Nonparametric empirical Bayes Tweedie's estimator for normal means with heteroscedastic errors

Luella J. Fu

University of Southern California, Los Angeles, USA

Gareth M. James

University of Southern California, Los Angeles, USA

Wenguang Sun

University of Southern California, Los Angeles, USA

Summary. The problem of estimating a vector of normal means, subject to a fixed variance term σ^2, has been extensively studied, but many practical situations instead involve a heterogeneous variance, σ_i^2. Hence, we consider the problem of estimating a vector of normal means with heteroscedastic variances and propose the "Nonparametric Empirical Bayes SURE Tweedie's" (NEST) estimator. As NEST is neither a James-Stein type linear estimator nor a discrete grouping method, its form is unlike previous heteroscedastic approaches. NEST estimates the marginal density of the data $f_{\sigma_i}(x_i)$, for any pair (x_i, σ_i), using a smoothing kernel that weights observations according to their distance from both x_i and σ_i. NEST then applies the estimated density to a generalized version of Tweedie's formula to estimate the corresponding mean vector. NEST is simple to calculate but flexible enough to accommodate general settings. Additionally, a Stein-type unbiased risk estimate (SURE) criterion is developed to select NEST's tuning parameters. Our theoretical results show that NEST is asymptotically optimal, while simulation studies show that it outperforms competitive methods, with substantial efficiency gains in many settings. The method is further demonstrated on a data set measuring the performance gap between social-economically advantaged and disadvantaged students in elementary to secondary school math scores.

Keywords: compound decision, empirical Bayes, kernel smoothing, shrinkage estimation, SURE, Tweedie's formula

1. Introduction

Suppose that we are interested in estimating a population mean vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^T$ based on a vector of summary statistics $\boldsymbol{X} = (X_1, \cdots, X_n)^T$. The problem is a classic low-dimensional problem that has reemerged in the high-dimensional setting. Simultaneously estimating hundreds or thousands of means involves additional challenges 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 an influential and effective approach to simultaneous estimation problems in the high-dimensional setting. 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 full Bayes or empirical Bayes (EB) 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, which made its first appearance in Robbins (1956). Tweedie's rule is an elegant shrinkage estimator that has recently received renewed interest (Brown and Greenshtein, 2009; Efron, 2011; Koenker and Mizera, 2014). 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 in particular appealing for large-scale estimation problems where nonparametric density estimates can be easily constructed from data. The resultant EB estimator enjoys optimality properties (Brown and Greenshtein, 2009), and delivers superior numerical performance (Efron, 2011). Efron (2011) further convincingly demonstrated that Tweedie's formula provides an effective tool for removing selection bias when estimating thousands of means simultaneously.

Most of the research in this area, in particular the study of Tweedie's formula, has been restricted to homoscedastic models of the form $X_i|\mu_i, \sigma \overset{ind}{\sim} N(\mu_i, \sigma^2)$. Due to their difficulty, few methodologies are available for the heteroscedastic case. However, the heteroscedastic setting has many practical applications. For example, microarray data (Erickson and Sabatti, 2005), returns on mutual funds (Brown et al., 1992), and the state-wide school performance gaps, considered in Section 4.2, are all examples of large-scale data where genes, funds, or schools have heterogeneous variances. Moreover, heteroscedastic errors often arise in ANOVA analysis and linear regression (Weinstein et al., 2017). Even the classic baseball data set (Brown, 2008) and the toxoplasmosis study (Efron and Morris, 1975) display heteroscedasticity due to unequal sample sizes in each unit.

The heteroscedastic problem 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 by

$$X_i|\mu_i, \sigma_i \overset{ind}{\sim} N(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n, \tag{1.1}$$

where μ_i and σ_i^2 are independent and drawn from unspecified priors

$$\mu_i \overset{iid}{\sim} G_\mu(\cdot),\ \sigma_i^2 \overset{iid}{\sim} G_\sigma(\cdot). \tag{1.2}$$

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 are known (take for example, Robbins (1951), Brown and Greenshtein (2009), Xie et al. (2012), and Weinstein et al. (2017)) and for implementation, use a consistent estimator, as discussed in Weinstein et al. (2017). An alternative to estimating σ^2 involves placing an objective prior on σ^2 as done in Jing et al. (2016), which extends the model in Xie et al. (2012) 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. However, this implicitly changes the loss function being minimized and hence produces inferior estimates, as we demonstrate in our empirical results. More advanced methods for dealing with heterogeneous variances have been developed, but many existing techniques may not be fully efficient in general settings. 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, which leads to loss of efficiency when the prior is misspecified. Recently, Weinstein et al. (2017) proposed to first group data by heterogeneous variances and then employ a linear shrinkage estimator within each group. The grouping method improves the classical James-Stein estimator by capturing the heteroscedasticity in the data. However, it involves discretizing the variances and prohibits information pooling across groups, both

of which tend to lead to efficiency loss. Moreover, it is unclear how to choose the "optimal" grouping to tradeoff the bias and variance.

By contrast, we propose an approach, "Nonparametric Empirical Bayes SURE Tweedie" (NEST), which adopts a two–step approach, first estimating the marginal distribution of X_i, $f_\sigma(x)$, and its derivative, and second predicting μ_i using a generalized version of Tweedie's formula. Hence, NEST departs from linear shrinkage and other techniques that focus on the form of the prior distribution. A significant challenge in the heterogeneous setting is that $f_\sigma(x)$ varies with σ, 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 σ dimensions. The intuition here is that the density function should change smoothly as a function of σ, so observations with variability close to σ can be used to estimate $f_\sigma(x)$. Once we obtain this two–dimensional density function, we simply apply Tweedie's formula to estimate the mean corresponding to any particular combination of x and σ. Compared to the grouping method (Weinstein et al., 2017), NEST incorporates the heteroscedasticity in a simpler and more accurate manner, and 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 about $G_\mu(\mu)$ or $G_\sigma(\sigma)$. In fact it makes no explicit assumptions about the priors since it directly estimates the marginal density $f_\sigma(x)$ 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 under the heteroscedastic setting. Finally, we demonstrate, via both simulated and real data settings, that NEST can provide high levels of estimation and prediction 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, presents the NEST decision rule and algorithm, and describes the SURE criterion for choosing bandwidths. Section 3 describes the asymptotic setup, lists assumptions needed, and provides the main theorem establishing the asymptotic optimality of NEST. Section 4 concludes with a comparison of methods in several simulations and a data application. The proofs are given in Section 5.

2. Tweedie's Formula and the Empirical Bayes Approach

This section describes our proposed NEST approach. Section 2.1 reviews Tweedie's formula as it has been developed for univariate models and then generalizes the formula to multivariate models; a special case of this general result gives the oracle estimator for the heteroscedastic setting. In Section 2.2, we consider an empirical Bayes framework for estimating the oracle rule and discuss new challenges in the heteroscedastic case. Finally Section 2.3 presents our SURE criterion for selecting the tuning parameters.

2.1. Tweedie's formula for heteroscedastic models

Consider the hierarchical model (1.1) and (1.2) with homoscedastic errors $\sigma_i^2 = \sigma^2$. Let $f_\sigma(x) = \int \phi_\sigma(x-\mu) dG_\mu(\mu)$ be the marginal density of X_i and $f_\sigma^{(1)}(x) = \frac{d}{dx} f_\sigma(x)$ its derivative. Robbins (1956) demonstrated that the estimator minimizing the expected squared error loss

is given by $\boldsymbol{\delta}^{TF} = (\delta_i^{TF} : 1 \leq i \leq n)$, where

$$\delta_i^{TF} = \mathbb{E}(\mu_i | x_i) = x_i + \sigma^2 \frac{f_\sigma^{(1)}(x_i)}{f_\sigma(x_i)}. \tag{2.1}$$

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 in particular appealing in large-scale studies where one observes thousands of X_i, making it possible to obtain an accurate estimate of the marginal density.

