Statistica Sinica Preprint No: SS-2019-0075	
Title	Simultaneous estimation of normal means with side information
Manuscript ID	SS-2019-0075
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0075
Complete List of Authors	Sihai Dave Zhao
Corresponding Author	Sihai Dave Zhao
E-mail	dave.zhao@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica

Simultaneous estimation of normal means with side information

Sihai Dave Zhao

Department of Statistics, University of Illinois at Urbana-Champaign

Abstract: The integrative analysis of multiple datasets is an important strategy in data analysis. It is increasingly popular in genomics, which enjoys a wealth of publicly available datasets that can be compared, contrasted, and combined in order to extract novel scientific insights. This paper studies a stylized example of data integration for a classical statistical problem: leveraging side information to estimate a vector of normal means. This task is formulated as a compound decision problem, an oracle integrative decision rule is derived, and a data-driven estimate of this rule based on minimizing an unbiased estimate of its risk is proposed. The data-driven rule is shown to asymptotically achieve the minimum possible risk among all separable decision rules, and it can outperform existing methods in numerical properties. The proposed procedure leads naturally to an integrative high-dimensional classification procedure, which is illustrated by combining data from two independent gene expression profiling studies.

Key words and phrases: Compound decision problem, Data integration, Gaussian sequence problem, Integrative genomics, Nonparametric empirical Bayes.

1. Introduction

Methods for the integrative analysis of multiple datasets are becoming increasingly important. This is especially true in genetics and genomics, where petabytes of public data are readily available for integrative analysis (Richardson et al., 2016; Ritchie et al., 2015). For example, Pickrell et al. (2016) analyzed summary statistics from genome-wide association studies of 42 human traits and found that multiple traits were influenced by several hundred common genetic variants. In a cross-species example, Shpigler et al. (2017) combined results from a honey bee gene expression study with a database of autism-associated genetic variants and found evidence for evolutionary conservation of genes associated with both honey bee sociality and human autism spectrum disorder. Comparing and contrasting existing data, or combining them with new data, can lead to novel insights that would have been difficult or impossible to uncover with a single dataset alone (Tseng et al., 2015).

Integrative analysis strategies can take many forms, and one particularly common implementation is to leverage side information from one or several auxiliary studies for the purpose of improving the analysis of some primary dataset of interest. Examples abound in the multiple testing literature, where methods such as p-value weighting and false discovery

rate regression incorporate auxiliary information to improve the power to detect true signals in a primary dataset (Genovese et al., 2006; Ramdas et al., 2017). In the genomic risk prediction literature, Hu et al. (2017) and Zhao (2017) showed that summary statistics from previously conducted genome-wide association studies can be used to improve the performance of polygenic risk scores.

Growing interest in these ideas gives rise to an important statistical question: what is the best way to leverage side information? This paper studies this question in a simple but nontrivial problem: the simultaneous estimation of a vector of normal means. The classical version of this problem considers a sequence of independent $X_{i1} \sim N(\theta_{i1}, \sigma_1^2)$ for $i = 1, \ldots, n$ with known σ_1^2, where the goal is to estimate the θ_{i1} (Johnstone, 2017). The integrative version, studied here, investigates how a auxiliary sequence of Gaussian random variables can be used to improve estimation of the means θ_{i1} of the primary Gaussian sequence.

This classical Gaussian sequence model is simplistic, but studying data integration in this setting is nevertheless instructive. First, the model is still important for many applications (Cai, 2012; Johnstone, 2017). Second, more accurate estimation of the mean vector has immediate implications for high-dimensional classification in genomics (Greenshtein and Park, 2009),

which will be demonstrated in Section 6. Finally, this simple problem can reveal general statistical phenomena that arise in integrative data analysis. More complicated variations of the Gaussian sequence model have been studied, for example involving unknown variances that differ across different indices i; see Section 2.2. Extensions of the present work to these more realistic settings are important directions for future work.

Section 2 formalizes this integrative estimation task as a compound decision problem and summarizes previous related work. The optimal way to leverage side information is derived in Section 3, which presents an oracle integrative decision rule that achieves the best risk within a certain class of estimators. This section also introduces a regularized version of the oracle rule that has the same asymptotic risk. A data-driven estimate of this regularized oracle rule is introduced in Section 4, and is shown to asymptotically achieve the optimal risk. Its good performance is illustrated in simulations in Section 5 and in two genomic risk prediction problems in Section 6. A discussion is presented in Section 7 and additional simulations and proofs can be found in the Supplementary Materials.

2. Normal means problem with side information

2.1 Problem statement

As in the classical Gaussian sequence problem, consider a sequence of independent $X_{i1} \sim N(\theta_{i1}, \sigma_1^2)$ for $i = 1, \ldots, n$, with σ_1^2 known. The side information problem studied in this paper further supposes that a second sequence of independent $X_{i2} \sim N(\theta_{i2}, \sigma_2^2), i = 1, \ldots, n$ is available, with σ_2^2 known. The goal is to estimate the θ_{i1}, just as in the classical problem, but here both the X_{i1} and the X_{i2} can be used for estimation. In this sense, the X_{i1} play the role of a primary dataset, and the X_{i2} act as side information from an auxiliary dataset. This paper assumes that the X_{i1} are independent of the X_{i2} for each i, though extensions to dependent X_{i1} and X_{i2} are discussed in Section 7.

This formulation is motivated by applications in integrative genomics. The indices i represent different genomic features, such as different genes, and the X_{i1} and X_{i2} represent different measurements on feature i from different studies. For example, in the genomics classification problem described in Section 6, each X_{i1} estimates a classifier parameter θ_{i1} corresponding to the ith gene from a primary study of interest, and each X_{i2} is the Z-score for the ith gene reported by an auxiliary study of a related

phenotype. The goal is to improve classification accuracy in the primary study by leveraging both X_{i1} and X_{i2} to better estimate the θ_{i1}.

In the above example, the X_{i1} and X_{i2} are paired for each i, as both correspond to the same genomic feature. The informativeness of this pairing is crucial for the good performance of data integration. For example, because the phenotypes considered by the two studies in Section 6 are related, genes with significant Z-scores in the auxiliary study are also likely to be important features for classification in the primary study, so combining the studies is likely to be fruitful. In contrast, if the phenotypes were unrelated, X_{i2} would likely not be informative about θ_{i1}. The challenge is to develop an estimation procedure that can make optimal use of X_{i2}, incorporating them when appropriate and discarding them otherwise. This is addressed by the method proposed in this paper.

To more formally state the problem, define $\mathbf{X}_{\cdot d} = (X_{1d}, \ldots, X_{nd})$, $\boldsymbol{\theta}_{\cdot d} = (\theta_{1d}, \ldots, \theta_{nd})$ for $d = 1, 2$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\cdot 1}, \boldsymbol{\theta}_{\cdot 2})$. Then the normal means problem with side information is to find a decision rule $\boldsymbol{\delta}(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}) = \{\delta_1(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}), \ldots, \delta_n(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2})\} : \mathbb{R}^{2n} \to \mathbb{R}^n$ that minimizes the risk function

$$R_n(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^{n} E[\{\theta_{i1} - \delta_i(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2})\}^2] \tag{2.1}$$

over some class of decision rules. An important class, namely the class

of separable estimators, will be considered in this paper and is discussed in Section 3. This paper adopts the frequentist framework where the $\boldsymbol{\theta}_{\cdot d}$ are fixed nonrandom constants. The auxiliary data are thus statistically independent of the primary data, and it is interesting that they can still provide useful information for estimating $\boldsymbol{\theta}_{\cdot 1}$.

