Statistica Sinica Preprint No: SS-2024-0063	
Title	Empirical Bayes Estimation with Side Information: A Nonparametric Integrative Tweedie Approach
Manuscript ID	SS-2024-0063
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0063
Complete List of Authors	Jiajun Luo, Trambak Banerjee, Gourab Mukherjee and Wenguang Sun
Corresponding Authors	Trambak Banerjee
E-mails	trambak@ku.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

EMPIRICAL BAYES ESTIMATION WITH SIDE INFORMATION: A NONPARAMETRIC INTEGRATIVE TWEEDIE APPROACH

Jiajun Luo[1], Trambak Banerjee[2], Gourab Mukherjee[1] and Wenguang Sun[3]

University of Southern California[1], University of Kansas[2] and Zhejiang University[3]

Abstract:

We investigate the problem of compound estimation of normal means while accounting for the presence of side information. Leveraging the empirical Bayes framework, we develop a nonparametric integrative Tweedie (NIT) approach that incorporates structural knowledge encoded in multivariate auxiliary data to enhance the precision of compound estimation. Our approach employs convex optimization tools to estimate the gradient of the log-density directly, enabling the incorporation of structural constraints. We conduct theoretical analyses of the asymptotic risk of NIT and establish the rate at which NIT converges to the oracle estimator. As the dimension of the auxiliary data increases, we accurately quantify the improvements in estimation risk and the associated deterioration in convergence rate. The numerical performance of NIT is illustrated through the analysis of both simulated and real data, demonstrating its superiority over existing methods.

Key words and phrases: Compound Decision Problem, Convex Optimization, Kernelized Stein's Discrepancy, Side Information, Tweedie's Formula.

1. Introduction

In data-intensive fields, such as genomics, neuroimaging, and signal processing, vast amounts of data are collected, often accompanied by various types of side information. We consider a compound estimation problem where $\boldsymbol{Y} = (Y_i : 1 \leq i \leq n)$ is a vector of summary statistics and serves as the primary data for analysis. In addition, we collect K auxiliary sequences

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

$\boldsymbol{S}^{(k)} = (S_i^{(k)} : 1 \le i \le n)$, $1 \le k \le K$, alongside the primary data. Suppose the elements in $\boldsymbol{Y}$ follow normal distributions

$$Y_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad 1 \le i \le n, \tag{1.1}$$

where $\theta_i = \mathbb{E}(Y_i)$ represents the true underlying effect size for the ith study unit. Following Brown and Greenshtein (2009); Efron (2011); Ignatiadis et al. (2023) we assume that σ^2 is known or can be well estimated from the data. For instance, in practical applications we often observe replicates for some of the observations using which σ^2 can be consistently estimated. Moreover, if we are in a rapid trend changing environments where the variances are stationary then σ^2 can be estimated from past data (see, for instance, Section 2.4 of Banerjee et al. (2021)). Let $\boldsymbol{S}_i = (S_i^1, \cdots, S_i^K)^T$ denote the side information associated with unit i and $\boldsymbol{S} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_n)^T$ the auxiliary data matrix. Assume that $\boldsymbol{S}_i$ follow some unspecified multivariate distribution F_S. Our task is to estimate the high-dimensional parameter $\boldsymbol{\theta} = (\theta_i : 1 \le i \le n)$ given both primary and auxiliary data.

Conventional meta-analytical methods frequently encounter two limitations. Firstly, these methods often assign equal importance to the primary and auxiliary data, relying on weighting strategies to calculate an overall effect by integrating data from multiple sources. Nevertheless, this approach may result in biased estimates of θ_i when the distributions of $\boldsymbol{Y}$ and $\boldsymbol{S}^{(k)}$ differ. Secondly, conventional techniques, which are designed to handle a small number of parameters, can become highly inefficient for large-scale estimation problems. This inefficiency is particularly pronounced when $\boldsymbol{\theta}$ is in high dimensions, where valuable structural knowledge of $\boldsymbol{\theta}$ can be extracted from both primary and auxiliary data and exploited to construct more efficient inference procedures.

This article presents an empirical Bayes approach to integrative compound estimation

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

with side information. The framework provides a flexible and powerful tool that can effectively integrate information from multiple sources. Our method capitalizes on the structural knowledge present in auxiliary data, which can be highly informative and has the potential to greatly enhance estimation accuracy when properly assimilated into the decision-making process. In what follows, we begin by presenting an overview of the progress made in this research direction and identify relevant issues. This will be followed by an exposition of our methodology for addressing the challenges. Finally, we discuss related works and highlight our contributions.

1.1 Compound decisions, structural knowledge and side information

Consider a compound decision problem where we make simultaneous inference of n parameters $(\theta_i : 1 \le i \le n)$ based on summary statistics $(Y_i : 1 \le i \le n)$ from n independent experiments. Let $\boldsymbol{\delta} = (\delta_i : 1 \le i \le n)$ be the decision rule, i.e., our estimate of θ_i. Several classical ideas, such as the compound decision theory (Robbins, 1951), empirical Bayes (EB) methods (Robbins, 1964), and James-Stein shrinkage estimator (Stein, 1956), alongside more recent multiple testing methodologies (Efron et al., 2001; Sun and Cai, 2007), have demonstrated that structural information of the data can be leveraged to construct more efficient classification, estimation, and multiple testing procedures. For example, the subminimax rule in Robbins (1951) has shown that the disparity in the proportions of positive and negative signals can be incorporated into inference to reduce the misclassification rate. Similarly, the adaptive z-value procedure in Sun and Cai (2007) has demonstrated that the shape of the alternative distribution can be utilized to construct more powerful false discovery rate (FDR, Benjamini and Hochberg, 1995) procedures.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

When auxiliary data is taken into account, the inference units become heterogeneous. This heterogeneity provides new structural knowledge that can be leveraged to further improve the efficiency of existing methods. For instance, in genomics research, prior data and domain knowledge can be used to define a prioritized subset of genes. Roeder and Wasserman (2009) proposed to up-weight the p-values in prioritized subsets where genes are more likely to be associated with the disease. Structured multiple testing is a crucial area which has garnered considerable attention. A partial list of references, including Lei and Fithian (2018); Cai et al. (2019); Li and Barber (2019); Ignatiadis and Huber (2021); Ren and Candès (2020), demonstrates that the power of existing FDR methods can be substantially improved by utilizing auxiliary data to assign differential weights or to set varied thresholds to corresponding test statistics. Similar ideas have been adopted by some recent works on shrinkage estimation. For instance, Weinstein et al. (2018) and Banerjee et al. (2020) propose to incorporate side information into inference by first creating groups, then constructing group-wise linear shrinkage or soft-thresholding estimators.

1.2 Nonparametric integrative Tweedie

Tweedie's formula is an elegant and celebrated result that has received renewed interests recently (Jiang and Zhang, 2009; Brown and Greenshtein, 2009; Efron, 2011; Koenker and Mizera, 2014; Ignatiadis et al., 2023; Saha and Guntuboyina, 2020; Kim et al., 2022; Zhang et al., 2022; Gu and Koenker, 2023). Under the nonparametric empirical Bayes framework, the formula is particularly appealing for large-scale estimation problems for it is simple to implement, removes selection bias (Efron, 2011) and enjoys frequentist optimality properties (Xie et al., 2012).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

The EB implementation of Tweedie's formula has been extensively studied in the literature. Zhang (1997) demonstrated that a truncated generalized empirical Bayes (GEB) estimator asymptotically achieves both Bayes and minimax risks. Additionally, the nonparametric maximum likelihood estimate (NPMLE, Kiefer and Wolfowitz, 1956) approach and the broader class of g-modeling approaches (Efron, 2016; Shen and Wu, 2022) implement Tweedie's formula by estimating the unknown prior distribution G through the Kiefer-Wolfwitz estimator (Jiang and Zhang, 2009). The NPMLE approach enjoys desirable asymptotic optimality properties in a wide range of problems (Jana et al., 2023; Jiang and Zhang, 2010; Soloff et al., 2021; Polyanskiy and Wu, 2020). In contrast to the NPMLE, Brown and Greenshtein (2009) proposed the f-modeling strategy, which implements Tweedie's formula directly by estimating the marginal density of observations using Gaussian kernels. This nonparametric EB estimator achieves asymptotic optimality in both dense and sparse regimes. Empirically, NPMLE outperforms the kernel method by Brown and Greenshtein (2009). However, the algorithm for NPMLE by Jiang and Zhang (2009) cannot handle data-intensive applications due to its computational complexity. The connection between compound estimation and convex optimization was established by Koenker and Mizera (2014), which casts NPMLE as a convex program, resulting in fast and stable algorithms that outperform competing methods; see Gu and Koenker (2017), Koenker and Gu (2017b) and Saha and Guntuboyina (2020) for recent works in this direction. However, in the context of the g-modeling strategy, direct non-parametric assimilation of covariates has not been thoroughly investigated, and these approaches can exhibit significant computational complexity, particularly when dealing with covariates of moderate to high dimensions.

To effectively extract and incorporate useful structural information from both primary

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

and auxiliary data, we propose a nonparametric integrative Tweedie (NIT) approach to compound estimation of normal means. NIT utilizes the f-modeling strategy, which involves directly estimating the log-gradient of the conditional distribution of Y given $\boldsymbol{S}$, also known as the score function, thereby eliminating the need for a deconvolution estimator for the unknown mixing distribution. We recast compound estimation via NIT as a convex program using a well-designed reproducing kernel Hilbert space (RKHS) representation of Stein's discrepancy. By searching for feasible score embeddings in the RKHS, we obtain the optimal shrinkage factor, resulting in a computationally efficient and scalable algorithm that exhibits superior empirical performance, even in high-dimensional covariate settings. The kernelized optimization framework also provides a rigorous and powerful mathematical interface for theoretical analysis. Leveraging the RKHS theory and concentration theories of V-statistics, we derive the approximate order of the kernel bandwidth, establish the asymptotic optimality of the data-driven NIT procedure, and explicitly characterize the impact of covariate dimension on the rate of convergence.