We demonstrate that Tweedie's formula can be extended to the multivariate setting. Assume that $\boldsymbol{X} = (X_1, \cdots, X_n)^T$ follows a multivariate normal distribution

$$\boldsymbol{X} | (\boldsymbol{\mu}, \Sigma) \sim N_n(\boldsymbol{\mu}, \Sigma), \tag{2.2}$$

where $\boldsymbol{\mu} \sim G_{\boldsymbol{\mu}}(\cdot)$ and Σ is known. Denote by $f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu})$ the density of $\boldsymbol{X}$ given $\boldsymbol{\mu}$, and $f_\Sigma(\boldsymbol{x}) = \int f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu}) dG_{\boldsymbol{\mu}}(\boldsymbol{\mu})$ the marginal density of $\boldsymbol{X}$.

Let $\boldsymbol{\delta} = \{\delta_1(\boldsymbol{x}), \cdots, \delta_n(\boldsymbol{x})\}^T$ be an estimator for $\boldsymbol{\mu}$. The compound Bayes risk of $\boldsymbol{\delta}$ under squared error loss is

$$r(\boldsymbol{\delta}, G_{\boldsymbol{\mu}}) = \int \int \frac{1}{n} \sum_{i=1}^{n} \{\delta_i(\boldsymbol{x}) - \mu_i\}^2 f_\Sigma(\boldsymbol{x}|\boldsymbol{\mu}) d\boldsymbol{x} dG_{\boldsymbol{\mu}}(\boldsymbol{\mu}). \tag{2.3}$$

The next theorem derives the optimal estimator under risk (2.3).

THEOREM 1. ***(The multivariate Tweedie's formula)***. *Under Model 2.2, the optimal estimator that minimizes* (2.3) *is*

$$\boldsymbol{\delta}^\pi(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{\mu}|\boldsymbol{x}, \Sigma) = \boldsymbol{x} + \Sigma \frac{\boldsymbol{f}_\Sigma^{(1)}(\boldsymbol{x})}{f_\Sigma(\boldsymbol{x})}, \tag{2.4}$$

where $\boldsymbol{f}_\Sigma^{(1)}(\boldsymbol{x})$ *is the partial derivative* $\boldsymbol{f}_\Sigma^{(1)}(\boldsymbol{x}) = \left\{ \frac{d}{dx_1} f_\Sigma(\boldsymbol{x}), \ldots, \frac{d}{dx_n} f_\Sigma(\boldsymbol{x}) \right\}^T$.

Consider the special case where $\Sigma = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$ 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 heteroscedastic 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*

$$\delta_i^\pi = \mathbb{E}(\mu_i | X_i = x_i, \sigma_i) = x_i + \sigma_i^2 \frac{f_{\sigma_i}^{(1)}(x_i)}{f_{\sigma_i}(x_i)}. \tag{2.5}$$

The proof of the corollary is straightforward and omitted.

The estimator (2.5) envelopes previous work. If we assume homoscedastic errors $\sigma_i = \sigma$, then $f_{\sigma_i}(x_i) = f_\sigma(x_i)$ and (2.5) reduces to the univariate Tweedie's formula. The Bayes rule (2.5) 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.

2.2. Weighted kernel density estimation

In order to implement (2.5) 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

$$\hat{f}_{\sigma,\boldsymbol{h}}(x) := \sum_{j=1}^{n} w_j \phi_{h_{xj}}(x - x_j), \tag{2.6}$$

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 (2.6) 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. Silverman, 1986, 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 Brown and Greenshtein (2009), 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 2.3.

We follow the standard method in the literature (e.g. Wand and Jones, 1994) to obtain the estimate of the derivative $\hat{f}^{(1)}_{\sigma,\boldsymbol{h}}(x)$:

$$\hat{f}^{(1)}_{\sigma,\boldsymbol{h}}(x) := \frac{d}{dx}\hat{f}_{\sigma,\boldsymbol{h}}(x) = \sum_{j=1}^{n} w_j \phi_{h_{xj}}(x - x_j)\left(\frac{x_j - x}{h^2_{xj}}\right). \tag{2.7}$$

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

$$\delta_{\boldsymbol{h},i} = x_i + \sigma_i^2 \frac{\hat{f}^{(1)}_{\sigma_i,\boldsymbol{h}}(x_i)}{\hat{f}_{\sigma_i,\boldsymbol{h}}(x_i)}. \tag{2.8}$$

2.3. Selecting the tuning parameters

Implementing (2.8) 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; Stein, 1981) 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:

$$S_i(\boldsymbol{h}) \coloneqq S(\boldsymbol{h}; x_i, \sigma_i^2) = \sigma_i^2 + {\sigma_i}^4 \left[\frac{2\hat{f}_{\sigma_i,\boldsymbol{h}}(x_i)\hat{f}^{(2)}_{\sigma_i,\boldsymbol{h}}(x_i) - \left\{\hat{f}^{(1)}_{\sigma_i,\boldsymbol{h}}(x_i)\right\}^2}{\left\{\hat{f}_{\sigma_i,\boldsymbol{h}}(x_i)\right\}^2} \right], \tag{2.9}$$

where the second derivative is computed as

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

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 (2.9): $S(\boldsymbol{h}) = \sum_{i=1}^n S_i(\boldsymbol{h})$. 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 (2.8) provides the proposed "Nonparametric Empirical Bayes SURE Tweedie" (NEST) estimator, denoted $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \cdots, \hat{\delta}_n)$.

Remark 1. 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 Brown and Greenshtein (2009) (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. Wand and Jones, 1994). Our numerical results show that the SURE criterion leads to much improved performance.

Finally we prove that (2.9) 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 (2.8) 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 (2.9).

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.*

3. Asymptotic Properties of NEST

This section studies the risk behavior of the proposed NEST estimator and establishes its asymptotic optimality. Section 3.1 describes the basic setup. The main theorem is presented in Section 3.2, where we also explain the main steps and provide some intuitions behind the proofs.

3.1. 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 (Robbins, 1951, 1956) 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), $j \neq i$ (Robbins, 1951). Denote $l_n(\boldsymbol{\delta}, \boldsymbol{\mu}) = n^{-1} \|\boldsymbol{\delta} - \boldsymbol{\mu}\|_2^2$ the squared error loss of estimating $\boldsymbol{\mu}$ using $\boldsymbol{\delta}$. Let $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Then the (compound) Bayes risk is

$$r(\boldsymbol{\delta}, G) = \mathbb{E}_{\boldsymbol{D}} \left\{ l_n(\boldsymbol{\delta}, \boldsymbol{\mu}) \right\}, \tag{3.1}$$

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 2.5], 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):

$$r(\boldsymbol{\delta}^\pi, G) = r(\delta^\pi, G) \coloneqq \mathbb{E}_{X,\mu,\sigma} \left\{ (\delta^\pi - \mu)^2 \right\}. \tag{3.2}$$

The oracle risk (3.2) characterizes the optimal performance of all decision rules. Moreover, it is easy to analyze as it can be explicitly written as the following integral

$$\mathbb{E}_{X,\mu,\sigma} \left\{ (\delta^\pi - \mu)^2 \right\} = \int \int \int (\delta^\pi - \mu)^2 \phi_\sigma(x - \mu) dx dG_\mu(\mu) dG_\sigma(\sigma). \tag{3.3}$$

We focus on the setting where the oracle risk is bounded below by a constant: $r(\delta^\pi, G) \geq C > 0$. Following Robbins (1964), we call a decision rule $\boldsymbol{\delta}$ *asymptotically optimal* if

$$\lim_{n \to \infty} r(\boldsymbol{\delta}, G) = r(\delta^\pi, G). \tag{3.4}$$

The major goal of our theoretical analysis is to show that the NEST estimator $\hat{\boldsymbol{\delta}}$ is asymptotically optimal in the sense of (3.4). 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 (3.1) cannot be explicitly written as simple integrals as done in (3.3) 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.