To illustrate the complexities of this problem, first suppose that it were known that $\theta_{i2} = \theta_{i1}$ for all $i = 1, \ldots, n$ and that $\sigma_1 = \sigma_2$. The best way to integrate the auxiliary dataset would clearly be to apply existing optimal estimation methods for a single Gaussian sequence to the sequence of averaged observations $(X_{i1} + X_{i2})/2$. Next consider a slightly more complicated setting: $\theta_{i2} = \theta_{i1}$ for all but one i, but the i for which $\theta_{i2} \neq \theta_{i1}$ is unknown. The auxiliary sequence is clearly still informative for estimating the $\boldsymbol{\theta}_{\cdot 1}$, but how it should be used is no longer obvious. Finally, consider an even more complicated scenario where $\theta_{i2} = h(\theta_{i1}) + e_i$ for some unknown function $h(t)$, where the e_i are unknown perturbations that exhibit no patterns with respect to θ_{i1}. If the magnitudes of the e_i are small relative to the θ_{i1}, $\mathbf{X}_{\cdot 2}$ should still be useful when estimating $\boldsymbol{\theta}_{\cdot 1}$, but it is even less clear how to optimally integrate it into the estimation procedure. This paper provides one approach.

2.2 Previous work

The classical normal means estimation problem without side information, which aims to minimize the risk function (2.1) using decision rules that can depend only on $\mathbf{X}_{\cdot 1}$ and not $\mathbf{X}_{\cdot 2}$, has inspired an enormous literature (Johnstone, 2017). Stein (1956) found that the maximum likelihood estimator $\delta_i(\mathbf{X}_{\cdot 1}) = X_{i1}$ is inadmissible, and since then research has focused on finding alternative estimators with better risk properties. Several different but intimately related perspectives on this problem have been developed.

The shrinkage perspective is exemplified by the James-Stein estimator (James and Stein, 1961; Stigler, 1990), which estimates θ_{i1} by scaling X_{i1} towards zero. The empirical Bayes perspective (Robbins, 1964) treats the θ_{i1} as random draws from a prior distribution, uses the X_{i1} to estimate any unknown parameters in the prior, then estimates each θ_{i1} by its posterior mean conditional on X_{i1}. Efron and Morris (1973) showed that the James-Stein estimator is an empirical Bayes estimator assuming a normal prior for the θ_{i1}. The compound decision perspective (Robbins, 1951; Zhang, 1997) treats the θ_{i1} as nonrandom constants and directly derives the decision rule that minimizes the risk. Under certain conditions, the optimal solution from this perspective is closely related to nonparametric empirical Bayes estimators (Brown and Greenshtein, 2009; Jiang et al., 2009; Zhang, 2003).

More complicated versions of the classical normal means problem have also been intensely studied. For example, specialized methods have been developed for estimating sparse normal means, where most of the θ_{i1} are assumed to equal zero (Castillo et al., 2012; Donoho and Johnstone, 1994, 1995; Martin et al., 2014). Heteroscedastic normal sequences, where the X_{i1} can have different variances for different indices i, have also been considered, both when the variances are known (Fu et al., 2019; Tan, 2016; Weinstein et al., 2018; Xie et al., 2012; Zhang and Bhattacharya, 2017) and when they are unknown but estimates are available (Feng and Dicker, 2018; Gu and Koenker, 2017; Jing et al., 2016).

So far, however, most work on the normal means problem and its variants has considered only a single sequence of observations X_{i1}, and it appears that the side information problem (2.1) has not yet been widely studied. Jiang et al. (2010), Cohen et al. (2013), Tan (2016), and Kou and Yang (2017) proposed methods that can integrate X_{i2}, but these essentially require knowledge of the nature of the relationship between θ_{i1} and X_{i2}, and may not work well when this relationship is misspecified. Banerjee et al. (2018) studied the side information problem, but only for sparse $\boldsymbol{\theta}_{\cdot 1}$. Very recently Saha and Guntuboyina (2017) and Koudstaal and Yao (2018) considered two or more Gaussian sequences, but minimized the risk of es-

timating the means of all of the sequences, rather than the means of just one of them as in (2.1). In contrast to existing work, this paper studies the optimal use of $\mathbf{X}_{\cdot 1}$ and $\mathbf{X}_{\cdot 2}$ for estimating possibly non-sparse $\boldsymbol{\theta}_{\cdot 1}$.

3. Oracle integrative separable rules

Without any restrictions, the optimal decision rule is simply $\delta_i(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}) = \theta_{i1}$, which is not useful because the performance of this rule cannot realistically be achieved using the observed data alone. Instead, this paper only considers rules in the class

$$\mathcal{S} = \{\boldsymbol{\delta}(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}) : \delta_i(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}) = f(X_{i1}, X_{i2})\}, \tag{3.2}$$

where f is some fixed real-valued function that is applied to each pair (X_{i1}, X_{i2}) in order to estimate θ_{i1}. In other words, the estimate of θ_{i1} is calculated by applying $f(x_1, x_2)$ to only the ith pair of observations (X_{i1}, X_{i2}), and $f(x_1, x_2)$ cannot vary with i.

Rules in $\mathcal{S}$, called "separable" rules, are appealing because of their simplicity and have been extensively studied (Brown and Greenshtein, 2009; Cai, 2012; Robbins, 1951; Zhang, 2003). The maximum likelihood estimator $\boldsymbol{\delta}(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}) = X_{i1}$ belongs to $\mathcal{S}$, and the James-Stein estimator approximates the optimal separable rule that is linear in X_{i1} (Jiang et al., 2009). The minimum risk among all separable estimators has been shown to be

asymptotically equivalent, in a certain sense, to the minimum achievable risk over the larger class of permutation invariant estimators (Greenshtein and Ritov, 2009).

The following proposition describes the oracle optimal integrative rule in $\mathcal{S}$ for estimating $\boldsymbol{\theta}_{\cdot 1}$, which will perform no worse than any separable rule that relies only on $\mathbf{X}_{\cdot 1}$. It is a direct consequence of the fundamental theorem of compound decision problems (Robbins, 1951; Jiang et al., 2009). Let $\phi(x)$ denote the standard normal density and define

$$\begin{aligned} p(x_1, x_2; t_1, t_2) =& \frac{1}{\sigma_1}\phi\left(\frac{x_1 - t_1}{\sigma_1}\right)\frac{1}{\sigma_2}\phi\left(\frac{x_2 - t_2}{\sigma_2}\right), \\ p_i^0(x_1, x_2) =& p(x_1, x_2; \theta_{i1}, \theta_{i2}), \end{aligned} \tag{3.3}$$

so that the density of (X_{i1}, X_{i2}) can be abbreviated by $p_i^0(x_1, x_2)$. As mentioned in the problem statement in Section 2.1, this paper assumes that the X_{i1} and X_{i2} are independent, but the following result is easily extended to settings where X_{i1} and X_{i2} are correlated; see Section 7.

Proposition 1. *Define the decision rule* $\boldsymbol{\delta}^\star = (\delta_1^\star, \ldots, \delta_n^\star)$ *where* $\delta_i^\star(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}) = f^\star(X_{i1}, X_{i2})$ *and*

$$f^\star(x_1, x_2) = \frac{\sum_{j=1}^n \theta_{j1} p_j^0(x_1, x_2)}{\sum_{j=1}^n p_j^0(x_1, x_2)}. \tag{3.4}$$

Then $R_n(\boldsymbol{\theta}, \boldsymbol{\delta}) \geq R_n(\boldsymbol{\theta}, \boldsymbol{\delta}^\star)$ *for any* $\boldsymbol{\delta} \in \mathcal{S}$ (3.2) *for* $R_n(\boldsymbol{\theta}, \boldsymbol{\delta})$ *in* (2.1).