In recent years, the theoretical foundations of score estimation approaches—particularly in the context of diffusion models used in generative AI for image generation—have garnered significant attention (see, for instance,Wibisono et al. (2024); Zhang et al. (2024); Dou et al. (2024) and the references therein). While these methods are primarily designed for diffusion models, their connections to our approach for estimating the score function are mainly theoretical. In particular, we note that the theoretical convergence rates established in our paper (Section 3) are weaker than those achieved for score estimation in Wibisono et al. (2024); Zhang et al. (2024). However, like these works, we also capture the exponential decay in convergence rates as the dimension K of the side information vector $\boldsymbol{S}_i$ increases

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

(see Theorem 1). Furthermore, both Wibisono et al. (2024) and Zhang et al. (2024) rely on ratio estimators to learn the score function and are similar in spirit to the approach pursued in Brown and Greenshtein (2009). The numerical experiments in Section 4 demonstrate superior performance of our proposed method compared to such ratio estimators across various regimes.

1.3 Our contributions

Methodological contributions. NIT offers several advantages over existing shrinkage estimators. Firstly, it provides a nonparametric framework for assimilating auxiliary data from multiple sources, setting it apart from existing works such as (Ke et al., 2014; Cohen et al., 2013; Kou and Yang, 2017; Ignatiadis and Wager, 2019). Unlike these methods, NIT does not require the specification of any functional relationship, and its asymptotic optimality holds for a wider class of prior distributions. Secondly, NIT has the ability to incorporate various types of side information and effectively handle multivariate covariates. By contrast, Weinstein et al. (2018); Banerjee et al. (2020) only focus on the variance or sparsity structure, and both methods can only deal with univariate covariates by adopting a grouping approach. However, under the multivariate covariate setting, it may be infeasible to determine the optimal number of groups and to search for the ideal grouping structure. Furthermore, grouping involves discretizing a continuous variable, leading to a loss of efficiency. Finally, NIT is a fast, scalable, and flexible tool that can incorporate various structural constraints and produce stable estimates.

Theoretical contributions. Firstly, we establish the convergence rates of NIT to the oracle integrative Tweedie estimator, which explicitly characterizes the improvements in estimation

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

risk by leveraging auxiliary data. Secondly, our theoretical analysis precisely quantifies the deterioration in convergence rates as the dimension of the side information increases, providing important caveats on utilizing high-dimensional auxiliary data. For this theoretical analysis, we introduce new analytical tools that formalize the L_p risk properties of Kernelized Stein Discrepancy (KSD) based estimators. To rigorously prove results for the L_p risk, we establish a local isometry between the L_p risk and the RKHS norm of the proposed estimator. Related KSD-based works (Liu et al., 2016; Chwialkowski et al., 2016; Banerjee et al., 2021) assume the existence of such local isometries without providing a formal analysis. The probability tools developed here can be of independent interest for decision theorists; particularly our techniques have demonstrated their usefulness for analyzing the L_p risk of recently proposed KSD methods, in the context of both heteroskedastic normal means problem (Banerjee et al., 2024) and mixed effects models (Banerjee and Sharma, 2025). For instance, in the context of the heteroskedastic normal means problem, Banerjee et al. (2024) (BFJMS24) consider the following hierarchical model: $Y_i|(\theta_i, \sigma_i^2) \overset{ind.}{\sim} N(\theta_i, \sigma_i^2),\ \theta_i \mid \sigma_i \overset{ind.}{\sim} G_\mu(\cdot|\sigma_i),\ \sigma_i \overset{i.i.d.}{\sim} H_\sigma(\cdot)$, where $G_\mu(\cdot \mid \sigma_i)$ and $H_\sigma(\cdot)$ are unspecified prior distributions, and develop novel techniques to address non-exchangeability in coordinate-wise rules arising from heterogeneous variances. We note that this setting does not fall within the scope of the homoskedastic framework considered in our paper since the models described in equations (1.1) and (2.1) assume a relationship between the location parameter θ_i and the side information $\boldsymbol{S}_i$, but they do not account for relationships involving variances in a heteroskedastic setup. However, the theory in our paper aligns with the theoretical derivations for Bayes-optimal rules in BFJMS 2024, and the convergence rates are related.

The article is organized as follows. In Section 2, we discuss the empirical Bayes esti-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

mation framework, NIT estimator, and computational algorithms. Section 3 delves into the theoretical properties of the NIT estimator. Sections 4 and 5 investigate the performance of NIT using simulated and real data, respectively. Additional technical details, numerical illustrations and proofs are provided in the online Supplementary Material. All R codes for reproducing the numerical experiments conducted in this paper are available at the following GitHub repository: https://github.com/jiajunluo121/NIT.

2. Methodology

Let $\boldsymbol{\delta}(\boldsymbol{y}, \boldsymbol{s}) = (\delta_i : 1 \leq i \leq n)$ be an estimator of $\boldsymbol{\theta}$, and $\mathcal{L}_n^2(\boldsymbol{\delta}, \boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n (\theta_i - \delta_i)^2$ be the corresponding loss function. We define the risk as $\mathbb{R}_n(\boldsymbol{\delta}, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Y}, \boldsymbol{S}|\boldsymbol{\theta}} \{\mathcal{L}_n^2(\boldsymbol{\delta}, \boldsymbol{\theta})\}$ and the Bayes risk as $B_n(\boldsymbol{\delta}) = \int \mathbb{R}_n(\boldsymbol{\delta}, \boldsymbol{\theta}) d\Pi(\boldsymbol{\theta})$, where $\Pi(\boldsymbol{\theta})$ is an unspecified prior distribution for $\boldsymbol{\theta}$.

We assume that the primary and auxiliary data are related through a latent vector $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)^T$ according to the following hierarchical model:

$$\begin{aligned} \theta_i &= g_\theta(\xi_i, \eta_{y,i}), \quad 1 \leq i \leq n, \\ s_i^{(j)} &= g_j^s(\xi_i, \tilde{\eta}_{j,i}), \quad 1 \leq j \leq K, \end{aligned} \tag{2.1}$$

where g_θ and g_j^s are unspecified functions, and $\eta_{y,i}$ and $\tilde{\eta}_{j,i}$ are random perturbations that are independent from $\boldsymbol{\xi}$. This hierarchical model assumes that the shared information between θ_i and $\boldsymbol{S}_i$ is encoded by a common latent variable ξ_i. The relevance of the auxiliary information hinges on the noise level as well as the functional forms of g_θ and g_j^s. Our methodology does not require prior knowledge of g_θ and g_j^s, offering a versatile framework for integrating both continuous and discrete auxiliary data, and accommodates various types of side information ranging from entirely non-informative to perfectly informative. As g_θ and g_j^s are unknown,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

we propose to incorporate covariate information nonparametrically. This section initially presents an oracle rule that optimally utilizes information from $\boldsymbol{S}$, followed by a discussion on a data-driven non-parametric rule designed to mimic the oracle rule. Subsequently, in Section 3.2, we delve into the frequentist risk properties of the proposed methods for a fixed sequence of $\boldsymbol{\theta}$.

Remark 1. Equations (1.1) and (2.1) can also be conceptualized as a Bayesian hierarchical model as follows:

$$Y_i \mid (\theta_i, \boldsymbol{S}_i) \overset{ind.}{\sim} N(\theta_i, \sigma^2),\ (\theta_i, \boldsymbol{S}_i) \mid \xi_i \overset{ind.}{\sim} G_\theta(\cdot \mid \xi_i) G_{\boldsymbol{s}}(\cdot \mid \xi_i),\ \xi_i \overset{i.i.d.}{\sim} G_\xi(\cdot),$$

where $G_\theta, G_{\boldsymbol{s}}$ and G_ξ are unknown distributions. In particular, the above representation includes the hierarchical model of Ignatiadis et al. (2023) (see Equation 1) as a special case where, marginalizing out ξ_i, the conditional distribution of θ_i given $\boldsymbol{S}_i$ is assumed to be Gaussian.

2.1 Oracle integrative Tweedie estimator

We consider an oracle with access to $f(y|\boldsymbol{s})$. The oracle rule that minimizes the Bayes risk among all decision rules with the full data set, comprising both primary and auxiliary data, is referred to as the *integrative Tweedie rule*. In Section 3, we will quantify its improved performance over the Bayes rule relying solely on the primary data. The integrative Tweedie rule is presented in following proposition. It can be derived from Tweedie's formula in (Efron, 2011) and is provided in the supplement for completeness.

