}_{\sigma,\hat{\boldsymbol{h}}}$ are constructed from $\{(x_i, \sigma_i^2) : 1 \le i \le n\}$. As the data used to construct the NEST estimator $\hat{\delta}$ are independent of the new observation, the Bayes risk for estimating μ can be expressed as

$$r(\hat{\delta}, G) = \mathbb{E}_{\boldsymbol{D}} \mathbb{E}_{X,\mu,\sigma} \left\{ (\hat{\delta} - \mu)^2 \right\}. \tag{4.5}$$

The risk (4.5) is relatively easy to analyze because $\mathbb{E}_{X,\mu,\sigma} \left\{ (\hat{\delta} - \mu)^2 \right\}$ can be evaluated explicitly as $\int\int\int (\hat{\delta} - \mu)^2 \phi_\sigma(x - \mu) dx dG_\mu(\mu) dG_\sigma(\sigma)$.

The following proposition, which constitutes a key step in establishing the asymptotic optimality, demonstrates that $r(\hat{\delta}, G)$ is asymptotically equal to the oracle risk.

PROPOSITION 2. *Suppose we apply two scalar decisions: the oracle estimator δ^π and the NEST estimator $\hat{\delta}$, to a new pair (X, σ^2) obeying Models 1.1 and 1.2. Then we have*

$$\mathbb{E}_{\boldsymbol{D}} \mathbb{E}_{X,\mu,\sigma} \left\{ (\hat{\delta} - \delta^\pi)^2 \right\} = o(1).$$

It follows that $\lim_{n\to\infty} r(\hat{\delta}, G) = r(\delta^\pi, G)$.

Proposition 2 is proven via three lemmas, which introduce two intermediate estimators $\tilde{\delta}$ and $\bar{\delta}$, defined rigorously in Section 6.3, that help bridge the gap between the NEST and oracle estimators. Intuitively, $\tilde{\delta}$ is constructed based on a kernel density estimator that eliminates the randomness in X_i, and $\bar{\delta}$ is obtained as an approximation to $\tilde{\delta}$ by further teasing out the variability in σ_j^2. The analysis involves the study of the relationships of the risks of $\tilde{\delta}$ and $\bar{\delta}$ to the risks of the oracle rule δ^π and NEST method $\hat{\delta}$. The three lemmas respectively show that (i) the risk of $\bar{\delta}$ is close to that of δ^π; (ii) the risk of $\tilde{\delta}$ is close to that of $\bar{\delta}$, and (iii) the risk of $\hat{\delta}$ is close to that of $\tilde{\delta}$. Therefore Proposition 2 follows by combining (i) to (iii).

Next we show that the proof of the theorem essentially boils down to proving the asymptotic optimality of a jacknifed NEST estimator $\hat{\boldsymbol{\delta}}^- =$

$(\hat{\delta}^{-1}, \cdots, \hat{\delta}^{-n})$, where $\hat{\delta}^{-i}$ represents the *ith* decision, in which the density and its derivative are fitted using data $\left(\boldsymbol{X}_{-i}, \boldsymbol{\sigma}^2_{-i}\right) = \{(x_j, \sigma_j) : 1 \leq j \leq n, j \neq i\}$, and then applied to (x_i, σ_i). The risk function for the jacknifed estimator is

$$r\left(\hat{\boldsymbol{\delta}}^{-}, G\right) = \mathbb{E}_{\boldsymbol{D}}\left(n^{-1}\left\|\hat{\boldsymbol{\delta}}^{-} - \boldsymbol{\mu}\right\|_2^2\right). \tag{4.6}$$

The next proposition shows that the compound risk of $\hat{\boldsymbol{\delta}}^{-}$ is equal to the univariate risk $r(\hat{\delta}, G)$ from applying NEST to a new pair (X, σ^2).

PROPOSITION 3. *Consider the jackknifed NEST estimator* $\hat{\boldsymbol{\delta}}^{-}$. *Then under the assumptions and conditions of Theorem 2, we have*

$$r\left(\hat{\boldsymbol{\delta}}^{-}, G\right) = r(\hat{\delta}, G) = r(\delta^{\pi}, G) + o(1).$$

Finally, the following proposition shows that the jackknifed NEST estimator is asymptotically equivalent to the full NEST estimator.

PROPOSITION 4. *Consider the full NEST estimator* $\hat{\boldsymbol{\delta}}$ *and the jackknifed NEST estimator* $\hat{\boldsymbol{\delta}}^{-}$. *Then under the assumptions and conditions of Theorem 2, we have* $r\left(\hat{\boldsymbol{\delta}}, G\right) = r\left(\hat{\boldsymbol{\delta}}^{-}, G\right) + o(1)$.

Combining Propositions 2 to 4, we complete the proof of Theorem 2, thereby establishing the asymptotic optimality of NEST.

4.3. *NEST and selection bias.* In large-scale inference, attention has recently moved beyond multiple testing, namely selecting a few non-null cases from a large number of hypotheses, to a new direction that involves estimating the effect sizes of the selected non-null cases. A widely used formulation is the "post-selection inference" framework, which has been applied to, for example, the construction of simultaneous confidence intervals for selected parameters [7, 30] and the hierarchical testing of structured hypotheses [46, 3, 39]. These types of questions have been largely driven by the modern "big data" problems, which often involve first searching large data sets for potential targets for inference, and then making refined decisions on a subset of parameters that appear to be worthy of investigating.

A central issue in large-scale selective inference is that the classical inference tool fails to take into account the *selection bias* (or snooping bias) when estimating many means simultaneously. [18] demonstrated convincingly in his "exponential example" that the histogram of the uncorrected differences

$z_i - \mu_i$ of the largest 20 z-values from 2,000 observations is centered around some positive constant; hence we have a systematic bias due to the selection effect. Next we discuss two theoretical results: the first explicitly characterizes the magnitude of the selection bias under hierarchical Models 1.1 and 1.2, and the second explains why the bias can be removed asymptotically by the proposed NEST estimator. We also present numerical examples in Section 5.2 to illustrate the effectiveness of NEST in correcting the selection bias as well as the failure of the homoscedastic Tweedie method.

THEOREM 3. *Consider Models 1.1 and 1.2 and the selection event* $\{X > t\}$. *When applying the naive estimator* X, *the expected value of the selection bias is given by*

$$\mathbb{E}_{X,\mu|X>t,\sigma^2}(X-\mu) = \sigma^2 \frac{f_\sigma(t)}{1-F_\sigma(t)}, \tag{4.7}$$

where $f_\sigma(t) = \int \phi_\sigma(t-\mu)g(\mu)d\mu$ *and* $F_\sigma(t) = P(X > t|\sigma^2) = \int_t^\infty f_\sigma(x)dx$.

We omit results for other types of selection events such as $\{X < t\}$ and $\{X < t_1 \text{ or } X > t_2\}$, which can be derived in a similar fashion. The next theorem establishes the asymptotic unbiasedness of the NEST estimator.

THEOREM 4. *Under the assumptions and conditions of Theorem 2, the bias term of the NEST estimator is vanishingly small:*

$$\mathbb{E}_{\boldsymbol{D}|X_i>t}(\hat{\delta}_i - \mu_i) = o(1). \tag{4.8}$$

This remarkable result follows from the convergence of NEST w.r.t. the oracle estimator (2.1)

$$\mathbb{E}_{\boldsymbol{D}}\mathbb{E}_{X|X>t,\sigma^2}(\hat{\delta} - \delta^\pi) = o(1), \tag{4.9}$$

and one elegant property of the correction term in the oracle estimator, which says that the conditional expectation of the ratio

$$\mathbb{E}_{X,\mu|X>t,\sigma^2}\left\{\frac{f_\sigma^{(1)}(X)}{f_\sigma(X)}\right\} = \frac{f_\sigma(t)}{1-F_\sigma(t)}, \tag{4.10}$$

after multiplying by σ^2, precisely cancels out the expected selection bias in (4.7). Both (4.9) and (4.10) are proved in Section 6.7.