The oracle rule $\boldsymbol{\delta}^\star$ also has a useful interpretation as a Bayes rule. If the θ_{i1} are viewed as independent draws from the discrete prior distribution

$$G_n(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} I(\theta_{i1} \le t_1, \theta_{i2} \le t_2), \tag{3.5}$$

then the posterior expectation $E(\theta_{i1} \mid X_{i1}, X_{i2})$ of θ_{i1} is exactly equal to (3.4). This is an example of the close connection between compound decision problems and nonparametric empirical Bayes procedures. The dependence between θ_{i1} and θ_{i2} under G_n quantifies the amount of information that can be borrowed from X_{i2}.

While appealing, this Bayesian interpretation is not necessary for Proposition 1, which holds for fixed and constant $\boldsymbol{\theta}_{\cdot 1}$ and $\boldsymbol{\theta}_{\cdot 2}$. Interestingly, under this frequentist setting Proposition 1 shows that $\mathbf{X}_{\cdot 2}$ can improve the estimation of $\boldsymbol{\theta}_{\cdot 1}$ even though $\mathbf{X}_{\cdot 1}$ and $\mathbf{X}_{\cdot 2}$ are statistically independent, as long the sequences $\boldsymbol{\theta}_{\cdot 1}$ and $\boldsymbol{\theta}_{\cdot 2}$ are related in some sense. There need not be an obvious functional relationship between the two mean vectors.

The above view of side information is slightly different from that of existing frameworks. Previous methods (Jiang et al., 2010; Kou and Yang, 2017; Tan, 2016) posit some functional relationship, typically linear, between θ_{i1} and the observed X_{i2}, rather than between θ_{i1} and the true mean θ_{i2}. For example, Kou and Yang (2017) assume that $\theta_{i1} = h(X_{i2}) + e_i$ for

some error term e_i, where $h(x)$ must be known up to a finite-dimensional parameter. These methods treat the X_{i2} as fixed, while the proposed framework acknowledges that the X_{i2} are random variables. The difference between existing work and the present setting is akin to the difference between classical regression methods and those that take into account covariate measurement error.

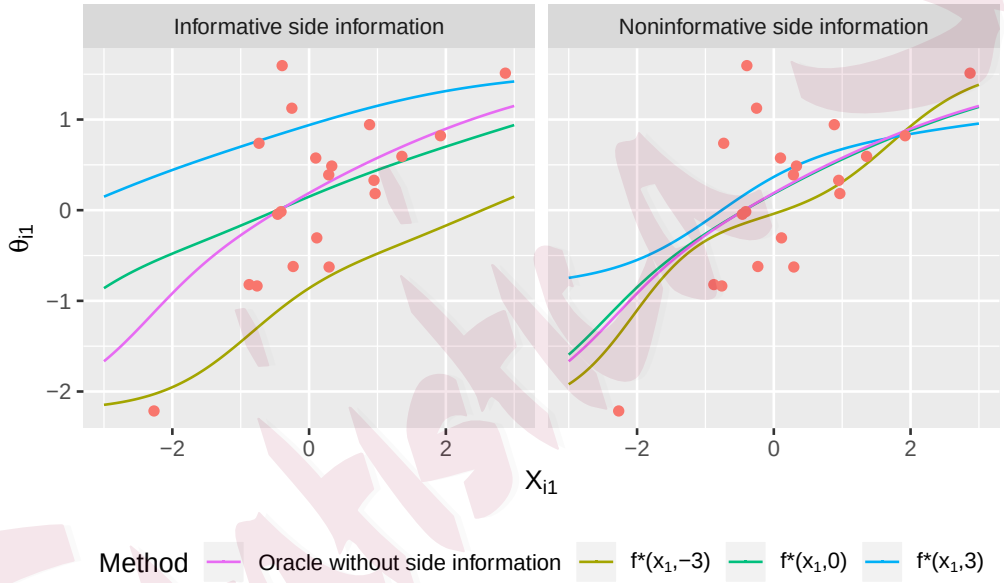

Figure 1: Oracle estimators with and without side information for $n = 20$ pairs (X_{i1}, X_{i2}). Each curve plots the estimate of θ_{i1} as a function of X_{i1}. Each dot corresponds to an observed X_{i1} along with its true mean θ_{i1}.

Figure 1 illustrates the oracle rule $\boldsymbol{\delta}^\star$ (3.4) and compares it to the best

separable estimator that does not use $\mathbf{X}_{\cdot 2}$, which is the posterior expectation of θ_{i1} under the prior G_n conditional only on X_{i1} (Zhang, 2003). In both panels, θ_{i1} was generated by drawing $n = 20$ values from a standard normal distribution. In the left panel, $\boldsymbol{\theta}_{\cdot 2} = \boldsymbol{\theta}_{\cdot 1}$, so $\mathbf{X}_{\cdot 2}$ was highly informative for $\boldsymbol{\theta}_{\cdot 1}$. Thus $f^\star(x_1, 3)$ gave the best estimates of θ_{i1} for large X_{i1} and $f^\star(x_1, -3)$ was most accurate for small X_{i1}. In the right panel of Figure 1, the θ_{i2} were generated from an independent standard normal so that $\mathbf{X}_{\cdot 2}$ was completely uninformative. In this non-informative setting, $\boldsymbol{\delta}^\star$ may not have the same performance as the optimal non-integrative separable rule for any given set of $\mathbf{X}_{\cdot 1}$ and $\mathbf{X}_{\cdot 2}$, but in expectation Proposition 1 guarantees that it will have equal or lower risk.

The oracle separable integrative rule $\boldsymbol{\delta}^\star$ described in (3.4) cannot be implemented in practice because it requires knowing the true $(\theta_{i1}, \theta_{i2})$ up to permutation of the indices. Section 4 will introduce a data-driven rule that targets the performance of $\boldsymbol{\delta}^\star$, though for technical reasons it will be more convenient to target a regularized version of the oracle rule. This will be denoted by $\boldsymbol{\delta}^\star_\rho = (\delta^\star_{\rho 1}, \ldots, \delta^\star_{\rho n})$, with $\delta^\star_{\rho i}(\mathbf{X}_{\cdot 1}, \mathbf{X}_{\cdot 2}) = f^\star_\rho(X_{i1}, X_{i2})$ for

$$f^\star_\rho(x_1, x_2) = x_1 + \frac{\sum_{j=1}^n (\theta_{j1} - x_1) p_j^0(x_1, x_2)}{\rho + \sum_{j=1}^n p_j^0(x_1, x_2)} \tag{3.6}$$

and ρ a small positive constant that prevents the denominator from being too close to zero. Under some assumptions, $\boldsymbol{\delta}^\star_\rho$ will have the same asymp-

totic risk as the oracle $\boldsymbol{\delta}^\star$.

Assumption 1. *There exist positive constants C and η such that $|\theta_{id}| \leq Cn^{1/4-\eta}$ for $i = 1, \ldots, n$ and $d = 1, 2$.*

Theorem 1. *Under Assumption 1,* $\lim_{n\to\infty}\{R_n(\boldsymbol{\theta}, \boldsymbol{\delta}^\star_\rho) - R_n(\boldsymbol{\theta}, \boldsymbol{\delta}^\star)\} = 0$.

Assumption 1 determines how quickly the magnitudes of θ_{id} can grow. To put this rate into perspective, if the θ_{id} were random draws from a normal distribution, then $\max_i |\theta_{id}|$ would be $O(\log^{1/2} n)$ almost surely. Related assumptions, which essentially restrict how variable the θ_{id} can be, have been made in previous work on normal means estimation without side information. For example, Xie et al. (2012) require $\lim n^{-1}\sum_i \theta_{i1}^2 < \infty$, and Jiang et al. (2009) and Zhang (2009) control the rate of the p-th weak moment of the distribution function $n^{-1}\sum_i I(\theta_{i1} \leq t_1)$.