Proposition 1 (Integrative Tweedie)**.** *Consider the hierarchical model (1.1) and (2.1). Let* $f(y|\boldsymbol{s})$ *be the conditional density of* Y *given* $\boldsymbol{S}$ *and denote* $\nabla_y \log f(y|\boldsymbol{s}) = \frac{\partial}{\partial y} \log f(y|\boldsymbol{s})$. *The*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

optimal estimator that minimizes the Bayes risk is $\boldsymbol{\delta}^{\pi}(\boldsymbol{y}, \boldsymbol{s}) = \{\delta^{\pi}(y_i, \boldsymbol{s}_i) : 1 \le i \le n\}$, *where*

$$\delta^{\pi}(y, \boldsymbol{s}) = y + \sigma^2 \nabla_y \log f(y|\boldsymbol{s}). \tag{2.2}$$

The integrative Tweedie rule (2.2) provides a versatile framework for integrating primary and auxiliary data. Existing literature on shrinkage estimation with side information typically requires a pre-specified form of the conditional mean function $m(\boldsymbol{S}_i) = E(Y_i|\boldsymbol{S}_i)$ (Ke et al., 2014; Cohen et al., 2013; Kou and Yang, 2017). In contrast, integrative Tweedie incorporates side information through a much wider class of functions $f(y|\boldsymbol{s})$ (where f conditioned on $\boldsymbol{s}$ is a mixture of Gaussian location densities), eliminating the need to pre-specify a fixed relationship between Y and $\boldsymbol{S}$. In Section S1 of the Supplementary Material, we present two toy examples that respectively demonstrate: (a) the reduction of integrative Tweedie to an intuitive data averaging strategy when the distributions of primary and auxiliary variables match perfectly, and (b) the effectiveness of integrative Tweedie in reducing the risk (relative to ignoring $\boldsymbol{S}$) even when the two distributions differ.

2.2 Nonparametric estimation via convex programming

This section proposes a data-driven approach to emulate the oracle rule. Let $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$ denote the set of all data, where the primary sequence is denoted by $x_{1i} = y_i$ and the $(k-1)$th auxiliary sequence is denoted by $x_{ki} = s_i^{(k-1)}$ for $k = 2, \ldots, K+1$ and $i = 1, \ldots, n$. Our objective is to estimate the shrinkage factor, which is given by

$$\begin{aligned}\boldsymbol{h}_f(\boldsymbol{X}) &= \left\{ \nabla_{u_1} \log f(u_1|u_2, \ldots, u_{K+1})\big|_{\boldsymbol{u}=\boldsymbol{x}_i} : 1 \le i \le n \right\} \\ &= \{\nabla_{y_i} \log f(y_i|\boldsymbol{s}_i) : 1 \le i \le n\}.\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

We present a convex program designed to estimate $\boldsymbol{h}_f(\boldsymbol{X})$. This program is motivated by the kernelized Stein's discrepancy (KSD) that we formally define in Section 3. The KSD measures the distance between a given $\boldsymbol{h}$ and the true score $\boldsymbol{h}_f$. It is always non-negative, and is equal to 0 if and only if $\boldsymbol{h} = \boldsymbol{h}_f$. Let $K_\lambda(\boldsymbol{x}, \boldsymbol{x}')$ be a kernel function that is integrally strictly positive definite, where λ is a tuning parameter. A detailed discussion on the construction of kernel $K(\cdot, \cdot)$ and choice of λ is provided in Section 2.3. Consider the following two $n \times n$ matrices:

$$(\boldsymbol{K}_\lambda)_{ij} = n^{-2}\, K_\lambda(\boldsymbol{x}_i, \boldsymbol{x}_j),\ (\nabla \boldsymbol{K}_\lambda)_{ij} = n^{-2}\, \nabla_{x_{1j}} K_\lambda(\boldsymbol{x}_i, \boldsymbol{x}_j)\ .$$

Given a fixed λ, we define $\hat{\boldsymbol{h}}_{\lambda,n}$ as the solution to the following quadratic program:

$$\hat{\boldsymbol{h}}_{\lambda,n} = \underset{\boldsymbol{h} \in \boldsymbol{V}_n}{\arg\min} \quad \boldsymbol{h}^T \boldsymbol{K}_\lambda \boldsymbol{h} + 2\boldsymbol{h}^T \nabla \boldsymbol{K}_\lambda \boldsymbol{1}, \tag{2.3}$$

where $\boldsymbol{V}_n$ is a convex subset of $\mathbb{R}^n$. Convex constraints, such as linearity and monotonicity, can be imposed through $\boldsymbol{V}_n$. Such constraints play an essential role in enhancing the stability and efficiency of compound estimation procedures (Koenker and Mizera, 2014); we provide a detailed discussion on these constraints in Section 2.3. In Section 3, we provide theory to demonstrate that solving the convex program (2.3) is equivalent to finding a shrinkage estimator, aided by side information, that minimizes the estimation risk.

Now combining (2.3) and Proposition 1, we propose the following class of nonparametric integrative Tweedie (NIT) estimators

$$\left\{\boldsymbol{\delta}_\lambda^{\mathrm{NIT}} : \lambda \in (0, \infty)\right\}, \qquad \text{where } \boldsymbol{\delta}_\lambda^{\mathrm{NIT}} = \boldsymbol{y} + \sigma^2\, \hat{\boldsymbol{h}}_{\lambda,n}. \tag{2.4}$$

In Section 3, we show that as $n \to \infty$ there exist choices of λ such that the resulting estimator (2.4) is asymptotically optimal.

The NIT estimator (2.4) marks a clear departure from existing NPMLE methods (Jiang

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

and Zhang, 2009; Koenker and Mizera, 2014; Gu and Koenker, 2017), which cannot be easily extended to handle multivariate auxiliary data. Furthermore, NIT has several additional advantages over existing empirical Bayes methods in both theory and computation. Firstly, in comparison with the NPMLE method (Jiang and Zhang, 2009), the convex program (2.3) is computationally efficient and easily scalable. Secondly, Brown and Greenshtein (2009) proposed estimating the score function using the ratio $\hat{f}^{(1)}/\hat{f}$, where $\hat{f}$ is a kernel density estimate and $\hat{f}^{(1)}$ is its derivative. However, the ratio estimate can be highly unstable. In contrast, our direct optimization approach produces more stable and accurate estimates. Finally, our convex program can be fine-tuned by selecting an appropriate λ, resulting in improved numerical performance and facilitating a disciplined theoretical analysis. The criterion in (2.3) can be rigorously analyzed to establish new rates of convergence (Sec. 3.1) that are previously unknown in the literature.

2.3 Computational details

This section presents several computational details: (a) a discussion on how to impose convex constraints; (b) a description of how to construct kernel functions capable of handling multivariate and potentially correlated covariates; and (c) a strategy on how to select the bandwidth λ.

1. Structural constraints. In the convex program, we impose the constraint $\mathbf{1}^T\boldsymbol{h} = 0$. This constraint ensures the "unbiasedness" of the estimator, as the expectation of the gradient of the log marginal density is theoretically zero. While other convex constraints can be readily integrated into the optimization, they may not be entirely appropriate. For instance, a monotonicity constraint, initially introduced in Koenker and Mizera (2014), has

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

been shown to be highly effective in improving estimation accuracy in sequence models without auxiliary variables. To facilitate ease of presentation, assume $y_1 \leq y_2 \leq \cdots \leq y_n$. The monotonicity constraint, expressed as $\sigma^2 h_{i-1} - \sigma^2 h_i \leq y_i - y_{i-1}$ for all i, can be formulated as $M\boldsymbol{h} \preceq \boldsymbol{a}$ and incorporated into $\boldsymbol{V}_n$ in (2.3) by selecting M as the upper triangular matrix $M_{ij} = \sigma^2\{I(i = j) - I(i = j - 1)\}$ and setting $\boldsymbol{a}^T = (y_2 - y_1, y_3 - y_2, \cdots, y_n - y_{n-1})$. However, the monotonicity constraint though satisfied by the Tweedie estimator for primary data is not satisfied by the integrative Tweedie estimator on the full data set as $y_1 \leq y_2$ does not imply $\mathbb{E}(\theta_1|Y_1, \boldsymbol{S}_1) \leq \mathbb{E}\theta_2|y_2, \boldsymbol{S}_2$. While our paper refrains from imposing such constraints, they can be easily integrated into the methodology if there is a preference for restricting the optimization space in (2.3). These additional constraints, though non-trivial, can significantly decrease the variance of the resulting estimator. Further discussion on this topic will be provided in Section 3.4.

2. Kernel functions. Constructing an appropriate kernel function is crucial when dealing with the complications that arise in the multivariate setting, where the auxiliary sequences may be correlated and have varying measurement units. We propose to use the Mahalanobis distance $\|\boldsymbol{x} - \boldsymbol{x}'\|_{\Sigma_{\boldsymbol{x}}} = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^T \Sigma_{\boldsymbol{x}}^{-1} (\boldsymbol{x} - \boldsymbol{x}')}$ in the kernel function, where $\Sigma_{\boldsymbol{x}}$ denotes the sample covariance matrix. Specifically, we employ the RBF kernel $K_\lambda(\boldsymbol{x}, \boldsymbol{x}') = \exp(-0.5\lambda^2 \|\boldsymbol{x} - \boldsymbol{x}'\|^2_{\Sigma_{\boldsymbol{x}}})$, where λ is the bandwidth parameter that can be selected through cross-validation. Mahalanobis distance is superior to the Euclidean distance as it is unit less, scale-invariant and accounts for the correlations in the data. When the auxiliary data contains both continuous and categorical variables, we propose to use the generalized Mahalanobis distance (Krusińska, 1987). While we illustrate the methodology for mixed types of variables in the numerical studies, we only pursue the theory in the case

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

where both $\boldsymbol{Y}$ and $\boldsymbol{S}$ are continuous.