5. Numeric Results. This section compares the performance of NEST relative to several competing methods using simulated data in Section 5.1, examines how NEST removes selection bias for the heteroscedastic setting in Section 5.2, and demonstrate all methods on a real dataset in Section 5.3.

5.1. *Simulation.* We compare seven approaches: the naive method (Naive), using $\boldsymbol{X}$ without shrinkage; Tweedie's formula (TF) from [9]; the scaling method mentioned in Section 1, which inputs scaled $X_i^{sc} = X_i/\sigma_i$ and outputs the rescaled mean (Scaled); the group linear method (Group L) of [44]; a grouping method (k-Groups), which casts the grouping idea of [44] into a Tweedie's formulation, by first creating k equal sized groups based on the variances and second applying the univariate Tweedie's formula within each group; the semi-parametric monotonically constrained SURE method (SURE-M) from [45]; and from the same paper, the semi-parametric grand-mean sure-shrinkage method (SURE-SG). Note, we implement a truncated version of NEST that has little impact on the numerical performance but is consistent with theory (discussed right after Assumption 1 in Section 4.2). Mean-squared errors (MSE), and associated standard errors (SE), are computed using the differences between the estimated mean from each method and the true $\boldsymbol{\mu}$, over a total of 50 simulation runs.

We simulate $X_i|\mu_i, \sigma_i \overset{iid}{\sim} N(\mu_i, \sigma_i^2)$ for $1 = 1, \ldots, n$, where n, g_μ and g_σ vary across settings with two sample sizes: $n = 5,000$ and $10,000$. Moreover, three scenarios for g_μ are investigated. In the first setting $\mu_i \sim N(3, 1^2)$. In the second setting, μ_i are drawn from a sparse mixture distribution with a 0.7 probability of a point mass at zero, and a 0.3 probability of a draw from $N(3, 0.3^2)$. Finally, under the third setting, μ_i are drawn from a two–point mixture model, with .5 chance of a point mass at 0 and .5 chance of a point mass at 3. For g_σ, we simulate $\sigma \overset{iid}{\sim} U[0.1, \sigma_M]$ and test three different values of σ_M for each g_μ setting. The values of σ_M are chosen so that the ratio between variances $var(\mu)/var(X)$ are reasonably close to .9, .75, and .5 despite the different models for g_μ. This provides an anchor for comparison across the three g_μ settings. In terms of the relationship between μ's variance and the average variance of X conditional on μ, a ratio of .9 means $var(\mu) = 9.5E\{var(X|\mu)\}$; a ratio of .75 means $var(\mu) = 3E\{var(X|\mu)\}$; and a ratio of .5 means $var(\mu) = E\{var(X|\mu)\}$. The first setting is relatively easy, the second is challenging, and the third is extreme. Because each g_μ setting has different hyperparameters, there are small deviations from the exact ratios of $var(\mu)/E\{var(X|\mu)\} = 9.5$, 3, and 1.

All methods, except for Naive, require tuning parameters. For the methods that use kernel estimators (NEST, TF, Scaled, and k-Groups) we minimize our $SURE$ criterion to select the kernel bandwidth, h_x or (h_x, h_s) for NEST, over a grid of possible values. The same bandwidth is used for both the density function and its derivatives. The k-Groups methods additionally must be tuned for the number of groups. To this end, we run the group method on evenly sized groups of 2, 5, and 10. Also, the linear shrinkage

TABLE 1
Normal Model with standard errors in parentheses. Bolded terms represent oracle, and best performing competitor(s).

n	Method	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 9.6$	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 3$	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 1$
5000	Oracle	**0.090 (0.000)**	**0.223 (0.001)**	**0.411 (0.002)**
	Naive	0.103(0.000)	0.336(0.001)	1.024(0.004)
	NEST	0.091(0.000)	0.226(0.001)	0.430(0.002)
	TF	0.091(0.000)	0.228(0.002)	0.444(0.002)
	Scaled	0.095(0.000)	0.270(0.001)	0.659(0.004)
	Group L	**0.090 (0.000)**	0.224(0.001)	0.416(0.002)
	2-Groups	0.091(0.001)	0.231(0.002)	0.462(0.012)
	5-Groups	0.092(0.001)	0.238(0.004)	0.519(0.032)
	10-Groups	0.093(0.001)	0.247(0.006)	0.570(0.053)
	SURE-M	**0.090 (0.000)**	**0.223 (0.001)**	**0.412 (0.002)**
	SURE-SG	**0.090 (0.000)**	0.224(0.001)	0.415(0.002)
10000	Oracle	**0.090 (0.000)**	**0.223 (0.001)**	**0.411 (0.001)**
	Naive	0.103(0.000)	0.335(0.001)	1.022(0.003)
	NEST	**0.090 (0.000)**	0.225(0.001)	0.424(0.001)
	TF	**0.090 (0.000)**	0.228(0.001)	0.461(0.015)
	Scaled	0.095(0.000)	0.267(0.001)	0.638(0.003)
	Group L	**0.090 (0.000)**	0.224(0.001)	0.415(0.001)
	2-Groups	**0.090 (0.000)**	0.228(0.001)	0.469(0.017)
	5-Groups	0.091(0.000)	0.233(0.003)	0.476(0.019)
	10-Groups	0.091(0.001)	0.237(0.004)	0.538(0.047)
	SURE-M	**0.090 (0.000)**	**0.223 (0.001)**	**0.411 (0.001)**
	SURE-SG	**0.090 (0.000)**	**0.223 (0.001)**	0.413(0.001)

methods (Group L, SURE-M, and SURE-SG) have a different set of tuning parameters. The tuning is handled via optimization techniques in the code provided by [44].

Table 1 shows mean squared errors for the normal model $\mu_i \overset{iid}{\sim} N(3, 1^2)$. For this setting, since the analytical form of the mixture density is known, we can additionally compare performance against the oracle rule (2.2). SURE-M consistently leads performance under the normal setting and uniformly achieves oracle performance. Since SURE-M is a linear shrinkage method that moves X towards a global center, it is expected to perform particularly well under this normal setting, which concentrates points around a center at 3. When $n = 5,000$, next best is SURE-SG, but NEST has an MSE only 1 - 5% higher than the oracle's. When $n = 10,000$, SURE-M and SURE-SG lead by a smaller gap, with several competing methods, including NEST, also achieving oracle performance when $var(\mu)$ is 9.6 times $E(var(X|\mu))$. In the extreme case where $var(\mu) = E(var(X|\mu))$ and the gap is largest between NEST and the best estimator, NEST is 3% higher than the oracle.