3.2. 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 the interesting scenarios in a range of large-scale inference problems such as signal detection (Donoho and Jin, 2004), sparse estimation (Abramovich et al., 2006), and false discovery rate analysis (Meinshausen and Rice, 2006; Cai and Sun, 2017). For example, in the signal detection problem considered by Donoho and Jin (2004), 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 5.4.1 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* (3.4).

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 $\{(x_i, \sigma_i^2) : 1 \leq i \leq n\}$. 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

$$r(\hat{\delta}, G) = \mathbb{E}_{\boldsymbol{D}} \mathbb{E}_{X,\mu,\sigma} \left\{ (\hat{\delta} - \mu)^2 \right\}. \tag{3.5}$$

The risk (3.5) 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 5.3, 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 $\bar{\delta}$ and $\tilde{\delta}$ 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 $\tilde{\delta}$ is close to that of $\hat{\delta}$. 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 $\left(\boldsymbol{X}_{-i}, \boldsymbol{\sigma}^2_{-i}\right) = \{(x_j, \sigma_j) : 1 \le j \le n, j \ne i\}$, and then applied to (x_i, σ_i). The risk function for the jacknifed estimator is

$$r\left(\hat{\boldsymbol{\delta}}^-, G\right) = \mathbb{E}_{\boldsymbol{D}}\left(n^{-1}\left\|\hat{\boldsymbol{\delta}}^- - \boldsymbol{\mu}\right\|_2^2\right). \tag{3.6}$$

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. Numeric Results

In this section we compare the performance of NEST relative to several competing methods using simulated data in Section 4.1 and a real dataset in Section 4.2. Our simulation results suggest that NEST improves upon other estimators in a wide variety of settings and has similar loss to the oracle estimator when μ_i comes from a normal distribution.

4.1. Simulation

In our simulations we compared five approaches: the naive method ("Naive"), using $\boldsymbol{X}$ without shrinkage; Tweedie's formula ("TF") from Brown and Greenshtein (2009); the scaling method mentioned in Section 1, which inputs scaled $X_i^{sc} = X_i/\sigma_i$ and outputs the rescaled mean ("Scaled"); and a grouping method ("k-Groups"), in the spirit of Weinstein et al. (2017), which first creates k equal sized groups based on the variances and second applies the univariate Tweedie's formula within each group. Finally, we implemented a truncated version of NEST to be consistent with theory (discussed right after Assumption 1 in Section 3.2). The truncation has little impact on the numerical performance. Mean-squared errors ("MSE"), and associated standard errors ("SE"), were computed using the differences between the estimated mean from each method and the true $\boldsymbol{\mu}$, over a total of 50 simulation runs.

In all settings 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. We consider two sample sizes: $n = 5,000$ and $10,000$. Moreover, three scenarios for g_μ are investigated. In the first setting $\mu_i \sim N(2, 0.5^2)$, while in the second setting $\mu_i \sim U[1, 3]$. Finally, under the third setting, μ_i are drawn from a mixture distribution, with a 0.7 probability of a point mass at zero, and a 0.3 probability of a draw from the $N(2, 0.5^2)$ distribution. For g_σ we simulate $\sigma \overset{iid}{\sim} U[0.5, \sigma_M]$ and test three different values, $\sigma_M = 2, 3$ and 4, allowing us to assess the impact of increasing heteroscedasticity. The three settings for each of μ_i and σ_i, along with the two values for n, generate 18 different simulation scenarios.

All methods require selecting the kernel bandwidth, h_x. We use the same bandwidth for both the density function and its derivatives, but select the bandwidth using two different methods. First, we choose h_x to minimize the true risk ("True Risk"), an optimal situation which could not be implemented in practical situations, and then, more realistically, select h_x to minimize our $SURE$ criterion over a grid of possible values ("SURE"). The grouping methods additionally must be tuned for the number of groups. To this end, we run the group linear method on even sized groups of 2, 5, and 10.

Table 1 shows mean squared errors for the normal model $\mu_i \overset{iid}{\sim} N(2, 0.5^2)$. For this setting, since the analytical form of the mixture density is known, we can additionally compare performance against the oracle rule (2.1). In the $5,000$ observation setting, Table 1 shows that for true risk, NEST performs closest to the oracle, followed by the 5-group method. For the level with lowest heteroscedasticity ($\sigma_M = 2$), NEST's true risk is only 20% higher than the oracle, while that of the 5-group method is 31% higher. Amongst the SURE-chosen risks, NEST is still best, but it is followed by the 2-group rather than the 5-group method. When the number of observations are increased to $10,000$, the homogeneous methods and 2-group performance remain approximately constant. The 10-group method sees the greatest improvement but NEST still consistently outperforms all other methods.

The simulations 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. The benefits of NEST become increasingly apparent with higher heterogeneity levels, as the difference between its performance and other methods widens.

Table 1. Normal Model with standard errors in parentheses. Bolded terms represent oracle, and best performances for true and SURE risk in each variability setting

n	h_x	Method	$\sigma_M = 2$	$\sigma_M = 3$	$\sigma_M = 4$
5000	True Risk	Oracle	**0.205 (0.001)**	**0.219 (0.001)**	**0.227 (0.001)**
		Naive	1.744(0.006)	3.578(0.012)	6.114(0.023)
		NEST	**0.245 (0.001)**	**0.312 (0.002)**	**0.404 (0.006)**
		TF	0.372(0.001)	0.653(0.002)	1.040(0.004)
		Scaled	0.577(0.003)	0.976(0.006)	1.399(0.008)
		2 Groups	0.294(0.007)	0.462(0.015)	0.699(0.026)
		5 Groups	0.268(0.005)	0.367(0.009)	0.505(0.015)
		10 Groups	0.285(0.004)	0.396(0.008)	0.548(0.014)
5000	SURE	NEST	**0.283 (0.012)**	**0.348 (0.010)**	**0.526 (0.032)**
		TF	0.373(0.001)	0.655(0.002)	1.048(0.004)
		Scaled	0.586(0.003)	0.996(0.006)	1.444(0.010)
		2 Groups	0.300(0.007)	0.491(0.019)	0.801(0.047)
		5 Groups	0.357(0.017)	0.594(0.045)	1.204(0.134)
		10 Groups	0.470(0.025)	0.935(0.114)	2.551(0.351)
10000	True Risk	Oracle	**0.205 (0.000)**	**0.220 (0.000)**	**0.226 (0.001)**
		Naive	1.746(0.005)	3.591(0.008)	6.093(0.018)
		NEST	**0.236 (0.001)**	**0.290 (0.002)**	**0.352 (0.003)**
		TF	0.372(0.001)	0.653(0.001)	1.030(0.003)
		Scaled	0.578(0.002)	0.964(0.004)	1.373(0.007)
		2 Groups	0.289(0.006)	0.453(0.014)	0.677(0.024)
		5 Groups	0.264(0.004)	0.350(0.006)	0.452(0.010)
		10 Groups	0.262(0.003)	0.349(0.006)	0.448(0.010)
10000	SURE	NEST	**0.247 (0.005)**	**0.321 (0.009)**	**0.428 (0.018)**
		TF	0.372(0.001)	0.655(0.001)	1.033(0.003)
		Scaled	0.584(0.002)	0.978(0.005)	1.401(0.008)
		2 Groups	0.293(0.007)	0.470(0.016)	0.760(0.038)
		5 Groups	0.288(0.008)	0.500(0.038)	0.994(0.136)
		10 Groups	0.358(0.012)	0.732(0.071)	1.381(0.156)