3. Modified cross-validation (MCV) for selecting λ. As the kernel bandwidth parameter, λ controls the classic bias-variance trade-off in the score function estimate. For instance, a larger value of λ allows unbiasedness but the resulting n dimensional score function estimator has more variance relative to a smaller λ which forces the estimated scores towards 0. Thus, to appropriately regularize the score function estimates, we propose to determine λ via the modified cross-validation (MCV) approach of Brown et al. (2013). Specifically, we introduce $\eta_i \sim \mathcal{N}(0, \sigma^2)$, $1 \leq i \leq n$, as *i.i.d.* noise variables that are independent of $\boldsymbol{y}$ and $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_K$. Define an uncorrelated pair $U_i = y_i + \alpha\eta_i$ and $V_i = y_i - \eta_i/\alpha$, which satisfies $\mathrm{Cov}(U_i, V_i) = 0$. The MCV strategy employs $\boldsymbol{U} = (U_1, \cdots, U_n)$ to construct estimators $\delta_\lambda^{\mathsf{IT}}(\boldsymbol{U}, \boldsymbol{S})$, while validating using $\boldsymbol{V} = (V_1, \cdots, V_n)$. Consider the following validation loss

$$\hat{L}_n(\lambda, \alpha) = \frac{1}{n}\sum_{i=1}^{n} \left\{\hat{\delta}_{\lambda,i}^{\mathsf{IT}}(\boldsymbol{U}, \boldsymbol{S}) - V_i\right\}^2 - \sigma^2(1 + 1/\alpha^2) \ . \tag{2.5}$$

For small α, λ is determined as the value that minimizes $\hat{L}_n(\lambda, \alpha)$, i.e., $\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \hat{L}_n(\lambda, \alpha)$ for any $\Lambda \subset \mathbb{R}^+$. The proposed NIT estimator is given by $\{y_i + \sigma_y^2 \hat{\boldsymbol{h}}_{\hat{\lambda},n}(i) : 1 \leq i \leq n\}$. Proposition 3 in Section 3.6 establishes that the validation loss converges to the true loss, justifying the MCV algorithm.

3. Theory

This section presents a large-sample theory for the data-driven NIT estimator in Equation (2.4) under the setting when σ^2 is known. Our theoretical analysis focuses on the continuous case, where $\boldsymbol{X}_i = (Y_i, S_{i1}, \cdots, S_{ik})^T$, $i = 1, \cdots, n$, are assumed to be i.i.d. samples from a continuous multivariate density f: $\mathbb{R}^{K+1} \to \mathbb{R}^+$. The focus on the continuous case enables an illuminating analysis of the convergence rates (Sections 3.3 and 3.4), providing valuable

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

insights into the role that side information plays in compound estimation.

Let $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T$ denote the $n \times (K+1)$ matrix of observations, with $\boldsymbol{X}^{(k)}$ being the kth column of $\mathbf{X}$. We define f as a density function on $\mathbb{R}^{K+1}$, and $\mathfrak{h}_f(\boldsymbol{x})$ as the first conditional score (FCS) $\nabla_1 \log f(x_1|\boldsymbol{x}_{-1})$, where $\boldsymbol{x}_{-1} = (x_j : 2 \leq j \leq K+1)$. By definition, the FCS is equivalent to the log-gradient of the joint density for the first coordinate, i.e., $\mathfrak{h}_f(\boldsymbol{x}) = \nabla_1 \log f(\boldsymbol{x})$. Recall our primary objective is to estimate the parameter vector $\boldsymbol{\theta} = \mathbb{E}\left\{\boldsymbol{X}^{(1)}\right\}$. As we have demonstrated in Proposition 1, the FCS $\mathfrak{h}_f(\boldsymbol{x})$ plays a crucial role in constructing the oracle estimator; hence accurately estimating the FCS enables us to obtain a precise estimate of $\boldsymbol{\theta}$.

Section 3.1 provides a detailed explanation of the relationship between compound estimation and kernelized Stein's discrepancy, and demonstrates how the FCS $\mathfrak{h}_f(\boldsymbol{x})$ can be accurately approximated by the solution to (2.3). The properties of the scores and convergence rates are established in Sections 3.2 to 3.4, with additional results presented in Sections 3.5 and 3.6.

3.1 Stein's discrepancy, shrinkage estimation and convex optimization

We begin by introducing a conditional version of the kernelized Stein's discrepancy measure, which is widely used to quantify the discrepancy between probability distributions. Specifically, we consider the Gaussian kernel function $K_\lambda : \mathbb{R}^{K+1} \times \mathbb{R}^{K+1} \to \mathbb{R}$ with bandwidth λ. Let P and Q be two distributions on $\mathbb{R}^{K+1}$, with densities denoted by p and q, and associated first conditional score functions denoted by $\mathfrak{h}_p$ and $\mathfrak{h}_q$, respectively. We define the conditional kernelized Stein's discrepancy (KSD; Liu et al. (2016); Chwialkowski et al. (2016); Banerjee et al. (2021)) between P and Q as the kernel-weighted distance between $\mathfrak{h}_p$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

and $\mathfrak{h}_q$, given by:

$$D_\lambda(P,Q) = \mathbb{E}_{(\boldsymbol{u},\boldsymbol{v}) \overset{\text{i.i.d.}}{\sim} P}\Big[\mathcal{K}_\lambda(\boldsymbol{u},\boldsymbol{v}) \times \Big\{\mathfrak{h}_p(\boldsymbol{u}) - \mathfrak{h}_q(\boldsymbol{u})\Big\} \times \Big\{\mathfrak{h}_p(\boldsymbol{v}) - \mathfrak{h}_q(\boldsymbol{v})\Big\}\Big].$$

The versatility and effectiveness of KSD make it a valuable tool for many statistical problems. For example, the minimization of KSD-based criteria has been a popular technique to solve a range of statistical problems, including new goodness-of-fit tests (Liu et al., 2016; Chwialkowski et al., 2016; Yang et al., 2018), Bayesian inference (Liu and Wang, 2016; Oates et al., 2017), and simultaneous estimation (Banerjee et al., 2021).

KSD is closely related to Maximum Mean Discrepancy (MMD, Gretton et al. (2012)); both are measures of discrepancy between probability distributions. While KSD involves the use of kernel functions to construct an unbiased estimate of the Stein operator, MMD uses kernel functions to map the distributions into a reproducing kernel Hilbert space where the distance between their mean embeddings can be computed. Compared to MMD, KSD is particularly well-suited for empirical Bayes estimation because it can be directly constructed based on the score functions, which, as shown in Proposition 1, yield optimal shrinkage factors. Moreover, it can be demonstrated (See Sec. 3 of Jitkrittum et al. (2020)) that the conditional KSD defined above satisfies the following properties:

$$D_\lambda(P,Q) \geq 0 \text{ and } D_\lambda(P,Q) = 0 \text{ if and only if } \int \big|P(u_1|\boldsymbol{u}_{-1}) - Q(u_1|\boldsymbol{u}_{-1})\big|\, p(\boldsymbol{u})\, d\boldsymbol{u} = 0.$$

These properties make KSD an attractive choice for testing and comparing distributions, as well as for other applications in which measuring the discrepancy between probability distributions is crucial.

We leverage the property that a value of 0 for the conditional KSD indicates the equality of the conditional distributions $P_1(\boldsymbol{u}) = P(u_1|\boldsymbol{u}_{-1})$ and $Q_1(\boldsymbol{u}) = Q(u_1|\boldsymbol{u}_{-1})$. How-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

ever, the direct evaluation of $D_\lambda(P, Q)$ is challenging. To overcome this difficulty, we introduce an alternative representation of the KSD, initially introduced in Liu et al. (2016) and Chwialkowski et al. (2016), that can be easily evaluated empirically. Specifically, we define a quadratic functional $\kappa_\lambda[\mathfrak{h}]$ over $\mathbb{R}^{K+1} \times \mathbb{R}^{K+1}$ for any functional $\mathfrak{h} : \mathbb{R}^{K+1} \to \mathbb{R}$, given by:

$$\kappa_\lambda[\mathfrak{h}](\boldsymbol{u}, \boldsymbol{v}) = K_\lambda(\boldsymbol{u}, \boldsymbol{v})\mathfrak{h}(\boldsymbol{u})\mathfrak{h}(\boldsymbol{v}) + \nabla_{\boldsymbol{v}} K_\lambda(\boldsymbol{u}, \boldsymbol{v})\, \mathfrak{h}(\boldsymbol{u}) + \nabla_{\boldsymbol{u}} K_\lambda(\boldsymbol{u}, \boldsymbol{v})\, \mathfrak{h}(\boldsymbol{v}) + \nabla_{\boldsymbol{u},\boldsymbol{v}} K_\lambda(\boldsymbol{u}, \boldsymbol{v}) \ . \tag{3.1}$$

Then the KSD, which solely involves $\mathfrak{h}_q$, can be equivalently represented by

$$D_\lambda(P, Q) = \mathbb{E}_{(\boldsymbol{u},\boldsymbol{v}) \overset{\text{i.i.d.}}{\sim} P}\{\kappa_\lambda[\mathfrak{h}_q](\boldsymbol{u}, \boldsymbol{v})\}. \tag{3.2}$$

We now turn to the compound estimation problem and discuss its connection to the KSD. Proposition 1 demonstrates that, when f is known, the optimal estimator is constructed using $\boldsymbol{h}_f(\boldsymbol{X}) = \{\mathfrak{h}_f(\boldsymbol{x}_1), \ldots, \mathfrak{h}_f(\boldsymbol{x}_n)\}^T$, the conditional score function evaluated at the n observed data points. Define

$$\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}) = \boldsymbol{h}^T \boldsymbol{K}_\lambda \boldsymbol{h} + 2\boldsymbol{h}^T \nabla \boldsymbol{K}_\lambda \mathbf{1} + \mathbf{1}^T \nabla^2 \boldsymbol{K}_\lambda \mathbf{1}, \tag{3.3}$$

where $(\nabla^2 \boldsymbol{K}_\lambda)_{ij} = n^{-2} \nabla_{x_{1j}} \nabla_{x_{1i}} K_\lambda(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{K}_\lambda$ is the $n \times n$ Gaussian kernel matrix with bandwidth λ, and $\nabla \boldsymbol{K}_\lambda$ is defined in Section 2.2. As the extra term $\mathbf{1}^T \nabla^2 \boldsymbol{K}_\lambda \mathbf{1}$ is independent of $\boldsymbol{h}$, it is easy to see that the convex program (2.3) is equivalent to minimizing $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h})$ over $\boldsymbol{h}$. When $\boldsymbol{h}$ is set to $\boldsymbol{h}_q = \{\mathfrak{h}_q(\boldsymbol{x}_1), \ldots, \mathfrak{h}_q(\boldsymbol{x}_n)\}$, (3.3) becomes the empirical version of the KSD defined in (3.2):

$$\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}_q) = D_\lambda(\hat{F}_n, Q) = \mathbb{E}_{(\boldsymbol{u},\boldsymbol{v}) \overset{\text{i.i.d.}}{\sim} \hat{F}_n}\{\kappa_\lambda[\mathfrak{h}_q](\boldsymbol{u}, \boldsymbol{v})\} = \frac{1}{n^2} \sum_{i,j=1}^{n} \kappa_\lambda[\mathfrak{h}_q](\boldsymbol{x}_i, \boldsymbol{x}_j) \ ,$$

where, $\hat{F}_n = n^{-1} \sum_{i=1}^n \delta_{\boldsymbol{x}_i}$ is the empirical distribution function.