Table 2 shows the performance of the sparse model for μ_i. For this set-

TABLE 2
Sparse model with standard errors in parentheses. Bolded terms represent best performing methods.

n	Method	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 9.5$	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 3$	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 1$
5000	Naive	0.369(0.002)	1.840(0.008)	6.011(0.026)
	NEST	**0.124 (0.001)**	**0.684 (0.004)**	1.298(0.011)
	TF	0.154(0.001)	0.775(0.003)	1.598(0.007)
	Scaled	0.182(0.002)	0.824(0.005)	1.713(0.014)
	Group L	0.286(0.001)	0.773(0.003)	1.180(0.004)
	2-Groups	0.139(0.002)	0.775(0.024)	1.704(0.132)
	5-Groups	0.168(0.019)	0.885(0.071)	2.833(0.996)
	10-Groups	0.211(0.040)	0.970(0.089)	2.248(0.277)
	SURE-M	0.285(0.001)	0.764(0.002)	**1.159 (0.003)**
	SURE-SG	0.286(0.001)	0.771(0.003)	1.174(0.004)
10000	Naive	0.369(0.001)	1.839(0.005)	6.009(0.015)
	NEST	**0.118 (0.001)**	**0.661 (0.003)**	1.230(0.006)
	TF	0.153(0.001)	0.773(0.002)	1.588(0.004)
	Scaled	0.178(0.001)	0.812(0.003)	1.629(0.010)
	Group L	0.286(0.001)	0.770(0.002)	1.170(0.002)
	2-Groups	0.134(0.001)	0.730(0.009)	1.473(0.023)
	5-Groups	0.136(0.005)	0.785(0.054)	1.727(0.238)
	10-Groups	0.149(0.011)	0.836(0.067)	1.902(0.287)
	SURE-M	0.285(0.001)	0.763(0.002)	**1.155 (0.002)**
	SURE-SG	0.286(0.001)	0.767(0.002)	1.165(0.002)

ting, we apply a stabilizing technique to shrinkage estimators that follow Tweedie's formula (2.7), which ensures that the estimate $\hat{\mu}$ has the same sign as the original data point x. In particular for (x, σ) we set $\hat{\mu}_S = \mathbb{I}\{\text{sgn}(x) = \text{sgn}(\hat{\mu})\} \times \hat{\mu}$, where $\mathbb{I}$ is an indicator function. We do not modify the linear shrinkage methods since they shrink X towards a specific location. In this sparse case, NEST does best in the relatively easy scenario where $Var(\mu) = 9.5E(Var(X|\mu)$ and in the challenging scenario where $Var(\mu) = 3E(Var(X|\mu)$. The linear shrinkage methods are the worst performers other than the naive method in the easiest scenario but perform best in the extreme scenario where the expected conditional variance of X is as large as the variance of μ. In this final scenario, the data is so noisy that the grand mean becomes the best estimator, which is essentially what the linear shrinkage methods produce. Even so, in this extreme setting, NEST has 22% the error of the naive method compared to best performer's 19%. NEST does particularly well in the 9.5 and 3 scenarios due to its ability to adapt to multiple centers and varying distributions.

The simulations also highlight a trade-off between the grouping and homogeneous methods. Grouping methods capture some of the heteroscedasticity in the data, but when the number of groups divides the data into clusters

that are much smaller than the full set, the estimator can become unstable, particularly for data with broad ranges of variance. Homogeneous methods have the opposite characteristics: their use of the whole data set makes them very stable, but can result in biased estimators in heterogeneous data settings. NEST can be seen as a continuous version of the discrete grouping method, and hence provides a compromise between the two approaches: it makes efficient use of all of the data but does not suffer from the bias introduced by using the homogeneous methods.

Table 3 corresponds to μ_i drawn from the two-point model, with half centered at 0 and half at 3. In this setting, the analytical oracle rule is available and included in the comparisons. NEST uniformly performs best in this case. When $n = 5,000$, it has errors that are 10% – 29% higher than the oracle MSE. Next best is the 2-Groups method with errors 23% – 64% higher than the oracle's. When n is $10,000$, the pattern is the same, with NEST leading shrinkage methods with errors between 7% – 22% while the next best method, the 2-Groups method, has errors 21% – 61% higher than the oracle's. Tweedie's-formula type methods appear to do better in a setting like this one, where there are multiple underlying means and highly non-normal distributions.

5.2. *Selection bias.* We also explore the relative abilities of NEST and TF to correct for selection bias in heteroscedastic data. We demonstrate the selection bias problem through two simple settings. In the first, $X_i|\mu_i \sim N(\mu_i, \sigma_g^2)$, where $i = 1, \ldots, 5000$, $\mu_i \sim N(1, .5^2)$, and $\sigma_g = 1$ (or 3) with 70 (or 30%) probability. In the second setting, σ_g is the same but μ_i come from two centers $X_i \sim .7N(0, .5^2 + 1^2) + .3N(5, .5^2 + 3^2)$. In these two-group settings, NEST reduces to TF applied to separate groups. We plot histograms for the differences $\hat{\mu} - \mu$ of the 20 smallest X_i generated from each of 200 simulations. This gives a total of $4,000$ differences for each of the three estimators: Naive, TF, and NEST.

In Figure 1, the top row corresponds to the setting with a single center $\mu_0 = 1$. Naive is plotted with TF on the left while Naive is plotted with NEST on the right. The top row shows Naive shifted leftward, but both TF and NEST are centered around zero. However, while TF has effectively removed the bias, it is clear that NEST has significantly smaller variance than TF, and only slightly higher variance than Naive.

NEST is able to fully use the information from σ_i^2 and produce separate shrinkage effects for the two groups while TF must take a weighted average of the shrinkage. There is a further, and more serious, consequence to the average shrinkage approach that TF takes. While NEST removes selection

TABLE 3
Two-point model with standard errors in parentheses. Bolded terms represent best performing methods.

n	Method	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 9.2$	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 3.2$	$\frac{Var(\mu)}{E\{Var(X\|\mu)\}} = 1$
5000	Oracle	**0.037 (0.001)**	**0.279 (0.002)**	**0.758 (0.003)**
	Naive	0.242(0.001)	0.701(0.003)	2.161(0.010)
	NEST	**0.048 (0.001)**	**0.306 (0.003)**	**0.856 (0.006)**
	TF	0.072(0.001)	0.374(0.003)	0.957(0.004)
	Scaled	0.129(0.001)	0.480(0.003)	1.169(0.006)
	Group L	0.208(0.001)	0.475(0.002)	0.904(0.003)
	2-Groups	0.061(0.002)	0.343(0.006)	0.951(0.024)
	5-Groups	0.066(0.006)	0.392(0.032)	1.120(0.124)
	10-Groups	0.085(0.013)	0.459(0.066)	1.412(0.261)
	SURE-M	0.208(0.001)	0.472(0.002)	0.893(0.003)
	SURE-SG	0.209(0.001)	0.475(0.002)	0.901(0.003)
10000	Oracle	**0.037 (0.001)**	**0.281 (0.001)**	**0.764 (0.0.002)**
	Naive	0.243(0.001)	0.704(0.002)	2.171(0.007)
	NEST	**0.045 (0.001)**	**0.302 (0.002)**	**0.839 (0.003)**
	TF	0.071(0.001)	0.377(0.002)	0.962(0.002)
	Scaled	0.127(0.001)	0.480(0.002)	1.160(0.004)
	Group L	0.209(0.001)	0.476(0.001)	0.905(0.002)
	2-Groups	0.058(0.001)	0.343(0.004)	0.927(0.014)
	5-Groups	0.055(0.003)	0.347(0.016)	0.965(0.047)
	10-Groups	0.063(0.005)	0.387(0.034)	1.135(0.135)
	SURE-M	0.209(0.001)	0.474(0.001)	0.897(0.002)
	SURE-SG	0.210(0.001)	0.476(0.001)	0.903(0.002)

bias by shrinking each X_i by an appropriate amount, TF is over–shrinking some X_i and under–shrinking others. Specifically, TF's shrinkage is

$$\sigma_g^2 \frac{f'(x_i)}{f(x_i)} = (\mu_0 - x_i)\sigma_g^2 \left\{ \frac{1}{v_1^2} w_i + \frac{1}{v_2^2}(1 - w_i) \right\},$$

where $w_i = \{p\phi_{v_1}(x_i - \mu_0)\}/\{p\phi_{v_1}(x_i - \mu_0) + (1-p)\phi_{v_2}(x_i - \mu_0)\}$, $g = 1$ or 2, p is the proportion of X_i from $g = 1$, and $v_g^2 = \tau^2 + \sigma_g^2$. Since these weights are positive, and $1/v_g^2$ increases as σ_g decreases, x_i's with the smaller standard deviation will be under–shrunk while x_i's with the larger standard deviation will be over–shrunk.