4. Data-driven separable estimator

4.1 Existing nonparametric empirical Bayes approach

By Proposition 1 and Theorem 1, the regularized oracle $\boldsymbol{\delta}^\star_\rho$ (3.6) is asymptotically optimal within the class of separable estimators (3.2), but it cannot be implemented in practice. It therefore remains to develop a fully data-driven estimator for the θ_{i1}. Two classes of approaches already exist. They

have been termed f- and g-modeling (Efron, 2014, 2019) and are based on nonparametric empirical Bayes principles that pretend that the $(\theta_{i1}, \theta_{i2})$ are random variables with prior distribution $G_n(t_1, t_2)$ (3.5).

In f-modeling, the oracle estimator (3.4) would be re-expressed as

$$f^\star(x_1, x_2) = x_1 + \frac{\tilde{p}'(x_1, x_2)}{\tilde{p}(x_1, x_2)},$$

where $\tilde{p}'(x_1, x_2) = \partial\tilde{p}/\partial x_1$ and

$$\tilde{p}(x_1, x_2) = \int p(x_1, x_2; t_1, t_2) dG_n(t_1, t_2)$$

with $p(x_1, x_2; t_1, t_2)$ from (3.3). If the $(\theta_{i1}, \theta_{i2})$ were truly random, $\tilde{p}(x_1, x_2)$ could be interpreted as the marginal density of (X_{i1}, X_{i2}), and $\tilde{p}(x_1, x_2)$ and $\tilde{p}'(x_1, x_2)$ could be estimated nonparametrically using kernel density estimators. In g-modeling, the oracle estimator would be re-expressed as

$$f^\star(x_1, x_2) = x_1 + \frac{\int (t_1 - x_1) p(x_1, x_2; t_1, t_2) dG_n(t_1, t_2)}{\int p(x_1, x_2; t_1, t_2) dG_n(t_1, t_2)},$$

and if the $(\theta_{i1}, \theta_{i2})$ were truly random, a nonparametric estimate of $G_n(t_1, t_2)$ could be obtained by maximizing the marginal log-likelihood (Kiefer and Wolfowitz, 1956)

$$\arg\max_G \prod_{i=1}^{n} \int p(X_{i1}, X_{i2}; t_1, t_2) dG(t_1, t_2).$$

Both f- and g-modeling have been used in normal means problems without side information, where they are asymptotically optimal even in

the frequentist framework where the θ_{i1} and θ_{i2} are nonrandom (Brown and Greenshtein, 2009; Feng and Dicker, 2018; Fu et al., 2019; Jiang et al., 2009; Koenker, 2014; Koenker and Mizera, 2014; Saha and Guntuboyina, 2017; Zhang, 2009). However, neither approach directly estimates the oracle decision rule, with f-modeling proceeding through the intermediate quantity $\tilde{p}(x_1, x_2)$ and g-modeling proceeding through $G_n(t_1, t_2)$.

4.2 Proposed direct risk minimization approach

This paper explores a more direct approach to estimating the oracle integrative separable classifier. Motivated by the regularized oracle (3.6), consider separable rules of the form $\boldsymbol{\delta}^t_\rho = (\delta^t_{\rho 1}, \ldots, \delta^t_{\rho n})$ with

$$\delta^{\mathbf{t}}_{\rho i}(x_1, x_2) = x_1 + \frac{\sum_{j=1}^n (t_{j1} - x_1) p(x_1, x_2; t_{j1}, t_{j2})}{\rho + \sum_{j=1}^n p(x_1, x_2; t_{j1}, t_{j2})}, \tag{4.7}$$

for a given $\mathbf{t} = (t_{11}, \ldots, t_{n1}, t_{12}, \ldots, t_{n2})$. By Theorem 1, the optimal $\mathbf{t}$ equals $(\theta_{11}, \ldots, \theta_{n1}, \theta_{12}, \ldots, \theta_{n2})$, but the θ_{id} are not known. The challenge is to choose a $\mathbf{t}$ in a data-driven fashion that can still asymptotically achieve the optimal risk.

Choosing $\mathbf{t}$ to minimize the risk (2.1) of $\boldsymbol{\delta}^{\mathbf{t}}_\rho$ (4.7) should give an estimator with good performance. However, calculating the risk requires knowing the true θ_{id}. On the other hand, Stein's Lemma (Stein, 1981) can be used to

give an unbiased estimate of the true risk as a function only of $\mathbf{t}$:

$$\begin{aligned}&\textsc{sure}(\mathbf{t})\\ =&\frac{2}{n}\sum_{i=1}^{n}\frac{\sum_j(t_{j1}-X_{i1})^2p(X_{i1},X_{i2};t_{j1},t_{j2})}{\rho+\sum_j p(X_{i1},X_{i2};t_{j1},t_{j2})}-\frac{2}{n}\sigma_1^2\sum_{i=1}^{n}\frac{\sum_j p(X_{i1},X_{i2};t_{j1},t_{j2})}{\rho+\sum_j p(X_{i1},X_{i2};t_{j1},t_{j2})}-\\ &\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_j(t_{1j}-X_{i1})p(X_{i1},X_{i2};t_{j1},t_{j2})}{\rho+\sum_j p(X_{i1},X_{i2};t_{j1},t_{j2})}\right\}^2+\sigma_1^2.\end{aligned} \tag{4.8}$$

The following theorem shows that $\textsc{sure}(\mathbf{t})$ is also a good approximation to the actual loss

$$\ell_n(\mathbf{t})=\frac{1}{n}\sum_{i=1}^{n}\{\theta_{i1}-\delta^{\mathbf{t}}_{\rho i}(X_{i1},X_{i2})\}^2 \tag{4.9}$$

uniformly over the set

$$\mathcal{T}=\{\mathbf{t}:|t_{jd}|\le Cn^{1/4-\eta},j=1,\ldots,n,d=1,2\}. \tag{4.10}$$

Theorem 2. *Under Assumption 1, if* $0<\rho\le 1$, *then*

$$\lim_{n\to\infty}E\sup_{\mathbf{t}\in\mathcal{T}}|\textsc{sure}(\mathbf{t})-\ell_n(\mathbf{t})|=0.$$

The tuning parameter $\mathbf{t}$ can now be chosen by minimizing this estimated risk, as a proxy for minimizing the unknown true risk. The proposed estimator is therefore defined to be

$$\boldsymbol{\delta}^{\hat{\mathbf{t}}}_{\rho}\text{ as in (4.7) with }\hat{\mathbf{t}}=\arg\min_{\mathbf{t}\in\mathcal{T}}\textsc{sure}(\mathbf{t}). \tag{4.11}$$

This strategy of direct risk minimization is common in the compound decision literature (Jing et al., 2016; Kou and Yang, 2017; Tan, 2016; Weinstein

et al., 2018; Xie et al., 2012, 2016; Zhang and Bhattacharya, 2017), but has not yet been used to approximate an optimal separable rule like (3.6). The following theorem shows that (4.11) can asymptotically achieve the same the performance as the optimal separable decision rule.

Theorem 3. *Under the same conditions as Theorem 2,* $\lim_{n\to\infty}\{E\ell_n(\hat{\mathbf{t}}) - R_n(\boldsymbol{\theta}, \boldsymbol{\delta}^\star)\} \leq 0$*, where* $E\ell_n(\hat{\mathbf{t}})$ *is the risk of the proposed estimator* $\boldsymbol{\delta}_\rho^{\hat{\mathbf{t}}}$ (4.11).

4.3 Implementation

The proposed estimator has been implemented in the R package `cole`, available at `github.com/sdzhao/cole`. In practice, the exact value of ρ appears to make little difference, and $\rho = 0$ works well in most cases. When the range of the X_{id} is very large or the variances σ_d^2 are very small, problem may arise when calculating $\textsc{sure}(\mathbf{t})$ due to numerical precision, in which case setting $\rho = 10^{-12}$ seems to be sufficient. Throughout this paper, the proposed method was implemented with $\rho = 0$.