Table 2. Uniform model with standard errors in parentheses. Bolded terms represent best performances for true and SURE risk in each variability setting

n	h_x	Method	$\sigma_M = 2$	$\sigma_M = 3$	$\sigma_M = 4$
5000	True Risk	Naive	1.749(0.006)	3.569(0.013)	6.111(0.022)
		NEST	**0.297 (0.001)**	**0.373(0.003)**	**0.461 (0.005)**
		TF	0.409(0.001)	0.692(0.003)	1.072(0.004)
		Scaled	0.609(0.003)	1.000(0.006)	1.415(0.010)
		2 Groups	0.340(0.002)	0.511(0.003)	0.745(0.005)
		5 Groups	0.321(0.004)	0.426(0.007)	0.562(0.011)
		10 Groups	0.340(0.006)	0.453(0.011)	0.605(0.018)
5000	SURE	NEST	**0.369 (0.022)**	**0.572 (0.070)**	**0.573 (0.028)**
		TF	0.411(0.001)	0.695(0.003)	1.076(0.004)
		Scaled	0.624(0.005)	1.023(0.007)	1.463(0.013)
		2 Groups	0.468(0.040)	0.768(0.106)	1.094(0.137)
		5 Groups	0.504(0.073)	1.023(0.245)	1.243(0.254)
		10 Groups	0.657(0.111)	1.477(0.430)	2.555(0.846)
10000	True Risk	Naive	1.749(0.005)	3.583(0.009)	6.090(0.017)
		NEST	**0.287 (0.001)**	**0.353 (0.002)**	**0.420 (0.003)**
		TF	0.408(0.001)	0.692(0.002)	1.073(0.002)
		Scaled	0.602(0.002)	0.983(0.005)	1.394(0.007)
		2 Groups	0.335(0.001)	0.505(0.002)	0.731(0.003)
		5 Groups	0.304(0.002)	0.398(0.004)	0.508(0.007)
		10 Groups	0.314(0.003)	0.411(0.007)	0.518(0.011)
10000	SURE	NEST	**0.338 (0.015)**	**0.440 (0.032)**	**0.603 (0.049)**
		TF	0.409(0.001)	0.689(0.002)	1.071(0.002)
		Scaled	0.614(0.003)	1.001(0.005)	1.441(0.008)
		2 Groups	0.498(0.033)	0.779(0.033)	1.446(0.086)
		5 Groups	0.449(0.041)	1.072(0.086)	1.363(0.127)
		10 Groups	0.748(0.073)	1.052(0.132)	1.904(0.381)

Table 2 corresponds to the $\mu_i \overset{iid}{\sim} U[1,3]$ model. The analytical Bayes rule is not computable when using the uniform or sparse distributions, but we compare NEST to the remaining four methods. In this setting the homogeneous methods are relatively stronger but are still outperformed by the heterogeneous methods. When using the true risk with $5,000$ observations, NEST has only 17% of the error of the naive method in the $\sigma_M = 2$ setting, while the next best 5-groups method has 18% of the error of the naive method. Using SURE, NEST again does best with 21% of the error of the naive method, while Tweedie's formula does next best with 23% of the error of the naive method. Many of the same patterns from the normal setting hold true for the uniform setting, including the impact of doubling the number of observations to 10,000, though the increase in observations benefits the SURE 2-group method more under the uniform model than under the normal model.

Table 3 shows the performance of the sparse model for μ_i. For this setting, we apply a stabilizing technique to all estimators 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. The difference between the heterogeneous and homogeneous methods is much smaller in this setting. However, NEST still outperforms all the competing methods and provides significant improvements over the Naive approach. Interestingly, for this setting, there is only a 1%–4% decrease in performance for NEST when switching from true risk to SURE. This suggests that while SURE is sensitive to the number of observations, it does not necessarily suffer more than the true-risk estimator would in non-normal settings. Overall, NEST was statistically significantly superior to all other methods in every simulation setting we considered.

4.2. Real 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 Rogosa (2003) and analyzed previously by Efron (2008) and Sun and McLain (2012) 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 2002–2013, 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. Some schools could, by chance, show severe math proficiency differences that are not reflective of true school quality. It is fairer to evaluate schools after correcting for mean bias. Since the true means are unknown, we used each year of data to predict school performance for the next year. For example, we used 2002's data to estimate 2003's achievement gap, and 2003's achievement gap to estimate 2004's. Using years 2002–2012 as training data and years 2003–2013 as testing data, we ran our method on 11 pairs of year–after–year data.

The data was cleaned prior to analysis. The different passing rates in each school between socioeconomic groups, $100(\hat{p}_A - \hat{p}_D)$, 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 σ_i, for the X_i were calculated as

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

Table 3. Sparse model with standard errors in parentheses. Bolded terms represent best performances for true and SURE risk in each variability setting

n	h_x	Method	$\sigma_M = 2$	$\sigma_M = 3$	$\sigma_M = 4$
5000	True Risk	Naive	1.752(0.007)	3.569(0.015)	6.130(0.021)
		NEST	**0.529 (0.002)**	**0.701 (0.003)**	**0.856 (0.005)**
		TF	0.588(0.002)	0.845(0.003)	1.062(0.005)
		Scaled	0.599(0.003)	0.846(0.004)	1.068(0.007)
		2 Groups	0.553(0.003)	0.774(0.004)	1.002(0.007)
		5 Groups	0.546(0.005)	0.741(0.008)	0.937(0.012)
		10 Groups	0.563(0.007)	0.775(0.012)	0.990(0.020)
5000	SURE	NEST	**0.533 (0.003)**	**0.715 (0.004)**	**0.890 (0.005)**
		TF	0.600(0.002)	0.900(0.004)	1.266(0.006)
		Scaled	0.607(0.003)	0.864(0.005)	1.111(0.007)
		2 Groups	0.573(0.005)	0.824(0.014)	1.128(0.032)
		5 Groups	0.642(0.022)	0.911(0.050)	1.605(0.231)
		10 Groups	0.771(0.054)	1.295(0.208)	1.904(0.356)
10000	True Risk	Naive	1.745(0.004)	3.567(0.010)	6.079(0.018)
		NEST	**0.522 (0.001)**	**0.685 (0.002)**	**0.812 (0.003)**
		TF	0.588(0.001)	0.847(0.002)	1.040(0.003)
		Scaled	0.592(0.002)	0.832(0.003)	1.034(0.005)
		2 Groups	0.546(0.002)	0.770(0.003)	0.972(0.005)
		5 Groups	0.530(0.004)	0.715(0.005)	0.874(0.007)
		10 Groups	0.540(0.005)	0.732(0.008)	0.898(0.012)
10000	SURE	NEST	**0.524 (0.001)**	**0.691 (0.002)**	**0.827 (0.003)**
		TF	0.598(0.001)	0.902(0.003)	1.250(0.004)
		Scaled	0.596(0.002)	0.844(0.003)	1.062(0.005)
		2 Groups	0.556(0.003)	0.798(0.008)	1.057(0.026)
		5 Groups	0.573(0.013)	0.828(0.033)	1.183(0.109)
		10 Groups	0.664(0.039)	1.007(0.083)	1.560(0.260)

Table 4. Performance gap prediction errors. Bolded terms represent best performances for SURE risk

Method	MSE (SE)	Method	MSE (SE)
Naive	66.01(0.78)	2 Group	55.18(0.67)
NEST	**54.69 (0.69)**	3 Group	54.93(0.83)
TF	56.01(0.94)	4 Group	55.23(0.84)
Scaled	56.95(0.61)	5 Group	55.29(1.02)

After cleaning, there were approximately $5,500$ schools in each of the 11 pairs of year–after–year data, although the number varied slightly from year to year.

Competing methods were compared using the MSE, calculated by averaging the squared differences between estimated school performance each year and actual school performance the next year, and then taking the average across the 11 windows of such data. The standard errors of the MSE were also calculated. We compared NEST to the same methods as in the simulation study. The SURE criterion developed in this paper was used to tune the bandwidth for all approaches except the naive method, which has no bandwidth parameter. Each bandwidth was tuned over 20 evenly spaced points, where the endpoints were adjusted to each method. The grouping method was also tuned over 2, 3, 4, and 5 groups. Since the standard deviations were unimodal and did not suggest particular groupings, the groups were constructed based on evenly spaced quantiles over the data.