The phenomenon is well–captured by the bottom row of Figure 1. Here, the NEST differences shown in the right plot are still unimodally centered around zero, but TF shrinkage issues become visible in the presence of distinct centers. The average shrinkage effect is slightly too mild for the group that should be shrunk toward zero, shifting its peak slightly left of zero, and far too severe for the group that should be shrunk towards five, shifting its peak towards ten. Thus heteroscedasticity prevents even TF from removing

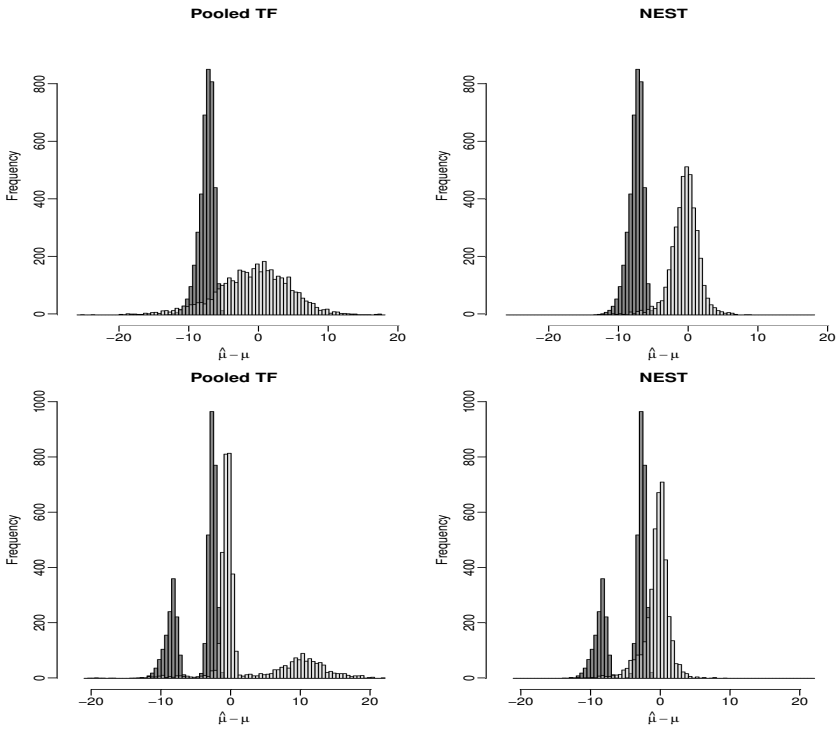

FIG 1. *Differences* $\hat{\mu} - \mu$ *for Naive (dark grey histograms) compared to TF (white histograms, left) and NEST (white histograms, right) for the smallest* 20 *observations across* 200 *simulations in two settings; top row has* $X_i \sim 0.7N(1, .5^2 + 1^2) + 0.3N(1, .5^2 + 3^2)$ *and bottom has* $X_i \sim .7N(0, .5^2 + 1^2) + .3N(5, .5^2 + 3^2)$.

selection bias, but NEST is still able to do so.

5.3. *California API data.* Next, we compare NEST and its competitors on California Academic Performance Index (API) school testing data. The API data files are publicly available on the California Department of Education's website, and the data were described in [35] and analyzed previously by [17] and [38] in the context of multiple testing (with heteroscedastic errors). The data focuses on the within-school achievement gap, for grades 2–12, between socio-economically advantaged (SEA) and socio-economically disadvantaged (SED) students as measured by the difference between the proportion in each group who passed California's standardized math tests, $\hat{p}_A - \hat{p}_D$. During 2002–2010, these test scores were collected in accordance with the No Child Left Behind act and used to unlock federal funding for high-performing schools. With about $7,000$ schools evaluated each year, these performance gaps are susceptible to mean bias. If the observed achievement gaps are not

corrected, small schools with fewer testers are particularly likely to fall in the extreme tails, obscuring large schools with lower variances from scrutiny. For this data, we compare the proportion of large schools selected into the top 100 achievement gaps estimated by each procedure.

The data was cleaned prior to analysis. Observed achievement gaps were calculated over three-year-periods so that the 9 years of data were aggregated into 7 windows: from 2002-2004, 2003-2005, ..., to 2008-2010. The achievement gaps were calculated using the difference of passing rates in each school between socioeconomic groups, $100(\hat{p}_A - \hat{p}_D)$, and served as X_i (in percentages). We chose schools where n_A and n_D were at least 30 students each. Additionally, we used schools that had at least 5 students who passed and 5 students who failed the math test for both SED and SEA groups. The standard deviations, corresponding to s_i, were estimates for σ_i and were calculated using

$$s = 100\sqrt{\hat{p}_A(1-\hat{p}_A)/n_A + \hat{p}_D(1-\hat{p}_D)/n_D}.$$

After cleaning, there were approximately $6,500$ schools in each of the 7 windows, although the number varied slightly from window to window.

To quantify the performance of each method, schools were classified as small, medium, or large based on the socioeconomic group with the smaller number of testers at each school, $n_{min} = \min(n_A, n_D)$. A school was small if the minimum number of testers fell into the bottom quartile of n_{min} amongst all schools, and large if it lay in the top quartile. Since we have defined large schools as those in the top 25% of the population, and, absent information to the contrary, it is reasonable to assume that school size is independent of achievement gap, this is the proportion of large schools expected to appear in the top 100 schools. Competing methods were compared using the squared difference between the observed fraction of large schools chosen in the top 100 by that method, relative to the nominal ideal fraction of 25%, averaged across the 7 three-year windows. We compared NEST to the same methods as in the simulation study. Our SURE criterion was used to tune the bandwidth when applicable. Each bandwidth was tuned on a grid of 0.1 intervals, where the endpoints were adjusted for each method. The k-Groups method was also tuned over 2, 3, 4, and 5 groups. Since the standard deviations were unimodal and did not suggest particular clusters, the groups were constructed based on evenly spaced quantiles over the data.

The mean squared errors (MSE) in Table 4 show that NEST adjusts achievement gaps in a way that most fairly represents large schools. In fact NEST had an average deviation from the benchmark 25% of only about 2.5%. By contrast the Naive method significantly under–counted large

TABLE 4
MSE for the proportion of large schools. Bolded terms represent best performances.