Because the value of $Cn^{1/4-\eta}$ that defines the feasible set $\mathcal{T}$ (4.10) is not known, in practice the minimization in (4.11) can be performed over

$$\hat{\mathcal{T}} = \prod_{i=1}^{n}[X_{i1} - M\sigma_1, X_{i1} + M\sigma_1] \times [X_{i2} - M\sigma_2, X_{i2} + M\sigma_2]$$

for some sufficiently large positive constant M, so that $\hat{\mathcal{T}}$ contains $(\boldsymbol{\theta}_{\cdot 1}, \boldsymbol{\theta}_{\cdot 2})$ with probability $\Phi(-M)^n$, where Φ is the cumulative distribution function

of a standard normal. By default, `cole` uses $M = 5$, so that $\hat{\mathcal{T}}$ contains $(\boldsymbol{\theta}_{\cdot 1}, \boldsymbol{\theta}_{\cdot 2})$ with probability 0.99 when $n = 10,000$. Optimizing $\text{SURE}(\mathbf{t})$ over $\hat{\mathcal{T}}$ is sensible because it is known from Theorem 1 that $E\{\text{SURE}(\mathbf{t})\}$ achieves a global minimum at $t_{jd} = \theta_{jd}$. This method works well, but bridging the gap between the theoretical procedure and its practical implementation is an important direction for future work.

Minimizing $\text{SURE}(\mathbf{t})$ is difficult because it is a nonconvex function. The implementation in `cole` performs a simple coordinate descent. At initialization, t_{id} is set to X_{id}, and at each iteration one t_{id} is updated by optimizing over K equally spaced candidates in $[X_{id} - M\sigma_d, X_{id} + M\sigma_d]$. By default, `cole` uses $K = 10$, and all analyses in this paper use $K = 10$ unless otherwise stated. The coordinates of $\mathbf{t}$ are updated in the order $t_{11}, t_{21}, \ldots, t_{n1}, t_{12}, \ldots, t_{n2}$, and convergence is reached when all of the coordinates have been cycled through once without changing the value of $\text{SURE}(\mathbf{t})$ by more than a small ϵ, which `cole` sets to 10^{-5} by default.

5. Simulations

5.1 Normal means problem without side information

The direct risk minimization approach proposed in this paper for estimating optimal separable decision rules appears to be novel in the compound

decision literature. This section thus first illustrates how this idea performs in the classical normal means problem without side information. The optimal separable estimator and its corresponding unbiased risk estimate will look like (3.4) and (4.8), respectively, but with the density $p(x_1, x_2; t_1, t_2)$ replaced by $\phi\{(x_1 - t_1)/\sigma_1\}/\sigma_1$, where $\phi(x)$ is the standard normal density. Similar to (4.11), a data-driven estimator of the oracle rule can be obtained by minimizing the risk estimate over $\mathbf{t}_1$ using the coordinate descent algorithm described in Section 4.3; this is available in the `cole` package. Analogs of Theorems 1–3 can also be proved.

The direct estimator was compared to the g-modeling procedure of Jiang et al. (2009), which is also asymptotically risk-optimal. One independent sequence $X_{i1}, i = 1, \ldots, 1,000$ was generated from $\mathrm{N}(\theta_{i1}, 1)$, with the goal of estimating $\boldsymbol{\theta}_{\cdot 1}$ using only $\mathbf{X}_{\cdot 1}$. The θ_{i1} equaled either 0 or μ and the number of nonzero θ_{i1} equaled either 5, 50, or 500. Table 1 displays the average total squared errors over 100 replications. Results for the estimator of Jiang et al. (2009) were taken directly from their Table 1, while the proposed estimator was implemented using a coordinate descent algorithm that optimized over $K = 50$ candidates for each t_{1j}. The results show that both estimators had almost identical performance.

# nonzero	5				50				500			
μ	3	4	5	7	3	4	5	7	3	4	5	7
GMLEB	39	34	23	11	157	105	58	14	459	285	139	18
Proposed	37	32	21	11	158	110	56	14	460	289	133	21

Table 1: Average total squared errors for the classical normal means problem without side information. GMLEB: the g-modeling method of Jiang et al. (2009).

5.2 Settings for normal means problem with side information

The primary data $\mathbf{X}_{\cdot 1} = (X_{11}, \ldots, X_{n1})$ were generated in four different ways, for three dense and one sparse configuration of their means $\boldsymbol{\theta}_{\cdot 1}$. To generate dense $\boldsymbol{\theta}_{\cdot 1}$, values of θ_{i1} were independently drawn from either a $\mathrm{N}(0,1)$, a $\mathrm{Unif}(-2,2)$, or an $\mathrm{Exp}(1)$ distribution. To generate the sparse configuration, 10% of the coordinates of $\boldsymbol{\theta}_{\cdot 1}$ were set to 1.5 and the rest were set to 0. The observed primary data were generated as $X_{i1} = \theta_{i1} + \epsilon_{i1}$, where the ϵ_{i1} were independently drawn from standard normal random variables. The θ_{i1} were fixed across all replications.

For each of these four settings, the auxiliary data $\mathbf{X}_{\cdot 2} = (X_{12}, \ldots, X_{n2})$ were generated in three different ways, to model different degrees of informativeness of $\boldsymbol{\theta}_{\cdot 2}$. First define e_i to be independent draws from a $\mathrm{Unif}(-4,4)$.

To generate strongly, weakly, and non-informative side information, θ_{i2} was set to be either $2\theta_{i1}^2$, $\theta_{i1}^2 + e_i$, or e_i, respectively. The observed auxiliary data were generated as $X_{i2} = \theta_{i2} + \epsilon_{i2}$, where the ϵ_i were again independently drawn from standard normal random variables. The θ_{i2} were fixed across all replications.

The proposed integrative normal means estimator (4.11) was compared to two existing approaches that can incorporate side information. One was the procedure of Banerjee et al. (2018). The other was estimator (1) of Kou and Yang (2017), defined as

$$\frac{\lambda}{\lambda + \sigma_1^2} X_{i1} + \frac{\sigma_1^2}{\lambda + \sigma_1^2} h(X_{i2})$$

for some function $h(x)$ known up to a finite number of parameters. These unknown parameters, as well as λ, are chosen by minimizing an unbiased estimate of the risk of this estimator. This estimator is motivated by the regression model $\theta_{i1} = h(X_{i2}) + e_i$ for some error terms e_i. However, it can be difficult to choose the correct regression function $h(x)$. For example, in some of the present simulation settings, the true relationship between the primary and auxiliary data is $\theta_{i2} = 2\theta_{i1}^2 + e_i$, which is difficult to translate into a regression model of θ_{i1} on X_{i2}. When implementing the method of Kou and Yang (2017), these simulations used both the nonlinear model $\theta_{i1} = \beta_0 + \beta_1 |X_{i2}|^{1/2} + e_i$ and the linear model $\theta_{i1} = \beta_0 + \beta_1 X_{i2} + e_i$.

Finally, two additional estimators for $\boldsymbol{\theta}_{\cdot 1}$ were also implemented to provide performance baselines. The first was the oracle (3.4), which attains the lowest possible risk of any separable decision rule that incorporates side information. The second was the g-modeling method of Jiang et al. (2009), which can asymptotically achieve the optimal risk of any separable rule that does not use side information.

5.3 Results for normal means problem with side information

Figure 2 illustrates the average losses, over 200 simulations, achieved by the competing methods for $\mathrm{N}(0,1)$, $\mathrm{Unif}(-2,2)$, $\mathrm{Exp}(1)$, or sparse $\boldsymbol{\theta}_{\cdot 1}$ and non-informative, weakly informative, or strongly informative $\boldsymbol{\theta}_{\cdot 2}$. Comparing the performances of the oracle rule (3.4) and the method of Jiang et al. (2009) shows that including auxiliary data does not degrade estimation accuracy asymptotically when $\boldsymbol{\theta}_{\cdot 2}$ is non-informative, and can greatly improve it when $\boldsymbol{\theta}_{\cdot 2}$ is informative.