The results, provided in Table 4, show that NEST outperforms all other methods, although the difference is small in some cases. All shrinkage methods clearly improve on the naive approach but still have relatively high errors. This is partly explained by the fact that we are comparing the estimators to the following year's achievement gap, which is itself only a noisy estimate of the true mean level of performance for each school. The higher error rates are also because, when the data is standardized, the standard deviations range from 1 to 12, with more than half the values above the maximum standard deviation of 4 from the simulation section. As we observed in Section 4.1, all methods suffer when standard deviations are very large. NEST needs more observations to cover such a large range well, while the grouping methods need more groups as well as many more observations.

5. Proofs

This section proves all theoretical results.

5.1. Proof of Theorem 1

The proof follows from similar arguments in Brown (1971) and Johnstone (2015). 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 simplifies 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.

5.2. Proof of Proposition 1

The proof involves a simple application of the following *Stein's lemma*.

LEMMA 1. *(Stein, 1981). 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 $\mathbb{E}\left\{h^{(1)}(X)\right\}$ exist, then we have*

$$\mathbb{E}\{h(X)(X-\mu)\} = \sigma^2 \mathbb{E}\left\{h^{(1)}(X)\right\}.$$

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.

5.3. 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 $\tilde{\delta}$ and $\bar{\delta}$. The estimator $\tilde{\delta}$ is based on $\tilde{f}$, which is defined as

$$\tilde{f}_\sigma(x) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}\left\{\hat{f}_\sigma(x)\right\}, \tag{5.7}$$

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

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

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 $\int_{A_\sigma^C}\{g_\mu * \phi_{\nu y}\}(x)\phi_{h_\sigma}(y-\sigma)dG_\sigma(y)$ 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

$$\int_{A_\sigma^C}\{g_\mu * \phi_{\nu y}\}(x)\phi_{h_\sigma}(y-\sigma)dG_\sigma(y) = O(n^{-\varepsilon}).$$

On A_σ, we apply the mean value theorem for definite integrals to conclude that there exists $\bar{\sigma} \in A_\sigma$ such that

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

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:

$$\bar{f}_\sigma(x) \coloneqq \{g * \phi_{\nu\bar{\sigma}}\}(x), \text{ for some } \bar{\sigma} \in A_\sigma. \tag{5.8}$$

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 5.4 to Section 5.6.

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,*

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

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

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

5.4. Proof of Lemma 2

We first argue in Section 5.4.1 that it is sufficient to prove the result over the following domain

$$\mathbb{R}_x \coloneqq \{x : C_n - \log n \le x \le C_n + \log n\}. \tag{5.9}$$

This simplification can be applied to the proofs of other lemmas.

5.4.1. Truncating the domain

Our goal is to show that $(\hat{\delta} - \delta^\pi)^2$ is negligible on $\mathbb{R}_x^C$. Since $|\mu| \le C_n$ by Assumption 1, the oracle estimator is bounded:

$$\delta^\pi = \mathbb{E}(X|\mu,\sigma^2) = \frac{\int \mu \phi_\sigma(x-\mu) dG_\mu(\mu)}{\int \phi_\sigma(x-\mu) dG_\mu(\mu)} < C_n.$$

Let $C_n' = C_n + \log n$. Consider the truncated NEST estimator $\hat{\delta} \wedge C_n'$. The two intermediate estimators $\tilde{\delta}$ and $\bar{\delta}$ are truncated correspondingly without altering their notations. Letting $\mathbb{1}_{\mathbb{R}_x}$ be the indicator function that is 1 on $\mathbb{R}_x$ and 0 elsewhere. Our goal is to show that

$$\int\int\int_{\mathbb{R}_{\boldsymbol{x}}^C} (\hat{\delta} - \delta^\pi)^2 \phi_\sigma(x-\mu) dx dG_\mu(\mu) dG_\sigma(\sigma) = O(n^{-\kappa}) \tag{5.10}$$

for some small $\kappa > 0$. Note that for all $x \in \mathbb{R}_x^C$, the normal tail density vanishes exponentially: $\phi_\sigma(x-\mu) = O(n^{-\epsilon'})$ for some $\epsilon' > 0$. The desired result follows from the fact that $(\hat{\delta} - \delta^\pi)^2 = o(n^\eta)$ for any $\eta > 0$, according to the assumption on C_n.

5.4.2. Proof of the lemma

We first apply triangle inequality to obtain

$$(\bar{\delta} - \delta^\pi)^2 \le \sigma^4 \left\{\frac{f_\sigma^{(1)}(x)}{f_\sigma(x)}\right\}^2 \left\{\frac{f_\sigma(x)}{\bar{f}_\sigma(x)}\right\}^2 \left[\left\{\frac{\bar{f}_\sigma^{(1)}(x)}{f_\sigma^{(1)}(x)} - 1\right\}^2 + \left\{\frac{\bar{f}_\sigma(x)}{f_\sigma(x)} - 1\right\}^2\right]^2.$$

Hence the lemma follows if we can prove the following facts for $x \in \mathbb{R}_x$.

(i) $f_\sigma^{(1)}(x)/f_\sigma(x) = O(C_n')$, where $C_n' = C_n + \log n$.

(ii) $\bar{f}_\sigma(x)/f_\sigma(x) = 1 + O(n^{-\varepsilon})$ for some $\varepsilon > 0$.

(iii) $\bar{f}_\sigma^{(1)}(x)/f_\sigma^{(1)}(x) = 1 + O(n^{-\varepsilon})$ for some $\varepsilon > 0$.

To prove (i), note that $\delta^{\pi} = O(C_n)$ as shown earlier, $x = O(C_n')$ if $x \in \mathbb{R}_x$. The oracle estimator satisfies $\delta^{\pi} = x + \sigma^2 f_{\sigma}^{(1)}(x)/f_{\sigma}(x)$. By Assumption 2, G_{σ} has a finite support, we claim that $f_{\sigma}^{(1)}(x)/f_{\sigma}(x) = O(C_n)$.

Now consider claim (ii). Let $\mathcal{A}_{\mu} \coloneqq \left\{\mu : |\mu - x| \leq \sqrt{\log(n)}\right\}$. Following similar arguments in previous sections, we apply the normal tail bounds to claim that $\phi_{\nu\bar{\sigma}}(\mu - x) = O\{n^{-1/(2\sigma^2+1)}\}$. Similar arguments apply to $f_{\sigma}(x)$ when $\mu \in \mathcal{A}_{\mu}$. Therefore

$$\frac{\bar{f}_{\sigma}(x)}{f_{\sigma}(x)} = \frac{\int_{\mu\in\mathcal{A}_{\mu}} \phi_{\nu\bar{\sigma}}(x-\mu)dG_{\mu}(\mu)}{\int_{\mu\in\mathcal{A}_{\mu}} \phi_{\sigma}(x-\mu)dG_{\mu}(\mu)} \left\{1 + O(n^{-\kappa_1})\right\} \tag{5.11}$$

for some $\kappa_1 > 0$. Next, we evaluate the ratio in the range of $\mathcal{A}_{\mu}$:

$$\frac{\phi_{\nu\bar{\sigma}}(\mu - x)}{\phi_{\sigma}(\mu - x)} = \frac{\sigma}{(\nu\bar{\sigma})} \exp\left[-\frac{1}{2}(\mu - x)^2 \left\{\frac{1}{(\nu\bar{\sigma})^2} - \frac{1}{\sigma^2}\right\}\right] = 1 + O(n^{-\kappa_2}) \tag{5.12}$$

for some $\kappa_2 > 0$. This result follows from our definition of $\bar{\sigma}$, which is in the range of $[\sigma - L_n, \sigma + L_n]$ for some $L_n \sim n^{-\eta_l}$. Since the result (5.12) holds for all μ in $\mathcal{A}_{\mu}$, we have

$$\begin{aligned}\int_{\mu\in\mathcal{A}_{\mu}} \phi_{\bar{\sigma}}(x-\mu)dG_{\mu}(\mu) &= \int_{\mu\in\mathcal{A}_{\mu}} \phi_{\sigma}(x-\mu)\frac{\phi_{\nu\bar{\sigma}}(\mu-x)}{\phi_{\sigma}(\mu-x)}dG_{\mu}(\mu) \\ &= \int_{\mu\in\mathcal{A}_{\mu}} \phi_{\bar{\sigma}}(x-\mu)dG_{\mu}(\mu)\left\{1 + O(n^{-\kappa_2})\right\}.\end{aligned}$$

Together with (5.11), claim (ii) holds true.