Method	MSE (SE)	Method	MSE (SE)
Naive	0.012(0.002)	2 Group	0.007(0.002)
NEST	**0.001 (0.001)**	3 Group	0.007(0.002)
TF	0.006(0.002)	4 Group	0.007(0.002)
Scaled	0.014(0.002)	5 Group	0.007(0.002)
SURE-M	0.006(0.002)	Group L	0.005(0.002)
SURE-SG	0.005(0.001)		

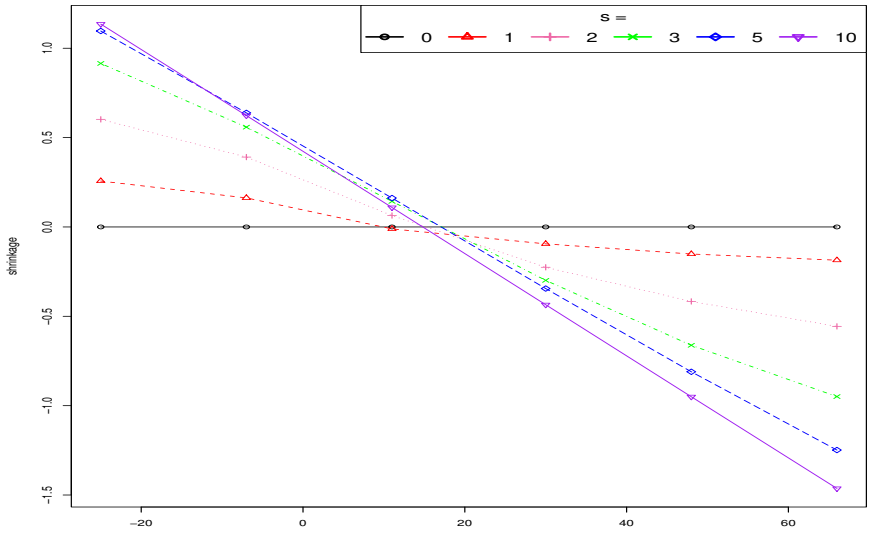

FIG 2. *NEST shrinkage as a function of x and s, plotted against x for different s.*

schools by approximately 12.5%. Even SURE-SG, which was the next most competitive method, had an average deviation over twice that of NEST, at 5.5%. Table 5 in the Appendix shows the full breakdown of proportions of large schools selected by each method on each of the seven windows of data.

Figure 2 visualizes the NEST shrinkage function for various values of s. The black line with circles, corresponding to $s = 0$, is flat since no shrinkage is needed for noiseless observations. We see that shrinkage is approximately linear in x, with the most negative x experiencing the greatest positive shrinkage and vice versa for positive x. Similarly, as s increases, the slope of the shrinkage line becomes steeper, and the shrinkage increases. However, notice that the shrinkage is not linear in s (or s^2), with the marginal change in shrinkage declining as s grows larger. This differs from the standardization approach (2.11) which, for a given value of x, must be linear in s^2, by construction.

6. Proofs. This section proves main theoretical results; additional proofs for some technical lemmas are provided in the Supplementary Material.

6.1. *Proof of Theorem 1.* As described in [18], by Bayes rule,

$$g_{\boldsymbol{\theta}}(\boldsymbol{\eta}|\boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\eta})g(\boldsymbol{\eta})/f_{\boldsymbol{\theta}}(\boldsymbol{x}).$$

But for exponential families of the form given by (2.3)

$$g_{\boldsymbol{\theta}}(\boldsymbol{\eta}|\boldsymbol{x}) = \exp\left\{\boldsymbol{x}^T\boldsymbol{\eta} - \lambda_{\boldsymbol{\theta}}(\boldsymbol{x})\right\}\left[g(\boldsymbol{\eta})\exp(-\psi(\boldsymbol{\eta}))\right],$$

where $\lambda_{\boldsymbol{\theta}}(\boldsymbol{x}) = \log\left(\frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})}{h_{\boldsymbol{\theta}}(\boldsymbol{x})}\right)$. This implies that $\boldsymbol{\eta}|\boldsymbol{x}$ is also an exponential family with canonical parameter $\boldsymbol{x}$ and cumulant generating function (cgf)

$$\log E_{\boldsymbol{\theta}}\left(\exp\left\{\boldsymbol{\eta}^T\mathbf{t}\right\}\right) = \lambda_{\boldsymbol{\theta}}(\boldsymbol{x}+\mathbf{t}) - \lambda_{\boldsymbol{\theta}}(\boldsymbol{x}).$$

We can differentiate the cgf to derive the moments of $\boldsymbol{\eta}|\boldsymbol{x}$. In particular

$$E_{\boldsymbol{\theta}}(\boldsymbol{\eta}|\boldsymbol{x}) = \frac{\partial}{\partial\mathbf{t}}\lambda_{\boldsymbol{\theta}}(\boldsymbol{x}+\mathbf{t})|_{\mathbf{t}=0} = l'_{f,\boldsymbol{\theta}}(\boldsymbol{x}) - l'_{h,\boldsymbol{\theta}}(\boldsymbol{x}). \tag{6.1}$$

The proof for the Gaussian case can be derived either as a special case of (6.1) or following similar arguments to those in [8] and [26]. We begin by expanding the formula for the partial derivatives of $f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})$:

$$\boldsymbol{f}_\Sigma^{(1)}(\boldsymbol{x}) = \int \Sigma^{-1}\boldsymbol{\mu} f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})G_{\boldsymbol{\mu}}(\boldsymbol{\mu}) - \int \Sigma^{-1}\boldsymbol{x} f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})dG_{\boldsymbol{\mu}}(\boldsymbol{\mu}).$$

Then after pulling out Σ^{-1} on the left side and dividing both sides by $f_\Sigma(\boldsymbol{x})$, we get:

$$\frac{\boldsymbol{f}_\Sigma^{(1)}(\boldsymbol{x})}{f_\Sigma(\boldsymbol{x})} = \Sigma^{-1}\left\{\frac{\int \boldsymbol{\mu} f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})dG_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{f_\Sigma(\boldsymbol{x})} - \boldsymbol{x}\right\}.$$

The right side can be simplified according to the Bayes rule:

$$\mathbb{E}(\boldsymbol{\mu}|\boldsymbol{x};\Sigma) = \frac{\int \boldsymbol{\mu} f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})dG_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\int f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})dG_{\boldsymbol{\mu}}(\boldsymbol{\mu})} = \frac{\int \boldsymbol{\mu} f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})dG_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{f_\Sigma(\boldsymbol{x})}.$$

Finally left–multiplying both sides by Σ and taking $\boldsymbol{x}$ to the left side yields the desired result.

6.2. *Proof of Proposition 1.* The proof involves a simple application of the following *Stein's lemma*.

LEMMA 1. *([37]). Consider $X \sim N(\mu, \sigma^2)$ and a differentiable function h with its derivative denoted by $h^{(1)}$. If $\mathbb{E}\{h(X)(X-\mu)\}$ and $\mathbb{E}\{h^{(1)}(X)\}$ exist, then we have $\mathbb{E}\{h(X)(X-\mu)\} = \sigma^2\mathbb{E}\{h^{(1)}(X)\}$.*