We will now present a heuristic explanation of the optimization criterion (2.3). As $n \to \infty$, $\hat{F}_n \xrightarrow{P} F$ and it follows that $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}_q) \xrightarrow{P} \mathcal{S}_\lambda(\mathfrak{h}_q) := D_\lambda(F, Q)$. While stronger versions

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

of these results, such as uniform convergence, are available, we will not delve into those intricate details in this heuristic explanation of the proposed method's working principle. Moreover, $D_\lambda(F, Q) = 0$ iff the first conditional distributions given the rest are equal, i.e., $F(u_1|\boldsymbol{u}_{-1}) = Q(u_1|\boldsymbol{u}_{-1})$. Thus, if we could have minimized $S_\lambda(\mathfrak{h}_q)$ over the class $\mathcal{H} = \{\mathfrak{h}_q = \nabla_1 \log q(\boldsymbol{x}) : q \text{ is any density on } \mathbb{R}^{K+1}\}$, then the minimum would be achieved at the true FCS $\mathfrak{h}_f$ and the minimum value would be 0. However, $\mathcal{S}_\lambda(\mathfrak{h}_q)$ involves the unknown true distribution F, making direct minimization impractical. Alternatively, we minimize the corresponding sample based criterion $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}_q)$ in (3.3) (or equivalently, (2.3)). In large-sample situations, we expect the sampling fluctuations to be small; hence, minimizing $\widehat{\mathcal{S}}_{\lambda,n}$ will lead to score function estimates very close to the true score functions.

3.2 Score estimation under the L_p loss

The next three subsections formulate a rigorous theoretical framework, in the context of compound estimation, to derive the convergence rates of the proposed estimator.

The criterion (2.3) involves minimizing the V-statistic $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h})$. Using standard asymptotics results for V-statistics (Serfling, 2009), it follows that for any density q, $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}_q) \to \mathcal{S}_\lambda(\mathfrak{h}_q)$ in probability as $n \to \infty$. Also, it follows from, for example, Liu et al. (2016), that $\hat{\boldsymbol{h}}_{\lambda,n}$, the solution to (2.3), satisfies:

$$n^{-2} \sum_{i,j} K_\lambda(\boldsymbol{x}_i, \boldsymbol{x}_j) \left\{\hat{\boldsymbol{h}}_{\lambda,n}(i) - \mathfrak{h}_f(\boldsymbol{x}_i)\right\} \left\{\hat{\boldsymbol{h}}_{\lambda,n}(j) - \mathfrak{h}_f(\boldsymbol{x}_j)\right\} = O_P(n^{-1}) \tag{3.4}$$

as $n \to \infty$, where $\hat{\boldsymbol{h}}_{\lambda,n}(i) = \hat{\boldsymbol{h}}_{\lambda,n}(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$.

For implementational ease, we relax the optimization space from the set of all conditional score functions $\mathcal{H}$ to the set all of all real functionals on $\mathbb{R}^{K+1}$. Due to the presence of structural constraints discussed in Section 2.1, this relaxation has little impact on the numerical performance of the NIT estimator. Simulations in Section 4 show that the solutions to (2.3) produce efficient estimates.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

While (3.4) shows that in the RKHS norm the estimates are asymptotically close to the true score functions, for most practical purposes we need to establish the convergence under the ℓ_p norm. For $p > 0$, define $\ell_p(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f) = n^{-1} \sum_{i=1}^{n} |\hat{\boldsymbol{h}}_{\lambda,n}(\boldsymbol{x}_i) - \mathfrak{h}_f(\boldsymbol{x}_i)|^p$. The case of $p = 2$ corresponds to Fisher's divergence. Denote the RKHS norm on the left side of (3.4) by $d_\lambda(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f)$. The essential difficulty in the analysis is that the isometry between the RKHS metric and ℓ_p metric may not exist. Concretely, for any $\lambda > 0$, we can show that $d_\lambda \leq C_1 \, \ell_2$, where C_1 is a constant. However, the inequality in the other direction does not always hold. We aim to show that $\ell_2 \leq C_2 \, d_\lambda$ for some constant C_2; this would produce the desired bound on the L_p risk. Next we provide an overview of the main ideas and key contributions of the theoretical analyses in later subsections.

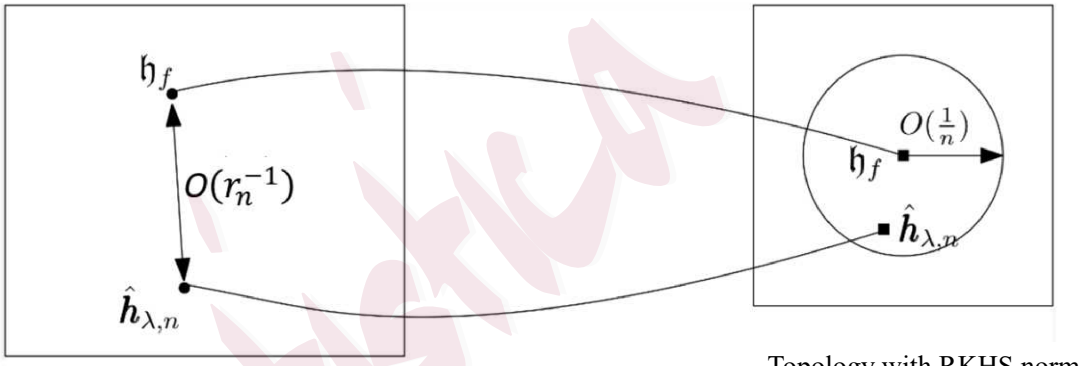


Figure 1: Schematic illustrating the relation between the RKHS and ℓ_1 risks of $\hat{\boldsymbol{h}}_{\lambda,n}$. The (approximate) isometry can be still established. However, the error rate is increased from n^{-1} to r_n^{-1} due to inversion.

In Sections 3.3 and 3.5, we show that as $n \to \infty$ then $\ell_2(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f) \leq c_{\lambda,n} \, d_\lambda(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f)\{1 + o_P(1)\}$, for some $c_{\lambda,n}$ that depends on λ and n only. This result, coupled with the convergence in the RKHS norm (3.4), produces a tolerance bound on the L_p risk of $\hat{\boldsymbol{h}}_{\lambda,n}$, which is

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

subsequently minimized over the choice of λ (Theorem 1 in Section 3.3). However, as a result of inverting, the n^{-1} error rate in (3.4) is increased to $r_n^{-1} \approx n^{-1/(K+2)}$ for the ℓ_p risk (in Theorem 1, we let $p = 1$). Figure 2 provides a schematic description of the phenomenon; further explanations regarding this error rate are provided after Theorem 1.

We point out that existing KSD minimization approaches, including the proposed NIT procedure, involve first mapping the observed data into RKHS and subsequently estimating unknown quantities under the RKHS norm. A tacit assumption for developing theoretical guarantees on the l_p risk is that the lower RKHS loss would also translate to lower ℓ_p loss; see, for example, Assumption 3 of Banerjee et al. (2021) and Section 5.1 of Liu et al. (2016). Heuristically, if $K_\lambda(\boldsymbol{u}, \boldsymbol{v}) = c_\lambda I(\boldsymbol{u} = \boldsymbol{v})$, with $c_\lambda \to \infty$ as $n \to \infty$, then (3.4) would imply $\ell_2(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f) \to 0$. By rigorously characterizing the asymptotic quasi-geodesic between the two topologies, it can be shown that there exists such choices of λ. We provide a complete analysis of the phenomenon that our score function estimates in the RKHS transformed space has controlled ℓ_p risk for the compound estimation problem. This analysis, which is new in the literature, also yields the rates of convergence for the ℓ_p error of the proposed NIT estimator in the presence of covariates.

3.3 Convergence rates for sub-exponential densities

To facilitate a simpler proof, in this section we assume that the true $(K + 1)$-dimensional joint density f as well as its score function $\mathfrak{h}_f$ are Lipschitz continuous. We first provide results for sub-exponential densities, which encompass the popular cases with Gaussian and exponential priors; the convergence rates for heavier-tail priors are discussed in Section 3.5.