To prove claim (iii), we first show that

$$f_{\sigma}^{(1)}(x) = \int \phi_{\sigma}(x-\mu)\frac{\mu - x}{\sigma^2}dG_{\mu}(\mu) = \int_{\mu\in\mathcal{A}_{\mu}} \phi_{\sigma}(x-\mu)\frac{\mu - x}{\sigma^2}dG_{\mu}(\mu)\left\{1 + O(n^{-\kappa_2})\right\}$$

for some $\kappa > 0$. The above claim holds true by using similar arguments of normal tails (as the term $(x - \mu)$ essentially has no impact on the rate). We can argue similarly that

$$\begin{aligned}\bar{f}_{\sigma}^{(1)}(x) &= \int_{\mathcal{A}_{\mu}} \frac{\sigma^2}{(\nu\bar{\sigma})^2}\frac{\phi_{\nu\bar{\sigma}}(\mu-x)}{\phi_{\sigma}(\mu-x)}\phi_{\sigma}(\mu-x)\frac{\mu-x}{\sigma^2}dG_{\mu}(\mu) \\ &= f_{\sigma}^{(1)}(x)\{1 + O(n^{-\varepsilon})\}\end{aligned}$$

for some $\epsilon > 0$. This proves (iii) and thus completes the proof of the lemma. Note that the proof is done without using the truncated version of $\bar{\delta}$. Since the truncation will always reduce the MSE, the result holds for the truncated $\bar{\delta}$ automatically.

5.5. Proof of Lemma 3

It is sufficient to prove the result over $\mathbb{R}_x$ defined in (5.9). Begin by defining $R_1 = \tilde{f}_{\sigma}^{(1)}(x) - \bar{f}_{\sigma}^{(1)}(x)$ and $R_2 = \tilde{f}_{\sigma}(x) - \bar{f}_{\sigma}(x)$. Then we can represent the squared difference as

$$(\tilde{\delta} - \bar{\delta})^2 = O\left(\left\{\frac{R_1}{\bar{f}_{\sigma}(x) + R_2}\right\}^2 + \left[\frac{R_2\bar{f}_{\sigma}^{(1)}(x)}{\bar{f}_{\sigma}(x)\{\bar{f}_{\sigma}(x) + R_2\}}\right]^2\right). \tag{5.13}$$

Consider L_n defined in the previous section. We first study the asymptotic behavior of R_2.

$$R_2 = \sum_{\sigma_j \in \mathcal{A}_\sigma} w_j \left\{ f_{\sigma_j}(x) - f_{\nu\bar{\sigma}}(x) \right\} + K_n(\sigma), \tag{5.14}$$

where the last term can be calculated as

$$K_n(\sigma) = \sum_{\sigma_j \in \mathcal{A}_\sigma^C} w_j \left\{ f_{\sigma_j}(x) - f_{\nu\bar{\sigma}}(x) \right\} = O\left(\sum_{\sigma_j \in \mathcal{A}_\sigma^C} w_j \right).$$

The last equation holds since both $f_{\sigma_j}(x)$ and $f_{\nu\bar{\sigma}}(x)$ are bounded according to our assumption $\sigma_l^2 \le \sigma_j^2 \le \sigma_u^2$ for all j. Consider $\mathcal{A}_\mu \coloneqq \left\{ \mu : |\mu - x| \le \sqrt{\log(n)} \right\}$. We have

$$\frac{f_{\nu\bar{\sigma}}(x)}{f_{\sigma_j}(x)} = \frac{\int_{\mu \in \mathcal{A}_\mu} \phi_{\nu\bar{\sigma}}(x-\mu) dG_\mu(\mu)}{\int_{\mu \in \mathcal{A}_\mu} \phi_{\sigma_j}(x-\mu) dG_\mu(\mu)} \left\{ 1 + O(n^{-\kappa_1}) \right\}$$

for some $\kappa_1 > 0$, and in the range of $\mu \in \mathcal{A}_\mu$, we have

$$\phi_{\nu\bar{\sigma}}(\mu - x) / \phi_\sigma(\mu - x) = 1 + O(n^{-\kappa_2})$$

for some $\kappa_2 > 0$ and all j such that $\sigma_j \in \mathcal{A}_\sigma$. We conclude that the first term in (5.14) is $O(n^{-\kappa})$ for some $\kappa > 0$ since $f_{\sigma_j}(x)$ is bounded and $\sum_{j \in \mathcal{N}_\sigma} w_j \le 1$.

Now we focus on the asymptotic behavior of $\sum_{\sigma_j \in A_\sigma^C} \omega_{\sigma_j}(\sigma)$. Let K_1 be the event that $n^{-1} \sum_{j=1}^n \phi_{h_\sigma}(\sigma_j - \sigma) < \frac{1}{2} \{ g_\sigma * \phi_{h_\sigma} \}(\sigma)$ and K_2 the event that

$$n^{-1} \sum_{j=1}^n \mathbb{1}_{\{\sigma_j \in A_\sigma^C\}} \phi_{h_\sigma}(\sigma_j - \sigma) > 2 \int_{A_\sigma^C} g_\sigma(y) \phi(y - \sigma) dy.$$

Let $Y_j = \phi_{h_\sigma}(\sigma_j - \sigma)$. Then for $a_j \le Y_j \le b_j$, we use the Hoeffding's inequality

$$\mathbb{P}\left(|\bar{Y} - \mathbb{E}(\bar{Y})| \ge t \right) \le 2 \exp\left\{ - \frac{2n^2 t^2}{\sum_{j=1}^n (b_i - a_i)^2} \right\}.$$

Taking $t = \frac{1}{2}\mathbb{E}(Y_i)$, we have

$$\mathbb{P}(K_1) \le 2 \exp\left\{ - \frac{(1/2) n^2 \{\mathbb{E}(Y_i)\}^2}{n \cdot O(h_\sigma^{-1})} \right\} = O(n^{-\epsilon})$$

for some $\epsilon > 0$. Similarly we can show that $\mathbb{P}(K_2) = O(n^{-\epsilon})$ for some $\epsilon > 0$. Moreover, on the event $K = K_1^C \cap K_2^C$, we have

$$\sum_{\sigma_j \in A_\sigma^C} \omega_{\sigma_j}(\sigma) \le \frac{4 \int_{A_\sigma^C} g_\sigma(y) \phi(y - \sigma) dy}{\{ g_\sigma * \phi_{h_\sigma} \}(\sigma)} = O(n^{-\epsilon})$$

for some $\epsilon > 0$. We use the same ϵ in the previous arguments, which can be achieved easily by appropriate adjustments (taking the smallest). Previously we have shown that the first term in (5.14) is $O(n^{-\epsilon})$. Hence on event K, $R_2 = O(n^{-\kappa})$ for some $\kappa > 0$.