Proof. Consider the class of estimators $\delta_{\boldsymbol{h}}$. Let $h(X_*) = \sigma_*^2 \frac{\hat{f}^{(1)}_{\sigma_*,\boldsymbol{h}}(X_*)}{\hat{f}_{\sigma_*,\boldsymbol{h}}(X_*)}$. Then $\delta^*_{\boldsymbol{h}} = X_* + h(X_*)$. Expanding the risk $R(\delta^*_{\boldsymbol{h}}, \mu_*) = E_{X_*|\mu_*}(\delta^*_{\boldsymbol{h}} - \mu_*)^2$, we get three terms inside the expectation:
$\mathbb{E}_{X_*|\mu_*}\left\{(X_*-\mu_*)^2 + 2h(X_*)(X_*-\mu_*) + h^2(X_*)\right\}$. The expectation of the first term is σ_*^2. Applying Stein's lemma to the second term, we get

$$\begin{aligned}\mathbb{E}_{X_*|\mu_*}\{2h(X_*)(X_*-\mu_*)\} &= 2\sigma_*^2\mathbb{E}_{X_*|\mu_*}\left\{h^{(1)}(X_*)\right\} \\ &= 2\sigma_*^4\mathbb{E}_{X_*|\mu_*}\left[\frac{\hat{f}_{\sigma_*,\boldsymbol{h}}(X_*)\hat{f}^{(2)}_{\sigma_*,\boldsymbol{h}}(X_*) - \left\{\hat{f}^{(1)}_{\sigma_*,\boldsymbol{h}}(X_*)\right\}^2}{\left\{\hat{f}_{\sigma_*,\boldsymbol{h}}(X_*)\right\}^2}\right].\end{aligned}$$

The third term can be easily computed as

$$\mathbb{E}_{X_*|\mu_*}\{h^2(X_*)\} = \sigma_*^4\mathbb{E}_{X_*|\mu_*}\left[\left\{\hat{f}^{(1)}_{\sigma_*,\boldsymbol{h}}(X_*)\right\}^2 / \left\{\hat{f}_{\sigma_*,\boldsymbol{h}}(X_*)\right\}^2\right].$$

Combining the three terms gives the desired equality $R\left(\delta^*_{\boldsymbol{h}}, \mu_*\right) = E_{X_*|\mu_*}S(\delta^*_{\boldsymbol{h}}; X_*, \sigma_*^2)$. The second part of the theorem follows directly from the first part.

6.3. *Three lemmas for Proposition 2.* Consider a generic triple of variables (X, μ, σ^2) from Model 1.1 and 1.2. The oracle estimator δ^π and NEST estimators $\hat{\delta}$ are respectively constructed based on $f_\sigma(x) = \int \phi_\sigma(x-\mu)dG_\mu(\mu)$ and $\hat{f}_{\sigma,\hat{\boldsymbol{h}}}(x)$. In our derivation, we use the notation $\hat{f}_\sigma(x)$, where the dependence of $\hat{f}$ on $\hat{\boldsymbol{h}}$ and observed data is suppressed.

We first define two intermediate estimators $\tilde{\delta}$ and $\bar{\delta}$ to bridge the gap between $\hat{\delta}$ and $\bar{\delta}$. The estimator $\tilde{\delta}$ is based on $\tilde{f}$, which is defined as

$$\tilde{f}_\sigma(x) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}\left\{\hat{f}_\sigma(x)\right\}, \tag{6.2}$$

to eliminate the variability in X_i and μ_i. The expression of $\tilde{f}_\sigma(x)$ can be obtained in two steps. The first step takes conditional expectation over $\boldsymbol{X}$ while fixing $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$:

$$\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\sigma}^2}\{\hat{f}_\sigma(x)\} = \sum_{j=1}^n \omega_j \int \phi_{h_x\sigma_j}(y-x)\phi_{\sigma_j}(y-\mu_j)dy = \sum_{j=1}^n \omega_j\phi_{\nu\sigma_j}(x-\mu_j),$$

where $\nu^2 = 1 + h_x^2$. The previous calculation uses the fact that

$$\int \phi_{\sigma_1}(y-\mu_1)\phi_{\sigma_2}(y-\mu_2)dx = \phi_{(\sigma_1^2+\sigma_2^2)^{1/2}}(\mu_1-\mu_2)dx.$$

The next step takes another expectation over $\boldsymbol{\mu}$ conditional on $\boldsymbol{\sigma}^2$. Using notation $\{g * f\}(x) = \int g(\mu)f(x-\mu)d\mu$, we have

$$\tilde{f}_\sigma(x) = \mathbb{E}_{\boldsymbol{\mu}|\boldsymbol{\sigma}^2} \sum_{j=1}^n \omega_j \phi_{v\sigma_j}(x-\mu_j) = \sum_{j=1}^n \omega_j \left\{g_\mu * \phi_{\nu\sigma_j}\right\}(x).$$

The second estimator $\bar{\delta}$ is based on $\bar{f}_\sigma$, which is defined as the limiting value of $\tilde{f}_\sigma$, to eliminate the variability from σ_j. Note that

$$\begin{aligned}\tilde{f}_\sigma(x) &= \frac{\sum_{j=1}^n \left\{g_\mu * \phi_{\nu\sigma_j}\right\}(x)\phi_{h_\sigma}(\sigma_j-\sigma)}{\sum_{j=1}^n \phi_{h_\sigma}(\sigma_j-\sigma)} \\ &= \frac{\int \left\{g_\mu * \phi_{\nu y}\right\}(x)\phi_{h_\sigma}(y-\sigma)dG_\sigma(y)}{\int \phi_{h_\sigma}(y-\sigma)dG_\sigma(y)} + K_n,\end{aligned}$$

where K_n is bounded and $\mathbb{E}_{\boldsymbol{\sigma}^2}(K_n) = O(n^{-\epsilon})$ for some $\epsilon > 0$. Let $L_n \sim n^{-\eta_l}$, $0 < \eta_l < \eta_s$. Define $A_\sigma := [\sigma - L_n, \sigma + L_n]$, with its complement denoted A_σ^C. Next we show that the integral $\int_{A_\sigma^C} \left\{g_\mu * \phi_{\nu y}\right\}(x)\phi_{h_\sigma}(y-\sigma)dG_\sigma(y)$ is vanishingly small. To see this, consider the value of $\phi_{h_\sigma}(\cdot)$ on the boundaries of A_σ^C, where the density is the greatest. Then,

$$\phi_{h_\sigma}(L_n) = (\sqrt{2\pi}h_\sigma)^{-1} \exp\left\{-\frac{1}{2}(L_n/h_\sigma)^2\right\}.$$

The choice of a polynomial rate L_n ensures that for all $y \in A_\sigma^C$, $\phi_{h_\sigma}(y-\sigma) = O(n^{-\varepsilon})$ for some $\varepsilon > 0$. Moreover, the support of σ_i is bounded. It follows that $\int_{A_\sigma^C} \left\{g_\mu * \phi_{\nu y}\right\}(x)\phi_{h_\sigma}(y-\sigma)dG_\sigma(y) = O(n^{-\varepsilon})$. On A_σ, we apply the mean value theorem for definite integrals to conclude that there exists $\bar{\sigma} \in A_\sigma$ such that

$$\int_{A_\sigma} \left\{g_\mu * \phi_{\nu y}\right\}(x)\phi_{h_\sigma}(y-\sigma)dG_\sigma(y) = \left\{g_\mu * \phi_{\nu\bar{\sigma}}\right\}(x) \int_{A_\sigma} \phi_{h_\sigma}(y-\sigma)g_\sigma(y)dy.$$

Following similar arguments we can show that

$$\int \phi_{h_\sigma}(y-\sigma)dG_\sigma(y) = \int_{A_\sigma} \phi_{h_\sigma}(y-\sigma)dG_\sigma(y)\{1+O(n^{-\epsilon})\}.$$

Cancelling the term $\int_{A_\sigma} \phi_{h_\sigma}(y-\sigma)dG_\sigma(y)$ from top and bottom, the limit of $\tilde{f}_\sigma(x)$ can be represented as:

$$\bar{f}_\sigma(x) \coloneqq \{g * \phi_{\nu\bar{\sigma}}\}(x), \text{ for some } \bar{\sigma} \in A_\sigma. \tag{6.3}$$

Finally, the derivatives $\tilde{f}_\sigma^{(1)}(x)$ and $\bar{f}_\sigma^{(1)}(x)$, as well as $\tilde{\delta}$ and $\bar{\delta}$, can be defined correspondingly.