Assumption 1. The $(K + 1)$ dimensional joint density f is sub-exponential.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

Our main result is concerned with the ℓ_1 risk of the solution from (2.3). The following theorem shows that the mean absolute deviation of the solution from the true score function is asymptotically negligible as $n \to \infty$. In the theorem we adopt the notation $a_n \asymp b_n$ for two sequences a_n and b_n, which means that $c_1 a_n \leq b_n \leq c_2 b_n$ for all large n and some constants $c_2 \geq c_1 > 0$.

Theorem 1. *Under Assumption 1, as $n \to \infty$ with $\lambda \asymp n^{-1/(K+2)}$,*

$$r_n \cdot \Big\{ \frac{1}{n} \sum_{i=1}^{n} \Big| \hat{\boldsymbol{h}}_{\lambda,n}(i) - \mathfrak{h}_f(\boldsymbol{x}_i) \Big| \Big\} \to 0 \quad \textit{in } L_1, \tag{3.5}$$

where, $r_n = n^{1/(K+2)} \{\log(n)\}^{-(2K+5)}$.

Remark 2. It follows immediately from Theorem 1 that the deviations between our proposed estimate and the oracle estimator in (2.2) obeys:

$$r_n \Big\{ \frac{1}{n} \sum_{i=1}^{n} \Big| \hat{\boldsymbol{\delta}}_{\lambda}^{\mathsf{IT}}(i) - \delta_i^{\pi}(y_i | \boldsymbol{s}_i) \Big| \Big\} \to 0 \text{ in } L_1 \text{ as } n \to \infty. \tag{3.6}$$

Under the classical setting with no auxiliary data ($K = 0$), we achieve the traditional $\sqrt{n}-$rate as established in Jiang and Zhang (2009). However, the convergence rate $r_n \sim n^{1/(K+2)}$ (barring poly-log terms) decreases polynomially in n as K increases. Noting that, for a $(K+1)$ dimensional Gaussian density, the rate of convergence of the mean integrated squared error for the optimally tuned kernel density estimate is $n^{4/(K+1)}$ (Wand and Jones, 1995), we see similar non-parametric (polynomial decay but different exponent) deterioration type in the convergence rate of our estimator as K increases. Adding additional structural constraints discussed in Section 2.3 can greatly improve this convergence rate but the resultant estimator might be highly sub-optimal under misspecification, i.e., when the structural constraints introduced in the model is not true for the data generation process. We pro-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

vide further discussions on the convergence rate of our proposed estimator, as well as its implications for transfer learning, in Section 3.4.

Next we sketch the outline of and main ideas behind the proof of Theorem 1; detailed arguments are provided in the supplement. Consider

$$\Delta_{\lambda,n} := \mathbb{E}\left\{d_\lambda(\boldsymbol{h}_{\lambda,n}, \mathfrak{h}_f)\right\} = \mathbb{E}_{\boldsymbol{X}_n}\left[K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_n)\left\{\hat{\boldsymbol{h}}_{\lambda,n}(1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right\}\left\{\hat{\boldsymbol{h}}_{\lambda,n}(n) - \mathfrak{h}_f(\boldsymbol{x}_n)\right\}\right],$$

where the expectation is taken over $\boldsymbol{X}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\boldsymbol{x}_i$ are i.i.d. samples from f. From (3.4) it follows that $\Delta_{\lambda,n} = O(n^{-1})$. For $\lambda \to 0$, $K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_n)$ is negligible only when $||\boldsymbol{x}_1 - \boldsymbol{x}_n||_2$ is small. Thus for studying the asymptotic behavior of $\Delta_{\lambda,n}$, we shall restrict ourselves on the event where $||\boldsymbol{x}_1 - \boldsymbol{x}_n||_2$ is small. Conditional on this event, we show that $\Delta_{\lambda,n}$ can be well approximated by $\kappa_{\lambda,n}\bar{\Delta}_{\lambda,n-1}$, where $\bar{\Delta}_{\lambda,n-1} = \mathbb{E}_{\boldsymbol{X}_{n-1}}[\{\hat{\boldsymbol{h}}_{\lambda,n}(1) - \mathfrak{h}_f(\boldsymbol{x}_1)\}^2 f(\boldsymbol{x}_1)]$ and the expectation is taken over $\boldsymbol{X}_{n-1} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1})$. To heuristically understand the genesis of $\bar{\Delta}_{\lambda,n-1}$, substitute $\boldsymbol{x}_1 + \boldsymbol{\epsilon}$ in place of $\boldsymbol{x}_n$ in the expression:

$$\Delta_{\lambda,n} = \int K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_n)\{\hat{\boldsymbol{h}}_{\lambda,n}(1) - \mathfrak{h}_f(\boldsymbol{x}_1)\}\{\hat{\boldsymbol{h}}_{\lambda,n}(n) - \mathfrak{h}_f(\boldsymbol{x}_n)\} f(\boldsymbol{x}_1)\ldots f(\boldsymbol{x}_n)\,\mathrm{d}\boldsymbol{x}_1 \ldots \mathrm{d}\boldsymbol{x}_n$$

and let $|\boldsymbol{\epsilon}| \to 0$. As $\lambda \to 0$, the contributions from the kernel weight K_λ can be separated out of the expression and subsequently accounted by constants $\kappa_{\lambda,n}$. Meanwhile the remaining terms produce $\bar{\Delta}_{\lambda,n-1}$. The rate at which $|\boldsymbol{\epsilon}| \to 0$ needs to be appropriately tuned with λ to get the optimal rate of convergence; a rigorous probability argument is provided in the supplement. We shall see that the intermediate quantity $\bar{\Delta}_{\lambda,n-1}$, which links the L_p and RKHS norms, can be explicitly characterized. The rate of convergences will be established by sandwiching $\bar{\Delta}_{\lambda,n-1}$ with functionals involving L_1 and L_2 norms.

Finally we present a result investigating the performance of the NIT estimator under the mean squared loss. Using sub-exponential tail bounds, the ℓ_2 loss of score functions

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

can be obtained by extending the results on ℓ_1 loss. The difference in the mean squared losses between the oracle and data-driven NIT estimators can be subsequently characterized. Lemma 1 below shows that this difference is asymptotically negligible.

Lemma 1. *For any unknown prior* Π *satisfying Assumption 1 and* $\lambda \asymp n^{-1/(K+2)}$,

$$\mathcal{L}_n^2(\hat{\boldsymbol{\delta}}_\lambda^{\mathsf{IT}}, \boldsymbol{\theta}) - \mathcal{L}_n^2(\boldsymbol{\delta}^\pi, \boldsymbol{\theta}) = o_p(r_n^{-1}) \text{ as } n \to \infty. \tag{3.7}$$

Combining (3.6) and (3.7), we have established the asymptotic optimality of the data-driven NIT procedure by showing that it achieves the risk performance of the oracle rule asymptotically when $n \to \infty$; this theory is corroborated by the numerical studies in Section 4.

3.4 Benefits and caveats in exploiting auxiliary data

The amount of efficiency gain of the data-driven NIT estimator depends on two factors: (a) the usefulness of the side information and (b) the precision of the approximation to the oracle. Intuitively when the dimension of the side information increases, the former increases whereas the latter deteriorates.

Consider the Tweedie estimator $y_i + \sigma^2 \nabla \log f_1(y_i)$ that only uses the marginal density f_1 of Y and no auxiliary information. The Fisher information based on the marginal f_1 and the conditional density $f(y|\boldsymbol{s})$ are

$$I_Y = \int \left\{ \frac{f_1'(y)}{f_1(y)} \right\}^2 f_1(y)\, dy \text{ and } I_{Y|\boldsymbol{S}} = \int \left\{ \frac{\nabla_y f(y|\boldsymbol{s})}{f(y|\boldsymbol{s})} \right\}^2 f(y, \boldsymbol{s})\, dy\, d\boldsymbol{s}\ .$$

The following proposition, which follows from Brown (1971) (for completeness a proof is provided in the supplement), shows that, under the oracle setting, utilizing side information is always beneficial, and the efficiency gain becomes larger when more columns of auxiliary data are incorporated into the estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

Proposition 2. *Consider hierarchical model (1.1)–(2.1). Let* $\boldsymbol{\delta}^{\pi}(\boldsymbol{y})$ *and* $\boldsymbol{\delta}^{\pi}(\boldsymbol{y}, \boldsymbol{S})$ *respectively denote the oracle estimator with only* $\boldsymbol{y}$ *and the oracle estimator with both* $\boldsymbol{y}$ *and* $\boldsymbol{S}$*. The efficiency gain due to usage of auxiliary information is*

$$B_n\left\{\boldsymbol{\delta}^{\pi}(\boldsymbol{y})\right\} - B_n\left\{\boldsymbol{\delta}^{\pi}(\boldsymbol{y}, \boldsymbol{S})\right\} = \sigma_y^4\left\{I_{(Y|\boldsymbol{S})} - I_Y\right\} \geq 0 \ .$$

The above equality is attained if and only if the primary variable is independent of all auxiliary variables.

Theorems 1 and Lemma 1 demonstrate that as the dimension K increases, the rate of convergence r_n decreases. This means that while adding more columns of auxiliary data (even if they are non-informative) theoretically never leads to a loss, there is still a tradeoff under our estimation framework. Specifically, the increase of K can widen the gap between the oracle and data-driven rules and potentially offset the benefits of including additional side information. To better understand this tradeoff, we present a numerical example that highlights two key aspects of the phenomenon.