The performance of the proposed data-driven estimator $\hat{\boldsymbol{\delta}}^{\mathbf{t}}_{\rho}$ (4.11) indeed appeared to converge to the oracle performance as the number of observations n increased, consistent with Theorem 3. Unlike the oracle, however, incorporating non-informative $\mathbf{X}_{\cdot 2}$ in $\hat{\boldsymbol{\delta}}^{\mathbf{t}}_{\rho}$ resulted in worse performance compared to the other methods when n was small. This is expected,

as non-informative X_{i2} add extra noise without decreasing bias, and the data-driven method requires enough samples to learn that the X_{i2} are not useful. In contrast, $\hat{\boldsymbol{\delta}}^{\mathbf{t}}_{\rho}$ regained its competitiveness for larger n, and when the auxiliary $\mathbf{X}_{\cdot 2}$ were at least weakly informative, it frequently achieved the lowest risk among all methods. These results suggest that incorporating $\mathbf{X}_{\cdot 2}$ using the proposed method is highly effective when $\mathbf{X}_{\cdot 2}$ is informative, and does not do too much harm when it is not.

The proposed $\hat{\boldsymbol{\delta}}^{\mathbf{t}}_{\rho}$ was sometimes outperformed by the two different implementations of the procedure of Kou and Yang (2017), for example when the θ_{i1} were generated from Exp(1). This may be because this setting was particularly difficult for the proposed method. Out of the four configurations of $\boldsymbol{\theta}_{\cdot 1}$, the maximum value of $|\theta_{i1}|$ was largest under the Exp(1) configuration, and Assumption 1 makes it clear that restricting this maximum value is important for the good performance of $\hat{\boldsymbol{\delta}}^{\mathbf{t}}_{\rho}$. On the other hand, when $n = 1,000$ $\hat{\boldsymbol{\delta}}^{\mathbf{t}}_{\rho}$ had essentially the same risk as the methods of Kou and Yang (2017), and for other configurations of θ_{i1}, $\hat{\boldsymbol{\delta}}^{\mathbf{t}}_{\rho}$ could perform significantly better.

Finally, the proposed rule performed extremely well with sparse $\boldsymbol{\theta}_{\cdot 1}$, even though it was not designed for this scenario. When the auxiliary data were strongly informative, it achieved the lowest risks among all im-

plemented methods when $n \geq 200$. It would be interesting to explore extensions of the proposed procedure to estimate sparse normal means.

6. Data analysis

High-dimensional classification is an important problem in genomics. Shi et al. (2010) studied the effectiveness of using gene expression microarray data to develop classification rules for various phenotypes. This section focuses on classification of two of these phenotypes: estrogen receptor status and treatment response status in breast cancer patients. The training and validation datasets they used are publicly available from the Gene Expression Omnibus (Edgar et al., 2002) under accession number GSE20194.

Integrating auxiliary data may help improve classification accuracy. Wang et al. (2005) developed a gene expression signature for distant metastasis-free survival in estrogen receptor-positive and -negative breast cancer patients. It may be possible to leverage data from Wang et al. (2005), publicly available under accession number GSE2034, to more accurately classify the two outcomes from Shi et al. (2010). However, it is not clear how to best integrate these auxiliary data.

The normal means estimation problem using side information, studied in this paper, provides one approach. Greenshtein and Park (2009) showed

that minimizing the squared error risk in the normal means problem is closely connected to minimizing the misclassification rate in high dimensional classification. Let $\bar{G}_i^Y$ denote the average expression level of gene i in across all training subjects in class $Y = 0, 1$ and $\hat{s}_i^Y$ denote its estimated standard deviation. Greenshtein and Park (2009) considered classifying an observed gene expression vector $(G_1, \ldots, G_n)$ using

$$I\left(\sum_{i=1}^{n} \hat{\theta}_i G_i / \hat{s}_i \geq c\right) \tag{6.12}$$

for some cutoff c, where $\hat{s}_i = \{(\hat{s}_i^1)^2/n_1 + (\hat{s}_i^0)^2/n_0\}^{1/2}$ and $\hat{\theta}_i$ is an estimate of the expected value of $Z_i = (\bar{G}_i^1 - \bar{G}_i^0)/\hat{s}_i$. They showed that using the f-modeling procedure of Brown and Greenshtein (2009) to obtain $\hat{\theta}_i$ can lead to more accurate classification compared to simply using $\hat{\theta}_i = Z_i$.

Combined with ideas in this paper, this framework leads to a natural integrative classifier. Let X_{i1} equal Z_i calculated for either estrogen receptor status or treatment response status from the Shi et al. (2010) study, and let X_{i2} be the differential expression Z-score of the ith gene with respect to either estrogen receptor status or distant metastasis-free survival from the Wang et al. (2005) study. Integrating X_{i2} into the estimate $\hat{\theta}_{i1}$ should lead to more accurate classification.

This integrative classification was implemented using the proposed rule $\hat{\boldsymbol{\delta}}_\rho^{\mathbf{t}}$ (4.11), the method of Kou and Yang (2017) using a model linear in X_{i2},

and the procedure of Banerjee et al. (2018) for sparse normal means. These were compared to five classifiers that do not make use of auxiliary information: 1) the method of Greenshtein and Park (2009) but implemented using the g-modeling procedure of Jiang et al. (2009), 2) the naive Bayes classifier, 3) logistic lasso using the R package `glmnet` (Friedman et al., 2010), 4) random forest using the R package `ranger` (Wright and Ziegler, 2017), and 5) the regularized optimal affine discriminant analysis of Fan et al. (2012) using the R package `TULIP` (Pan et al., 2019). Tuning parameters for lasso and the method of Fan et al. (2012) were chosen using 10-fold cross-validation while random forest was run using default parameters.

The integrative, naive Bayes, and Greenshtein and Park (2009) classifiers all assume that the X_{id} are independent across i. For these procedures, screening was thus first performed to ensure that the magnitude of the correlation between every pair of genes in the training data was small, similar to what was done in Dicker and Zhao (2016). Specifically, genes were sorted from most to least significantly associated with the outcome in the training data, with p-values calculated using the R package `limma` (Smyth, 2005). Starting from the most significant gene, any other gene with correlation greater than 0.2 in magnitude was removed from the dataset. No screening was performed for lasso, random forest, or the method of Fan et al. (2012).

Misclassification rates for estrogen receptor and treatment response status were assessed using the same training and testing datasets used in Shi et al. (2010), and classification was also repeated after swapping the roles of the training and testing data. The averages of the two resulting misclassification rates for the different methods are displayed in Figure 3.

The results suggest that integrative classification can be a useful strategy. Intuitively, the survival results from Wang et al. (2005) should be most informative for predicting treatment response, while the ER status data from Wang et al. (2005) should be most useful for predicting ER status. Indeed, the proposed integrative classifier using survival Z-scores to predict treatment response gave the lowest misclassification rate among all methods. The proposed method integrating ER status Z-scores to predict ER status performed better than every method except random forest and lasso.

7. Discussion

This paper assumes that the primary data $\mathbf{X}_{\cdot 1}$ and the auxiliary data $\mathbf{X}_{\cdot 2}$ are statistically independent. However, in some practical settings X_{i1} and X_{i2} may be correlated for each i, for example if $\mathbf{X}_{\cdot 1}$ and $\mathbf{X}_{\cdot 2}$ arise from case-control studies with shared controls (Zaykin and Kozbur, 2010). The ideas

proposed in this paper can be naturally extended to this correlated setting. Assuming (X_{i1}, X_{i2}) were bivariate normal with a known correlation, the oracle integrative rule would be similar to (3.4) and is given in (S1.1) in the Supplementary Materials. An asymptotically risk-optimal data-driven estimator could then be constructed by minimizing an unbiased risk estimate derived using Stein's lemma.