Now consider the domain $\mathbb{R}_x$. Define $\mathbb{S}_x := \{x : \bar{f}_\sigma(x) > n^{-\kappa'}\}$, where $0 < \kappa' < \kappa$. On $\mathbb{R}_x \cap \mathbb{S}_x^C$, we have

$$\int\int_{\mathbb{R}_x \cap \mathbb{S}_x^C} (\tilde{\delta} - \bar{\delta})^2 f_\sigma(x) dx dG_\sigma(\sigma) = O\{C_n'^2 \cdot \mathbb{P}(\mathbb{R}_x \cap \mathbb{S}_x^C)\} = O(n^{-\kappa}) \tag{5.15}$$

for some $\kappa > 0$. The previous claim holds true since the length of $\mathbb{R}_x$ is bounded by C_n', and both $\tilde{\delta}$ and $\bar{\delta}$ are truncated by C_n'.

Now we only need to prove the result for the region $\mathbb{R}_x \cap \mathbb{S}_x$. On event K, we have

$$\mathbb{E}_{\boldsymbol{\sigma}^2}\left(\mathbb{1}_K \cdot \int\int_{\mathbb{R}_x \cap \mathbb{S}_x} \left[\frac{R_2 \bar{f}_\sigma^{(1)}(x)}{\bar{f}_\sigma(x)\{\bar{f}_\sigma(x) + R_2\}}\right]^2 f_\sigma(x) dx dG_\sigma(\sigma)\right) = O(C_n'^2) O\left(n^{-(\kappa-\kappa')}\right),$$

which is $O(n^{-\eta})$ for some $\eta > 0$. On event K^C,

$$\mathbb{E}_{\boldsymbol{\sigma}^2}\left(\mathbb{1}_{K^C} \cdot \int\int_{\mathbb{R}_x \cap \mathbb{S}_x} (\tilde{\delta} - \bar{\delta})^2 f_\sigma(x) dx dG_\sigma(\sigma)\right) = O(C_n'^2) O(n^{-\epsilon}),$$

which is also $O(n^{-\eta})$. Hence the risk regarding the second term of (5.13) is vanishingly small. Similarly, we can show that the first term satisfies

$$\mathbb{E}_{\boldsymbol{\sigma}^2}\left(\int\int_{\mathbb{R}_x \cap \mathbb{S}_x} \left\{\frac{R_1}{\bar{f}_\sigma(x) + R_2}\right\}^2 f_\sigma(x) dx dG_\sigma(\sigma)\right) = O(n^{-\eta}).$$

Together with (5.15), we establish the desired result.

5.6. Proof of Lemma 4

Let $S_1 = \hat{f}_\sigma^{(1)}(x) - \tilde{f}_\sigma^{(1)}(x)$ and $S_2 = \hat{f}_\sigma(x) - \tilde{f}_\sigma(x)$. Then

$$(\tilde{\delta} - \hat{\delta})^2 \le 2\sigma^4 \left[\left\{\frac{\tilde{f}_\sigma^{(1)}(x)}{\tilde{f}_\sigma(x)}\right\}^2 \left\{\frac{S_2}{S_2 + \tilde{f}_\sigma(x)}\right\}^2 + \left\{\frac{S_1}{S_2 + \tilde{f}_\sigma(x)}\right\}^2\right]. \tag{5.16}$$

According to the definition of $\tilde{f}_\sigma(x)$ [cf. equation (5.7)], we have $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}(S_2) = 0$. By doing differentiation on both sides we further have $\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}(S_1) = 0$.

A key step in our analysis is to study the variance of S_2. We aim to show that

$$\mathbb{V}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}(S_2) = O(n^{-1} h_\sigma^{-1} h_x^{-1}). \tag{5.17}$$

To see this, first note that $\mathbb{V}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}(S_2) = \sum_{j=1}^n w_j^2 \mathbb{V}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}\{\phi_{h_{xj}}(x - X_j)\}$, where

$$\begin{aligned}
\mathbb{V}\{\phi_{h_{xj}}(x - X_j)\} &= \int \{\phi_{h_{xj}}(x - y)\}^2 \{g_\mu * \phi_{\sigma_j}\}(y) dy - \left\{\int \phi_{h_{xj}}(x - y)\{g_\mu * \phi_{\sigma_j}\}(y) dy\right\}^2 \\
&= \frac{1}{h_x \sigma_j^2} \int \phi^2(z) g_\mu * \phi(x + h_x \sigma_j z) dz - \left\{\frac{1}{\sigma_j} \int \phi(z) g_\mu * \phi_{\sigma_j}(x + h_x \sigma_j z) dz\right\}^2 \\
&= \frac{1}{h_x \sigma_j^2} \left\{\int \phi^2(z) dz\right\} f_{\sigma_j}(x)\{1 + o(1)\} - \left\{\frac{1}{\sigma_j} f_{\sigma_j}(x)\right\}^2 \{1 + o(1)\} \\
&= O(h_x^{-1}).
\end{aligned}$$

Next we shall show that

$$\mathbb{E}_{\boldsymbol{\sigma}^2}\left\{\sum_{j=1}^n w_j^2\right\} = O\left(n^{-1}h_\sigma^{-1}\right). \tag{5.18}$$

Observe that $\phi_{h_\sigma}(\sigma_j - \sigma) = O(h_\sigma^{-1})$ for all j. Therefore we have

$$\sum_{j=1}^n \phi_{h_\sigma}^2(\sigma_j - \sigma) = O(h_\sigma^{-1})\sum_{j=1}^n \phi_{h_\sigma}(\sigma_j - \sigma),$$

which further implies that

$$\sum_{j=1}^n w_j^2 = \frac{\sum_{j=1}^n \phi_{h_\sigma}^2(\sigma_j - \sigma)}{\left\{\sum_{j=1}^n \phi_{h_\sigma}(\sigma_j - \sigma)\right\}^2} = \frac{O(n^{-1}h_\sigma^{-1})}{n^{-1}\sum_{j=1}^n \phi_{h_\sigma}(\sigma_j - \sigma)}.$$

Let $Y_j = \phi_{h_\sigma}(\phi_i - \phi)$ and $\bar{Y} = n^{-1}\sum_{j=1}^n Y_i$. Then $0 \le Y_j \le (\sqrt{2\pi}h_\sigma)^{-1}$ and

$$\mathbb{E}(Y_j) = \{g_\sigma * \phi_{h_\sigma}\}(\sigma) = g_\sigma(\sigma) + O(h_\sigma^2).$$

Let E_1 be the event such that $\bar{Y} < \frac{1}{2}E(\bar{Y})$. We apply Hoeffding's inequality to obtain

$$\begin{aligned}
\mathbb{P}\left\{\bar{Y} < \frac{1}{2}E(\bar{Y})\right\} &\le \mathbb{P}\left\{|\bar{Y} - E(\bar{Y})| \ge \frac{1}{2}E(\bar{Y})\right\} \\
&\le 2\exp\left\{-\frac{2n^2 g_\sigma * \phi_{h_\sigma}(\sigma)}{n(2\pi)^{-1}h_\sigma^{-2}}\right\} \\
&\le 2\exp(Cnh_\sigma^2) = O(n^{-1}).
\end{aligned}$$

Note that $\sum_{j=1}^n w_j^2 \le \sum_{j=1}^n w_j = 1$. We have

$$\begin{aligned}
\mathbb{E}(\sum_{j=1}^n w_j^2) &= \mathbb{E}\left(\sum_{j=1}^n w_j^2 \mathbb{1}_E\right) + \mathbb{E}\left(\sum_{j=1}^n w_j^2 \mathbb{1}_E^C\right) \\
&= O(n^{-1}h_\sigma^{-1}) + O(n^{-1}) \\
&= O(n^{-1}h_\sigma^{-1}),
\end{aligned}$$

proving (5.18). Next, consider the variance decomposition

$$\mathbb{V}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}(S_2) = \mathbb{V}_{\boldsymbol{\sigma}^2}\{\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}(S_2)\} + \mathbb{E}_{\boldsymbol{\sigma}^2}\{\mathbb{V}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}(S_2)\}.$$