Consider the hierarchical model (1.1)–(2.1). We draw the latent vector $\boldsymbol{\xi}$ from a two-point mixture model, with equal probabilities on two atoms 0 and 2, *i.e.* $\xi_i \sim 0.5\delta_{\{0\}} + 0.5\delta_{\{2\}}$. The mean vectors are simulated as $\theta_i = \xi_i + \eta_{y,i}$ and $\mu_{k,i} = \xi_i + \eta_{k,i}$, $1 \leq k \leq K$ with $\eta_{y,i}, \eta_{k,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Finally we generate $Y_i \sim \mathcal{N}(\theta_i, 1)$ and $S_{k,i} \sim \mathcal{N}(\mu_{k,i}, 1)$, $1 \leq k \leq K$. We vary K from 1 to 12 and compare the oracle and data-driven NIT procedures in Figure 2. We can see that the increase of K has two effects: (a) the MSE of the oracle NIT procedure decreases steadily, while (b) the gap between the oracle and data-driven NIT procedures increases quickly. The combined effect initially leads to a rapid decrease in the MSE of the data-driven NIT procedure, but the decline slackens as $K \geq 5$.

In light of the above discussion, it follows that when we have a large number of auxiliary

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

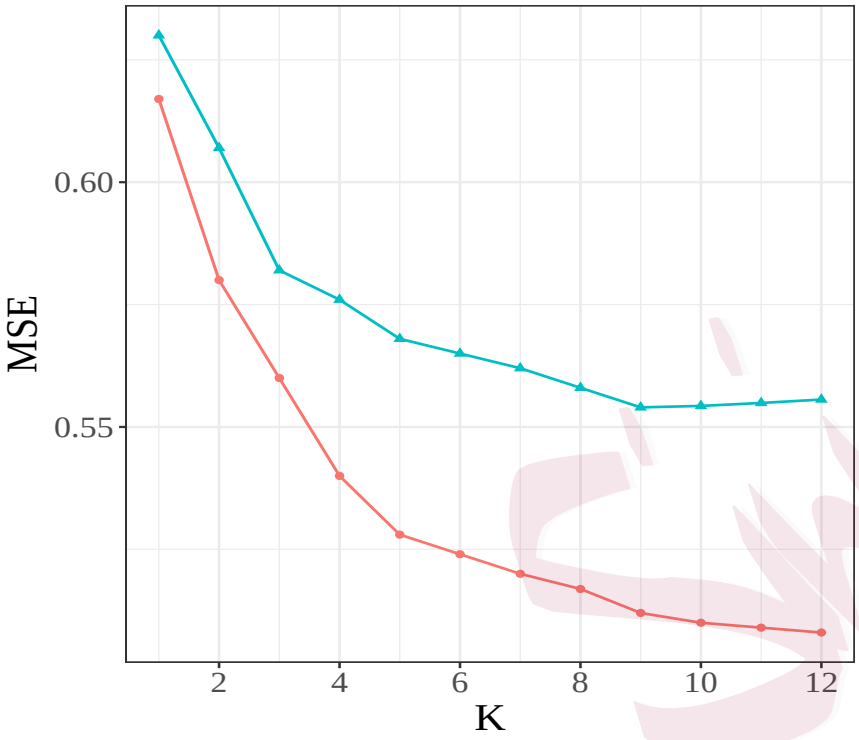


Figure 2: Mean squared error (MSE) of our proposed method (in sky blue) is plotted along with the oracle risk (in magenta) as the number of auxiliary variable (K) increases. The MSE of the oracle procedure always decreases but the MSE of the data-driven NIT procedure stops decreasing as $K \geq 9$.

variables, it may be beneficial to conduct compress the auxiliary data to lower dimensions before applying the NIT estimator. Another remedy can be to impose structural constraints such as monotonicity or lower-dimensional functional relationship between the primary and the auxiliary data akin to (Ignatiadis et al., 2023).

3.5 Convergence rates for heavy-tail densities

In this section, we expand upon the results presented in Section 3.3 to encompass a broader range of prior distributions. We still assume that the true $(K+1)$-dimensional joint density

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

f as well as its score function $\mathfrak{h}_f$ are Lipschitz continuous. Akin to conditions in Theorem 5.1 of Xie et al. (2012) we further assume that the density has $(2+\delta)$ moment bounded for some $\delta > 0$.

Assumption 2: For some $\delta > 0$, the $K+1$ dimensional joint density f has bounded $2+\delta$ moment, i.e., $\mathbb{E}_{\boldsymbol{x}\sim f}\|\boldsymbol{x}\|^{2+\delta} < \infty$.

The next theorem shows that, for suitably chosen bandwidth, the data-driven NIT estimator is asymptotically close to the oracle estimator and the difference in their losses also converges to 0 under any prior satisfying Assumption 2. The rate of convergence is slower than that of Theorem 1, which is mainly due to the larger terms needed to bound heavier tails. Similar to Theorem 1, the rate decreases with the increase of K.

Theorem 2. *Under Assumption 2, with* $\lambda \asymp n^{-1/(K+2)}$ *and*

$$r_n = n^{\delta(K+2)^{-1}(K+3+2\delta)^{-1}}(\log n)^{-K-3},$$

we have

$$r_n \cdot \left\{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\boldsymbol{h}}_{\lambda,n}(i) - \mathfrak{h}_f(\boldsymbol{x}_i)\right|\right\} \to 0 \quad \textit{in } L_2 \textit{ as } n \to \infty.$$

Additionally, we have $\mathcal{L}_n^2(\hat{\boldsymbol{\delta}}_\lambda^{\mathsf{IT}}, \boldsymbol{\theta}) - \mathcal{L}_n^2(\boldsymbol{\delta}^\pi, \boldsymbol{\theta}) = o_p(r_n^{-1})$ *as* $n \to \infty$.

The rate r_n above converges to the rate in Theorem 1 as $\delta \to \infty$, which heuristically translates to the existence of all possible moments. Also, theorems 1 and 2 are based on the same value of the bandwidth λ.

3.6 Consistency of the MCV criterion

In sections 3.3 and 3.5, we have established asymptotic risk properties of our proposed method as bandwidth $\lambda \to 0$. For finite sample sizes, it is important to select the "best"

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

bandwidth based on a data-driven criterion as provided in Section 2.3. The following proposition establishes the consistency of the validation loss to the true loss, justifying the effectiveness of the bandwidth selection rule.

Proposition 3. *For any fixed $\lambda > 0$ and n, we have*

$$\lim_{\alpha \to 0} \mathbb{E}\Big\{ \hat{L}_n(\lambda, \alpha) - \mathcal{L}_n^2(\hat{\boldsymbol{\delta}}_\lambda^{\mathsf{IT}}, \boldsymbol{\theta}) \Big\} = 0 \ .$$

provided that there is a unique solution to (2.3) *for* $\alpha = 0$.

4. Simulation

We consider three different settings where the structural information is encoded in (a) one *given* auxiliary sequence that shares structural information with the primary sequence through a common latent vector (Section 4.1); (b) one auxiliary sequence carefully *constructed within* the same data to capture the sparsity structure of the primary sequence (Section S8.1 of the Supplement); (c) *multiple* auxiliary sequences that share a common structure with the primary sequence (Section 4.2).

The following methods are considered in the comparison: (a) James-Stein estimator (JS); (b) the empirical Bayes Tweedie (EBT) estimator implemented using kernel smoothing as described in Brown and Greenshtein (2009); (c) the NPMLE method by Koenker and Mizera, 2014, implemented by the R-package `REBayes` in Koenker and Gu (2017a); (d) the empirical Bayes with cross-fitting (EBCF) method by Ignatiadis and Wager (2019); (e) the oracle NIT procedure (2.2) with known $f(y|\boldsymbol{s})$ (NIT.OR); (f) the data-driven NIT procedure (2.4) by solving the convex program (NIT.DD). The last three methods, which utilize auxiliary data, are expected to outperform the first three methods when the side information is informative.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

The MSE of NIT.OR is provided as the optimal benchmark for assessing the efficiency of various methods.

To implement NIT.DD, we employ the generalized Mahalanobis distance, as discussed in Section 2.3, to compute the RBF kernel with bandwidth λ. To select an optimal λ, we solve optimization problems 2.3 across a range of λ values and then compute the corresponding modified cross-validation (MCV) loss. The data-driven bandwidth is chosen as the value of λ that minimizes the validation loss (2.5).

4.1 Simulation 1: integrative estimation with one auxiliary sequence

Let $\boldsymbol{\xi} = (\xi_i : 1 \le i \le n)$ be a latent vector obeying a two-point normal mixture:

$$\xi_i \sim 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(1,1).$$

The primary data $\boldsymbol{Y} = (Y_i : 1 \le i \le n)$ in the target domain are simulated according to the following hierarchical model: $\theta_i \sim \mathcal{N}(\xi_i, \sigma^2)$, $\quad Y_i \sim \mathcal{N}(\theta_i, 1)$. By contrast, the auxiliary data $\boldsymbol{S} = (S_i : 1 \le i \le n)$ obeys $\zeta_i \sim \mathcal{N}(\xi_i, \sigma^2)$, $\quad S_i \sim \mathcal{N}(\zeta_i, \sigma_s^2)$. This data generating mechanism is a special case of the hierarchical model (2.1) where both the primary parameter θ_i and auxiliary parameter ζ_i are related to a common latent variable ξ_i, with σ controlling the amount of common information shared by θ_i and ζ_i. We further use σ_s to reflect the noise level when collecting data in the source domain. The auxiliary sequence $\boldsymbol{S}$ becomes more useful when both σ and σ_s decrease. We consider the following settings to investigate the impact of σ, σ_s and sample size n on the performance of different methods.

Setting 1: we fix $n = 1000$ and $\sigma \equiv 0.1$, then vary σ_s from 0.1 to 1.