As pointed out by one referee, this setting is especially interesting because when X_{i1} and X_{i2} are correlated, the $\mathbf{X}_{\cdot 2}$ provides useful information for estimating $\boldsymbol{\theta}_{\cdot 1}$ even when $\boldsymbol{\theta}_{\cdot 2}$ and $\boldsymbol{\theta}_{\cdot 1}$ are completely unrelated. This is not true when X_{i1} and X_{i2} are independent. This is verified by Figure 1 in the Supplementary Materials, where the oracle integrative rule performed better than the oracle non-integrative rule when $|cor(X_{i1}, X_{i2})| = 0.9$ even though $\boldsymbol{\theta}_{\cdot 2}$ was generated to be non-informative for $\boldsymbol{\theta}_{\cdot 1}$. Thus rules such as (3.4) can therefore take full advantage of information about $\boldsymbol{\theta}_{\cdot 1}$ contained the auxiliary $\mathbf{X}_{\cdot 2}$, whether that information comes in the form of informative θ_{i2}, correlated X_{i2}, or both.

This paper considered only a single sequence of auxiliary data, but it is straightforward to extend the proposed procedure to multiple auxiliary sequences. However, this would result in theoretical and computational difficulties, for given $D - 1$ auxiliary datasets, Assumption 1 would re-

quire $|\theta_{id}| \leq n^{1/(2D)-\eta}$ for $d = 1, \ldots, D$, and the proposed procedure would require optimizing over Dn parameters. It would be of great interest to study whether there exists a convex surrogate of the unbiased risk estimate (4.8). An alternative approach might be to use parametric or semiparametric methods, such as those proposed by Kou and Yang (2017), but to endow them with data-driven model selection capabilities.

It would be interesting to extend data integration ideas to other variants of the classical normal means problem, such as heteroscedastic sequences, sparse sequences, and non-normal observed data. It would also be interesting to consider broader applications of the compound decision framework beyond the simultaneous estimation of a mean vector, such as the integrative high-dimensional classification problem in Section 6.

Though this paper studied the highly stylized problem of normal means estimation with side information, its results reveal several general principles of integrative analysis. First, auxiliary data can be useful even if they are statistically independent of, and have no clearly expressible functional relationship with, primary data. The two datasets need only be related in the sense discussed in Section 3. Second, in principle, integrating auxiliary data can only help and not harm the primary analysis. This is because it is possible to learn from the data themselves the degree to which the auxiliary

data are informative, and thus the degree to which they should influence inference in the primary data. Third, nonparametric methods, such as the proposed (4.11), can asymptotically achieve ideal performance.

Supplementary Materials

The supplementary materials contain simulation results when the primary and auxiliary data are correlated, as well as proofs of the theoretical results.

Acknowledgements

The author thanks Professors Jiaying Gu, Gourab Mukherjee, and Roger Koenker for helpful discussions. This research was supported in part by NSF grant DMS-1613005.

References

Banerjee, T., G. Mukherjee, and W. Sun (2018). Adaptive sparse estimation with side information. Technical report, University of Southern California.

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*, 1685–1704.

Cai, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation.

Statistical Science, 31–50.

Castillo, I., A. van der Vaart, et al. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics 40*(4), 2069–2101.

Cohen, N., E. Greenshtein, and Y. Ritov (2013). Empirical bayes in the presence of explanatory variables. *Statistica Sinica*, 333–357.

Dicker, L. H. and S. D. Zhao (2016). High-dimensional classification via nonparametric empirical bayes and maximum likelihood inference. *Biometrika*, 21–34.

Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association 90*(432), 1200–1224.

Donoho, D. L. and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika 81*(3), 425–455.

Edgar, R., M. Domrachev, and A. E. Lash (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research 30*(1), 207–210.

Efron, B. (2014). Two modeling strategies for empirical bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics 29*(2), 285.

Efron, B. (2019). Bayes, Oracle Bayes, and Empirical Bayes. *Statistical Science*. to appear.

Efron, B. and C. Morris (1973). Stein's estimation rule and its competitorsan empirical bayes approach. *Journal of the American Statistical Association 68*(341), 117–130.

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: the

regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 745–771.

Feng, L. and L. H. Dicker (2018). Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. *Computational Statistics & Data Analysis*.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

Fu, L., W. Sun, and G. M. James (2019). Nonparametric empirical bayes estimation on heterogeneous data. Technical report, University of Southern California.

Genovese, C. R., K. Roeder, and L. Wasserman (2006). False discovery control with p-value weighting. *Biometrika 93*(3), 509–524.

Greenshtein, E. and J. Park (2009). Application of non parametric empirical Bayes estimation to high dimensional classification. *J. Mach. Learn. Res. 10*, 1687–1704.

Greenshtein, E. and Y. Ritov (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality: The Third Erich L. Lehmann Symposium*, pp. 266–275. Institute of Mathematical Statistics.

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

Hu, Y., Q. Lu, W. Liu, Y. Zhang, M. Li, and H. Zhao (2017). Joint modeling of genetically cor-

related diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS genetics 13*(6), e1006836.

James, W. and C. M. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 367–379. Berkeley and Los Angeles, University of California Press.

Jiang, W., C.-H. Zhang, et al. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics 37*(4), 1647–1684.

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

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

Johnstone, I. M. (2017). Gaussian estimation: Sequence and wavelet models. Technical report, Department of Statistics, Stanford University, Stanford.

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

Koenker, R. (2014). A gaussian compound decision bakeoff. *Stat 3*(1), 12–16.

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. and J. J. Yang (2017). Optimal shrinkage estimation in heteroscedastic hierarchical linear models. In *Big and Complex Data Analysis*, pp. 249–284. Springer.

Koudstaal, M. and F. Yao (2018). From multiple gaussian sequences to functional data and beyond: a stein estimation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(2), 319–342.

Martin, R., S. G. Walker, et al. (2014). Asymptotically minimax empirical bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics 8*(2), 2188–2206.

Pan, Y., Q. Mai, and X. Zhang (2019). Tulip: A toolbox for linear discriminant analysis with penalties. *arXiv preprint arXiv:1904.03469*.

Pickrell, J. K., T. Berisa, J. Z. Liu, L. Ségurel, J. Y. Tung, and D. A. Hinds (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics 48*(7), 709.

Ramdas, A., R. F. Barber, M. J. Wainwright, and M. I. Jordan (2017). A unified treatment of multiple testing with prior knowledge. *arXiv preprint arXiv:1703.06222*.

Richardson, S., G. C. Tseng, and W. Sun (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and its Application 3*, 181–209.

Ritchie, M. D., E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim (2015). Methods of inte-

grating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics 16*(2), 85.

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

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

Saha, S. and A. Guntuboyina (2017). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *arXiv preprint arXiv:1712.02009*.

Shi, L., G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T.-M. Chu, F. M. Goodsaid, L. Pusztai, et al. (2010). The microarray quality control (maqc)-ii study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology 28*(8), 827.

Shpigler, H. Y., M. C. Saul, F. Corona, L. Block, A. C. Ahmed, S. D. Zhao, and G. E. Robinson (2017). Deep evolutionary conservation of autism-related genes. *Proceedings of the National Academy of Sciences 114*(36), 9653–9658.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420. Springer.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal

distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 197–206. Berkeley and Los Angeles, University of California Press.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*, 1135–1151.

Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 147–155.

Tan, Z. (2016). Steinized empirical bayes estimation for heteroscedastic data. *Statistica Sinica*, 1219–1248.

Tseng, G. C., D. Ghosh, and X. J. Zhou (2015). *Integrating Omics Data*. Cambridge University Press.

Wang, Y., J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet 365*(9460), 671–679.