The first term is zero, and the second term is given by

$$\mathbb{E}_{\boldsymbol{\sigma}^2}\{\mathbb{V}_{\boldsymbol{X},\boldsymbol{\mu}|\boldsymbol{\sigma}^2}(S_2)\} = O(h_x^{-1})\mathbb{E}\left(\sum_{j=1}^n w_j^2\right) = O(n^{-1}h_\sigma^{-1}h_x^{-1}).$$

We simplify the notation and denote the variance of S_2 by $\mathbb{V}(S_2)$ directly. Therefore $\mathbb{V}(S_2) = O(n^{-\epsilon})$ for some $\epsilon > 0$. Consider the following space $\mathbb{Q}_x = \{x : \tilde{f}_\sigma(x) > n^{-\epsilon'}\}$, where $2\epsilon' < \epsilon$. In the proof of the previous lemmas, we have shown that on $\mathbb{R}_x$,

$$\tilde{f}_\sigma(x) = f_\sigma(x)\{1 + O(n^{-\epsilon})\} + K_n,$$

where K_n is a bounded random variable due to the variability of σ_j^2, and $E_{\boldsymbol{\sigma}^2}(K_n) = O(n^{-\varepsilon})$ for some $\varepsilon > 0$. Next we show it is sufficient to only consider $\mathbb{Q}_x$. To see this, note that

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{\sigma}^2}\left(\int\int_{\mathbb{R}_x\cap\mathbb{Q}_x^C}(\tilde{\delta}-\bar{\delta})^2 f_\sigma(x)dxdG_\sigma(\sigma)\right) \\ = \; & \mathbb{E}_{\boldsymbol{\sigma}^2}\left(\int\int_{\mathbb{R}_x\cap\mathbb{Q}_x^C}(\tilde{\delta}-\bar{\delta})^2\left[\tilde{f}_\sigma(x)\{1+O(n^{-\epsilon})\}+K_n\right]dxdG_\sigma(\sigma)\right) \\ = \; & O(C_n'^3)\left\{O(n^{-\epsilon'})+O(n^{-\epsilon'-\epsilon})+O(n^{-1/2})\right\}, \end{aligned}$$

which is also $O(n^{-\eta})$ for some $\eta > 0$. Let

$$Y_j = w_j\phi_{h_{xj}}(x-X_j) - w_j\left\{g_\mu * \phi_{\nu\sigma_j}\right\}(x)$$

and $\bar{Y} = n^{-1}\sum_{j=1}^n Y_j$. Then $\mathbb{E}(Y_j) = 0$, $S_2 = \sum_{j=1}^n Y_j$, and $0 \le Y_j \le D_n$, where $D_n \sim h_x^{-1}$. Let E_2 be the event such that $S_2 < -\frac{1}{2}\tilde{f}_\sigma(x)$. Then by applying Hoeffding's inequality,

$$\mathbb{P}(E_2) \le \mathbb{P}\left\{|\bar{Y}-E(\bar{Y})| \ge \frac{1}{2}\tilde{f}_\sigma(x)\right\} \le 2\exp\left\{-\frac{2n^2\{\frac{1}{2}\tilde{f}_\sigma(x)\}^2}{nD_n^2}\right\} = O(n^{-\epsilon})$$

for some $\epsilon > 0$. Note that on event E_2, we have

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left\{(\hat{\delta}-\tilde{\delta})^2\mathbb{1}_{E_2}\right\} = O(C_n^2)O(n^{-\epsilon}) = o(1).$$

Therefore, we only need to focus on the event E_2^C, on which we have $\tilde{f}_\sigma(x) + S_2 \ge \frac{1}{2}\tilde{f}_\sigma(x)$. It follows that on E_2^C, we have $\{S_2/(\tilde{f}_\sigma(x)+S_2)\}^2 \le 4S_2^2/\{\tilde{f}_\sigma(x)\}^2$. Therefore the first term on the right of (5.16) can be controlled as

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(\mathbb{1}_{E_2^C}\cdot\int\int_{\mathbb{R}_x\cap\mathbb{Q}_x}\left\{\frac{\tilde{f}_\sigma^{(1)}(x)}{\tilde{f}_\sigma(x)}\right\}^2\left\{\frac{S_2}{S_2+\tilde{f}_\sigma(x)}\right\}^2 f_\sigma(x)dxdG_\sigma(\sigma)\right) \\ = \; & O(C_n'^2)O(n^{-(\epsilon-2\epsilon')}) = O(n^{-\eta}) \end{aligned}$$

for some $\eta > 0$. Hence we show that the first term of (5.16) is vanishingly small.

For the second term in (5.16), we need to evaluate the variance term of S_1, which can be similarly shown to be of order $O(n^{-\eta})$ for some $\eta > 0$. Following similar arguments, we can prove that the expectation of the second term in (5.16) is also vanishingly small, establishing the desired result.

5.7. 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

$$r\left(\hat{\boldsymbol{\delta}}^-, G\right) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(n^{-1}\left\|\hat{\boldsymbol{\delta}}^- - \boldsymbol{\mu}\right\|_2^2\right) = n^{-1}\sum_{i=1}^n r_i\left(\hat{\delta}^{-i}, G\right). \tag{5.19}$$

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 (5.19) is equal to the univariate risk $r(\hat{\delta}, G)$ defined by (3.5), and thereby converting the compound risk of a vector decision to the risk of a scalar decision: $r\left(\hat{\boldsymbol{\delta}}^-, G\right) = r(\hat{\delta}, G)$.

5.8. 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

$$S_i = \sum_{j=1}^{n} \phi_{h_\sigma}(\sigma_i - \sigma_j), \quad S_i^- = \sum_{j \neq i} \phi_{h_\sigma}(\sigma_i - \sigma_j).$$

Some algebra shows the following relationships:

$$\hat{f}_\sigma(x_i) = \frac{S_i^-}{S_i}\hat{f}^-_{\sigma_i}(x_i) + \frac{1}{2S_i\pi h_\sigma h_x \sigma_i}, \quad \hat{f}^{(1)}_{\sigma_i}(x_i) = \frac{S_i^-}{S_i}\hat{f}^{(1),-}_{\sigma_i}(x_i). \tag{5.20}$$

The full and jackknifed NEST estimators are respectively given by:

$$\hat{\delta}_i = x_i + \sigma_i^2 \frac{\hat{f}^{(1)}_{\sigma_i}(x_i)}{\hat{f}_{\sigma_i}(x_i)}, \quad \hat{\delta}_i^- = x_i + \sigma_i^2 \frac{\hat{f}^{(1),-}_{\sigma_i}(x_i)}{\hat{f}^-_{\sigma_i}(x_i)}. \tag{5.21}$$

Then according to (5.20) and (5.21), 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 similarly as done in Section 5.3:

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

Let $Y_{ij} = \phi_{h_\sigma}(\sigma_i - \sigma_j)\phi_{h_{xj}}(x_i - x_j)$, and $\bar{Y}_i$ as 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 other hand,

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

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 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.

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.

Berger, J. O. (1976, 01). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist. 4*(1), 223–226.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics 42*(3), 855–903.

Brown, L. D. (2008). In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics 2*(1), 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, 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, T. and W. Sun (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*(1), 197–223.

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.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist. 32*, 962–994.

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

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. 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.

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.

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. (2015). Gaussian estimation: Sequence and wavelet models. *Draft version*.

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.

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. *Big and Complex Data Analysis: Methodologies and Applications (Springer International Publishing)*, 249–284.

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

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

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

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics. 35*, 1–20.

Rogosa, D. (2003). Accuracy of api index and school base report elements: 2003 academic performance index, california department of education. Report.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press.

Stein, C. M. (1981, 11). Estimation of the mean of a multivariate normal distribution. *Ann. Statist. 9*(6), 1135–1151.

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.

Wand, M. P. and M. C. Jones (1994). *Kernel Smoothing*, Volume 60 of *Chapman and Hall CRC Monographs on Statistics and Applied Probability*. Chapman and Hall CRC.

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2017). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association*. To appear.

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.

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