According to the triangle inequality, to prove Proposition 2, we only need to establish the following three lemmas, which are proved, in order, from Section B.1 to Section B.3 in the Supplementary Material.

LEMMA 2. *Under the conditions of Proposition 2,*

$$\int\int\int (\bar{\delta} - \delta^\pi)^2 \phi_\sigma(x-\mu)dxdG_\mu(\mu)dG_\sigma(\sigma) = o(1).$$

LEMMA 3. *Under the conditions of Proposition 2,*

$$\mathbb{E}_{\boldsymbol{\sigma}^2} \int\int\int (\tilde{\delta} - \bar{\delta})^2 \phi_\sigma(x-\mu)dxdG_\mu(\mu)dG_\sigma(\sigma) = o(1).$$

LEMMA 4. *Under the conditions of Proposition 2,*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2} \int\int\int (\hat{\delta} - \tilde{\delta})^2 \phi_\sigma(x-\mu)dxdG_\mu(\mu)dG_\sigma(\sigma) = o(1).$$

6.4. *Proof of Proposition 3.* Consider applying δ_i^- to estimate μ_i. Then

$$r_i(\hat{\delta}^{-i}, G) = \mathbb{E}_{\boldsymbol{X}_{-i},\boldsymbol{\sigma}^2_{-i}} \int\int\int (\hat{\delta}^{-i} - \mu)^2 \phi_\sigma(x-\mu)dxdG_\mu(\mu)dG_\sigma(\sigma).$$

Then the risk function for the jacknifed estimator is

$$r\left(\hat{\boldsymbol{\delta}}^-, G\right) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(n^{-1}\left\|\hat{\boldsymbol{\delta}}^- - \boldsymbol{\mu}\right\|_2^2\right) = n^{-1}\sum_{i=1}^n r_i\left(\hat{\delta}^{-i}, G\right). \tag{6.4}$$

Noting that $r_i(\hat{\delta}^{-i}, G)$ is invariant to i, and ignoring the notational difference between n and $n-1$, we can see that the average risk of $\hat{\boldsymbol{\delta}}^-$ from (6.4) is equal to the univariate risk $r(\hat{\delta}, G)$ defined by (4.5), thereby converting the compound risk of a vector decision to the risk of a scalar decision: $r\left(\hat{\boldsymbol{\delta}}^-, G\right) = r(\hat{\delta}, G)$.

6.5. *Proof of Proposition 4.* Consider the full and jackknifed density estimators $\hat{f}_{\sigma_i}(x_i)$ and $\hat{f}^{-}_{\sigma_i}(x_i)$. Denote $\hat{f}^{(1)}_{\sigma_i}(x_i)$ and $\hat{f}^{(1),-}_{\sigma_i}(x_i)$ the corresponding derivatives. Let $S_i = \sum_{j=1}^{n} \phi_{h_\sigma}(\sigma_i - \sigma_j)$, $S_i^{-} = \sum_{j\neq i} \phi_{h_\sigma}(\sigma_i - \sigma_j)$. Some algebra shows the following relationships:

$$\hat{f}_{\sigma}(x_i) = \frac{S_i^{-}}{S_i}\hat{f}^{-}_{\sigma_i}(x_i) + \frac{1}{2S_i\pi h_\sigma h_x \sigma_i}, \quad \hat{f}^{(1)}_{\sigma_i}(x_i) = \frac{S_i^{-}}{S_i}\hat{f}^{(1),-}_{\sigma_i}(x_i). \tag{6.5}$$

The full and jackknifed NEST estimators are respectively given by:

$$\hat{\delta}_i = x_i + \sigma_i^2 \frac{\hat{f}^{(1)}_{\sigma_i}(x_i)}{\hat{f}_{\sigma_i}(x_i)}, \quad \hat{\delta}_i^{-} = x_i + \sigma_i^2 \frac{\hat{f}^{(1),-}_{\sigma_i}(x_i)}{\hat{f}^{-}_{\sigma_i}(x_i)}. \tag{6.6}$$

Then according to (6.5) and (6.6), we have

$$\begin{aligned}(\hat{\delta}_i - \hat{\delta}_i^{-})^2 &= \left(\frac{\sigma_i}{2\pi h_x h_\sigma}\right)^2 \left\{\frac{\hat{f}^{(1),-}_{\sigma_i}(x_i)}{\hat{f}^{-}_{\sigma_i}(x_i)}\right\}^2 \left\{\frac{1}{S_i \hat{f}_{\sigma_i}(x_i)}\right\}^2 \\ &= O(h_x^{-2}h_\sigma^{-2})O(C_n'^2)\left\{S_i\hat{f}_{\sigma_i}(x_i)\right\}^{-2}\end{aligned}$$

Define $Q_i^{-} = S_i\hat{f}_{\sigma_i}(x_i) - (2\pi h_x h_\sigma \sigma_i)^{-1} = \sum_{j\neq i}\phi_{h_\sigma}(\sigma_i - \sigma_j)\phi_{h_{xj}}(x_i - x_j)$. Then $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(Q_i^{-}\right) = (n-1)\mathbb{E}_{X_i,\mu_i,\sigma_i^2} q(X_i, \sigma_i^2)$, where the q function can be derived using arguments similar to those in Section 6.3:

$$q(x,\sigma^2) = \int \{g_\mu * \phi_{\nu y}\}(x)\phi_{h_\sigma}(y-\sigma)dG_\sigma(y).$$

Let $Y_{ij} = \phi_{h_\sigma}(\sigma_i - \sigma_j)\phi_{h_{xj}}(x_i - x_j)$ and $\bar{Y}_i$ be its average. Then $\mathbb{E}(\bar{Y}_i) = \mathbb{E}\{q(X_i,\sigma_i^2)\}$. Let A_i be the event such that $\bar{Y} < \frac{1}{2}\mathbb{E}(\bar{Y})$. Then applying Hoeffding's inequality, we claim that

$$\mathbb{P}(A_i) \leq P\left\{|\bar{Y}_i - \mathbb{E}(\bar{Y}_i)| \geq \frac{1}{2}(\bar{Y}_i)\right\} = O(n^{-\epsilon})$$

for some $\epsilon > 0$. Note that on event A_i, we have

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left\{(\hat{\delta}_i - \hat{\delta}_i^{-})^2 \mathbb{1}_{A_i}\right\} = O(C_n'^2)O(n^{-\epsilon}) = o(1).$$

for all i. On the complement,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left\{(\hat{\delta}_i - \hat{\delta}_i^{-})^2 \mathbb{1}_{A_i^C}\right\} &\leq O(h_x^{-2}h_\sigma^{-2})O(C_n'^2)\left(Q_i^{-}\right)^{-2} \\ &\leq O(h_x^{-2}h_\sigma^{-2})O(C_n'^2)\left\{\frac{1}{2}(n-1)\mathbb{E}\{q(X_i,\sigma_i^2)\}\right\}^{-2}\end{aligned}$$

Our assumption on bounded σ_i^2 implies that $\mathbb{E}\{q(X_i, \sigma_i^2)$ is bounded below by a constant. Hence

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left\{(\hat{\delta}_i - \hat{\delta}_i^-)^2 \mathbb{1}_{A_i^C}\right\} = O(n^{-2}h_x^{-2}h_\sigma^{-2}C_n'^2) = o(1).$$

Finally we apply the triangle inequality and the compound risk definition to claim that

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(n^{-1}\left\|\hat{\boldsymbol{\delta}}^- - \boldsymbol{\mu}\right\|_2^2\right) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(n^{-1}\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\mu}\right\|_2^2\right) + o(1),$$

which proves the desired result.