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

Wright, M. N. and A. Ziegler (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software 77*(1), 1–17.

REFERENCES

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.

Xie, X., S. C. Kou, and L. Brown (2016). Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *Annals of statistics 44*(2), 564.

Zaykin, D. V. and D. O. Kozbur (2010). P-value based analysis for shared controls design in genome-wide association studies. *Genetic epidemiology 34*(7), 725–738.

Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statistica Sinica 7*(1), 181–193.

Zhang, C.-H. (2003). Compound decision theory and empirical bayes methods. *The Annals of Statistics 31*(2), 379–390.

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica 19*, 1297–1318.

Zhang, X. and A. Bhattacharya (2017). Empirical bayes, sure and sparse normal mean models. *arXiv preprint arXiv:1702.05195*.

Zhao, S. D. (2017). Integrative genetic risk prediction using non-parametric empirical bayes classification. *Biometrics 73*(2), 582–592.

Department of Statistics, University of Illinois at Urbana-Champaign

E-mail: sdzhao@illinois.edu

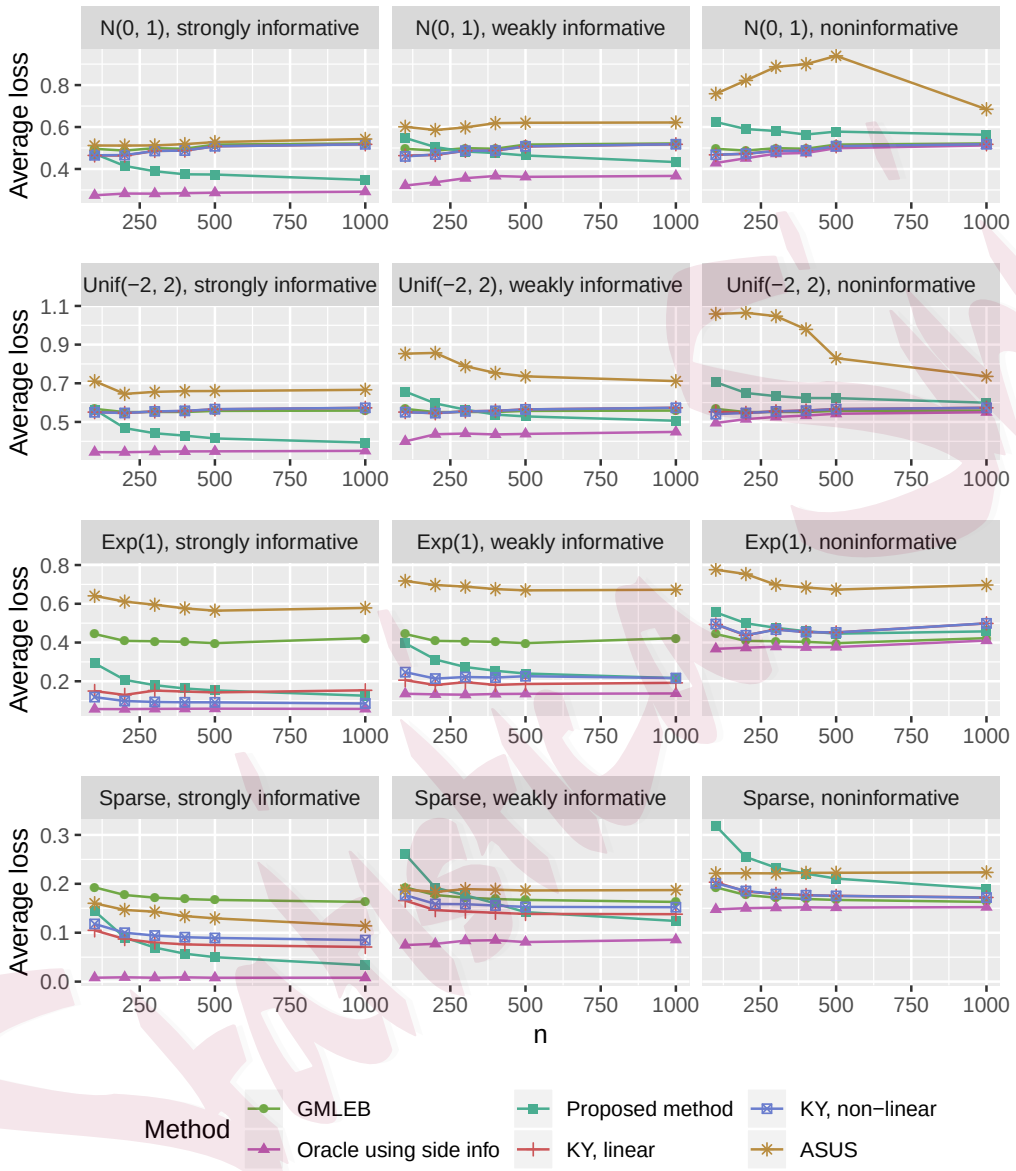

Figure 2: Average losses for four different configurations of $\boldsymbol{\theta}_{\cdot 1}$ and three degrees of informativeness of $\boldsymbol{\theta}_{\cdot 2}$. GMLEB: method of Jiang et al. (2009); KY, linear: method of Kou and Yang (2017) with model $\theta_{i1} = \beta_0 + \beta_1 X_{i2} + e_i$; KY, nonlinear: method of Kou and Yang (2017) with model $\theta_{i1} = \beta_0 + \beta_1 |X_{i2}|^{1/2} + e_i$; ASUS: method of Banerjee et al. (2018).

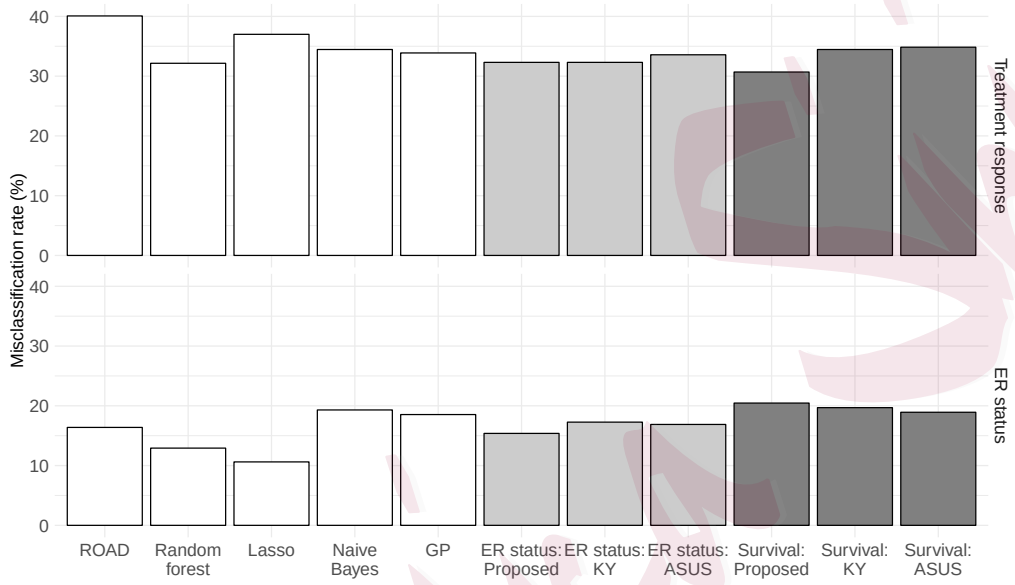

Figure 3: Average misclassification errors for treatment response status or estrogen receptor (ER) status from Shi et al. (2010). GP: method of Greenshtein and Park (2009); KY: method of Kou and Yang (2017); ASUS: method of Banerjee et al. (2018); ROAD: method of Fan et al. (2012). "+ ER status/survival": using differential expression with respect to either ER status or distant metastasis-free survival from Wang et al. (2005) as auxiliary data.