6.6. *Proof of Theorem 3.* Consider the distribution of X restricted to the region $[t, +\infty]$. The density of the truncated X is given by

$$f_\sigma(x|X > t) = \frac{d}{dx}\left\{\frac{\int_t^x f_\sigma(y)dy}{\int_t^\infty f_\sigma(y)dy}\right\} = \frac{f_\sigma(x)}{1 - F_\sigma(x)} \text{ for } x > t, \tag{6.7}$$

where f_σ and F_σ are defined in the theorem. The expected value of $X - \mu$ can be computed as

$$\begin{aligned}
\int_t^\infty (x-\mu) f_\sigma(x|X>t)dx &= \frac{\int_t^\infty (x-\mu) f_\sigma(x)dx}{1 - F_\sigma(x)} \\
&= \frac{\int\int_t^\infty (x-\mu)\phi_\sigma(x-\mu)dx dG(\mu)}{1 - F_\sigma(x)} \\
&= -\sigma^2 \frac{\int\int_t^\infty d\phi_\sigma(x-\mu)dG(\mu)}{1 - F_\sigma(x)} \\
&= \sigma^2 \frac{\int \phi_\sigma(t-\mu)dG(\mu)}{1 - F_\sigma(x)} \\
&= \sigma^2 \frac{f_\sigma(t)}{1 - F_\sigma(t)},
\end{aligned}$$

which gives the desired result.

6.7. *Proof of Theorem 4.* We first prove that the oracle estimator δ^π is unbiased. Using the density of the truncated variable (6.7), the conditional expectation of the ratio $f_\sigma^{(1)}(X)/f_\sigma(X)$ can be computed as

$$\begin{aligned}
E_{X,\mu|X>t}\left\{\frac{f_\sigma^{(1)}(X)}{f_\sigma(X)}\right\} &= \int_t^\infty \frac{f_\sigma^{(1)}(x)}{f_\sigma(x)} \frac{f_\sigma(x)}{1 - F_\sigma(t)} dx \\
&= \frac{-f_\sigma(t)}{1 - F_\sigma(t)}.
\end{aligned}$$

By Theorem 3, we claim that the oracle estimator is unbiased.

Next we discuss the general steps to show that $\mathbb{E}_{\boldsymbol{D}|X_i>t}(\hat{\delta}_i - \delta_i^\pi) = o(1)$. Details will be omitted. We only need to prove the convergence of NEST to the oracle Tweedie estimator under the L_2 risk:

$$\mathbb{E}_{\boldsymbol{D}|X_i>t}(\hat{\delta}_i - \delta_i^\pi)^2 = o(1). \tag{6.8}$$

Following similar arguments to those in previous sections, the proof of (6.8) essentially boils down to proving the result for a generic triple (X, μ, σ^2) and the corresponding decision rules $\hat{\delta}$ and δ^π: $\mathbb{E}_{\boldsymbol{D}}\left\{\mathbb{E}_{X|X>t}(\hat{\delta} - \delta^\pi)^2\right\} = o(1)$. The proof can be done similarly to that of Proposition 3 as the conditional density function $f_\sigma(x|X > t)$ is proportional to $f_\sigma(x)$. □

References.

[1] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879 (2008c:62012)

[2] Basu, P., Cai, T. T., Das, K. and Sun, W. (2017). Weighted False Discovery Rate Control in Large-Scale Multiple Testing. *Journal of the American Statistical Association* **0** 0-0.

[3] Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 297–318.

[4] Benjamini, Y. and Hochberg, Y. (1997). Multiple Hypotheses Testing with Weights. *Scandinavian Journal of Statistics* **24** 407-418.

[5] Benjamini, Y. and Yekutieli, D. (2011). False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters. *Journal of the American Statistical Association* **100** 71-81.

[6] Berger, J. O. (1976). Admissible Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss. *Ann. Statist.* **4** 223–226.

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

[8] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics* **42** 855–903.

[9] Brown, L. D. and Greenshtein, E. (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.

[10] Brown, S. J., Goetzmann, W., Ibbotson, R. G. and Ross, S. A. (1992). Survivorship Bias in Performance Studies. *The Review of Financial Studies* **5** 553-580.

[11] Cai, T. T. and Sun, W. (2009). Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481.

[12] Cai, T. and Sun, W. (2017). Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 197–223.

[13] Castillo, I. and van der Vaart, A. (2012). Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101.

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

[15] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 (2005e:62066)

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

[17] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866 (2009f:62015)

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

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

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

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

[22] Henderson, N. C. and Newton, M. A. (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.

[23] James, W. and Stein, C. (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* 361–379. University of California Press, Berkeley, Calif.

[24] Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684.

[25] Jing, B.-Y., Li, Z., Pan, G. and Zhou, W. (2016). On SURE-Type Double Shrinkage Estimation. *Journal of the American Statistical Association* **111** 1696–1704.

[26] Johnstone, I. M. (2015). Gaussian estimation: Sequence and wavelet models. *Draft version*.

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

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

[29] Kou, S. C. and Yang, J. J. (2017). *Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models* In *Big and Complex Data Analysis: Methodologies and Applications* 249–284. Springer International Publishing, Cham.

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

[31] Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393.

[32] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950* 131–148. University of California Press, Berkeley and Los Angeles. MR0044803 (13,480d)

[33] Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. on Math. Statistic. and Prob.* **1** 157–163.
[34] Robbins, H. (1964). The Empirical Bayes Approach to Statistical Decision Problems. *Annals of Mathematical Statistics.* **35** 1–20.
[35] Rogosa, D. (2003). Accuracy of API index and school base report elements: 2003 Academic Performance Index, California Department of Education. Report.
[36] Silverman, B. W. (1986). *Density estimation for statistics and data analysis* **26**. CRC press.
[37] Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Ann. Statist.* **9** 1135–1151.
[38] Sun, W. and McLain, A. C. (2012). Multiple Testing of Composite Null Hypotheses in Heteroscedastic Models. *Journal of the American Statistical Association* **107** 673–687.
[39] Sun, W. and Wei, Z. (2015). Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika* **102** 267–280.
[40] Tan, Z. (2015). Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli* **21** 574–603.
[41] Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98** 5116–5121.
[42] Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing. Chapman and Hall CRC Monographs on Statistics and Applied Probability* **60**. Chapman and Hall CRC.
[43] Weinstein, A., Fithian, W. and Benjamini, Y. (2013). Selection Adjusted Confidence Intervals With More Power to Determine the Sign. *Journal of the American Statistical Association* **108** 165-176.
[44] Weinstein, A., Ma, Z., Brown, L. D. and Zhang, C.-H. (2018). Group-Linear Empirical Bayes Estimates for a Heteroscedastic Normal Mean. *Journal of the American Statistical Association* **0** 1-13.
[45] Xie, X., Kou, S. and Brown, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association* **107** 1465–1479.
[46] Yekutieli, D. (2008). Hierarchical False Discovery Rate–Controlling Methodology. *J. Amer. Statist. Assoc.* **103** 309–316.
[47] Zhang, X. and Bhattacharya, A. (2017). Empirical Bayes, SURE, and sparse normal mean models. Preprint.