Setting 2: we fix $n = 1000$ and $\sigma_s \equiv 1$, then vary σ from 0.1 to 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

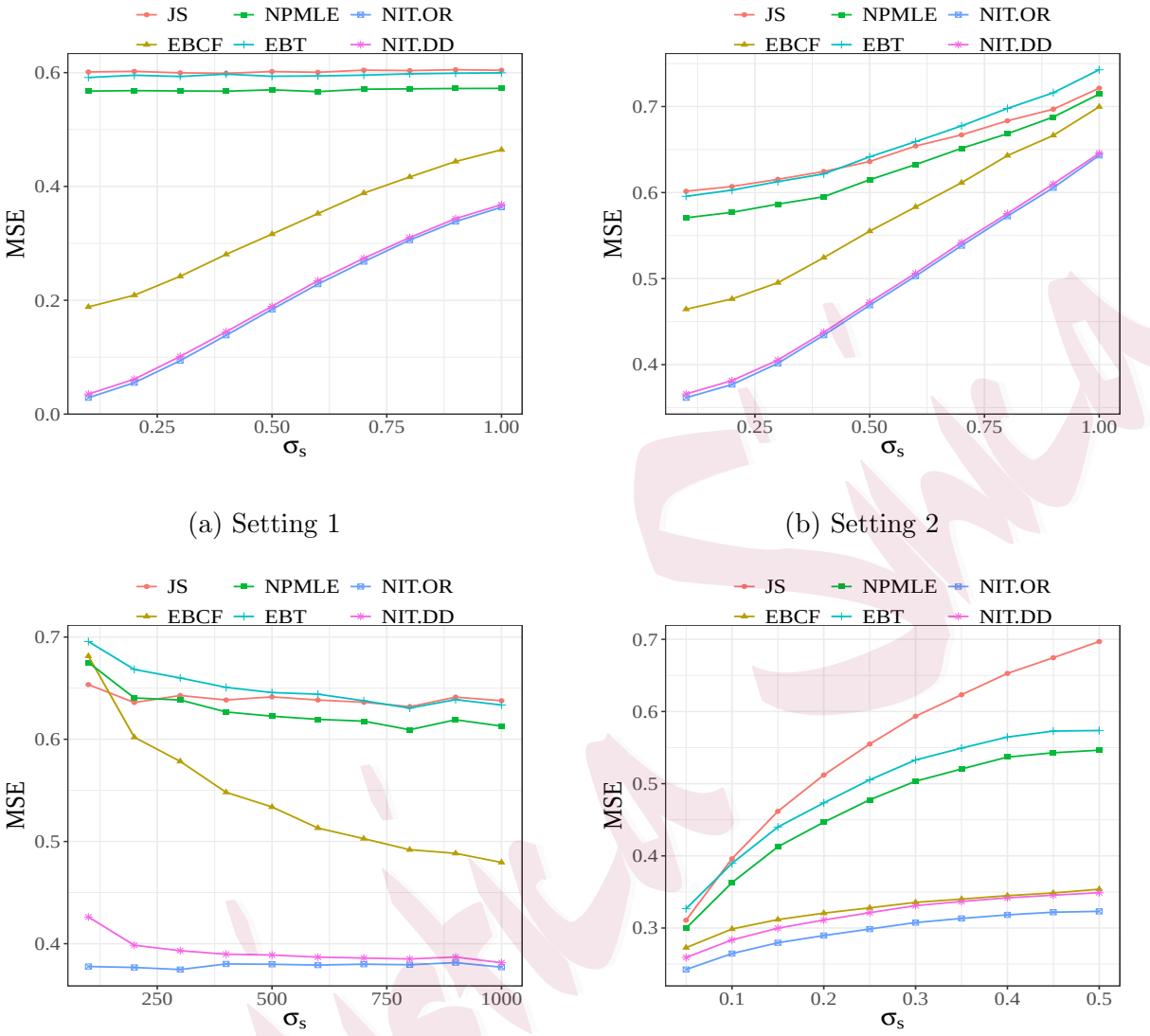
Setting 3: we fix $\sigma_s \equiv 0.5$ and $\sigma \equiv 0.5$, then vary n from 100 to 1000.

Finally we consider a setup where the auxiliary sequence is a binary vector. In the implementation of NIT.DD for categorical variables, we use indicator function to compute the pairwise distance between categorical variables. Precisely, assume that s_i and s_j are two categorical variables, then the distance $d(s_i, s_j) = \mathbf{1}(s_i = s_j)$.

Setting 4: Let $\boldsymbol{\xi} = (\xi_i : 1 \leq i \leq n)$ be a latent vector obeying a Bernoulli distribution $\xi_i \sim \text{Bernoulli}(p)$. The primary sequence in the target domain is generated according to a hierarchical model: $\theta_i \sim \mathcal{N}(2\xi_i, 0.25)$, $y_i \sim \mathcal{N}(\theta_i, 1)$. The auxiliary vector is a noisy version of the latent vector: $s_i \sim (1 - \xi_i)\text{Bernoulli}(0.05) + \xi_i\text{Bernoulli}(0.9)$. We fix $n = 1000$ and vary p from 0.05 to 0.5.

We apply different methods to simulated data generated by the models described above and calculate the MSEs over 100 replications. Figure 3 presents the simulation results for Settings 1-4, from which we observe several important patterns. First, the integrative methods (NIT.DD, EBCF) outperform univariate methods (JS, NPMLE, EBT) that do not incorporate auxiliary information in most settings. Moreover, NIT.DD consistently outperforms EBCF, with substantial efficiency gains observed in many cases. This is not unexpected since under the data generating scheme of Simulation 1, the conditional distribution of θ_i given S_i under Settings 1-4 is not necessarily Gaussian and this represents a deviation from the hierarchical model of Ignatiadis et al. (2023) upon which EBCF relies. Second, the efficiency gain of the integrative methods decreases as σ and σ_s increase (i.e., when the auxiliary data become less informative or more noisy), as indicated by Settings 1-2. Third, as shown in Setting 3, sample size has a significant impact on integrative empirical Bayes estimation, with larger sample sizes being essential for effectively integrating side information. EBCF

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063



(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Figure 3: Simulation results for one given auxiliary sequence.

may under-perform univariate methods when n is small. Fourth, the gap between NIT.OR and NIT.DD narrows as n increases. Finally, Setting 4 demonstrates that side information can be highly informative even when the types of primary and auxiliary data differs.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

4.2 Simulation 3: integrative estimation with multiple auxiliary sequences

This section considers a setup where auxiliary data are collected from multiple source domains. Denote $\boldsymbol{Y}$ the primary sequence and $\boldsymbol{S}^j$, $1 \le j \le 4$, the auxiliary sequences. In our simulation, we assume that the primary vector $\boldsymbol{\theta}_Y = \mathbb{E}(\boldsymbol{Y})$ share some common information with auxiliary vectors $\boldsymbol{\theta}_S^j = \mathbb{E}(\boldsymbol{S}^j)$, $1 \le j \le 4$ through a latent vector $\boldsymbol{\eta}$, which obeys a mixture model with two point masses at 0 and 2 respectively:

$$\eta_i \sim 0.5\delta_{\{0\}} + 0.5\delta_{\{2\}}, \quad 1 \le i \le n.$$

There can be various ways to incorporate auxiliary data from multiple sources. We consider, in addition to NIT.DD that utilizes all sequences, an alternative strategy that involves firstly constructing a new auxiliary sequence $\bar{\boldsymbol{S}} = \frac{1}{4}(\boldsymbol{S}^1 + \boldsymbol{S}^2 + \boldsymbol{S}^3 + \boldsymbol{S}^4)$ to reduce the dimension and secondly applying NIT.DD to the pair $(\boldsymbol{Y}, \bar{\boldsymbol{S}})$; this strategy is denoted by NIT1.DD. Intuitively, if all auxiliary sequences share identical side information, then data reduction via $\bar{\boldsymbol{S}}$ is lossless. However, if the auxiliary data are collected from heterogeneous sources with different structures and measurement units, then NIT1.DD may distort the side information and lead to substantial efficiency loss.

To illustrate the benefits and caveats of different data combination strategies, we first consider the scenario where all sequences share a common structure via the same latent vector (Settings 1-2). Then we turn to the scenario where the auxiliary sequences share information with the primary data in distinct ways (Settings 3-4). In all simulations below we use $n = 1000$ and 100 replications.

Setting 1: The primary and auxiliary data are generated from the following models:

$$Y_i = \theta_i^Y + \epsilon_i^Y, \quad S_i^j = \theta_i^j + \epsilon_i^j, \tag{4.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

where $\theta_i^Y \sim \mathcal{N}(\eta_i, \sigma^2)$, $\theta_i^j \sim \mathcal{N}(\eta_i, \sigma^2)$, $1 \le j \le 4$, $\epsilon_i^Y \sim \mathcal{N}(0,1)$ and $\epsilon_i^j \sim \mathcal{N}(0, \sigma_s^2)$, $1 \le i \le n$. We fix $\sigma = 0.5$ and vary σ_s from 0.1 to 1.

Setting 2: the data are generated using the same models as in Setting 1 except that we fix $\sigma_s = 0.5$ and vary σ from 0.1 to 1.

Setting 3: We generate $\boldsymbol{Y}$ and $\boldsymbol{S}^j$ using model (4.1). However, we now allow $\boldsymbol{\theta}^j$ to have different structures across j. Specifically, let $\boldsymbol{\eta}^1[1:500] = \boldsymbol{\eta}[1:500]$, $\boldsymbol{\eta}^1[501:n] = 0$, $\boldsymbol{\eta}^2[1:500] = 0$ and $\boldsymbol{\eta}^2[501:n] = \boldsymbol{\eta}[501:n]$. The following construction implies that only the first two sequences are informative in inference:

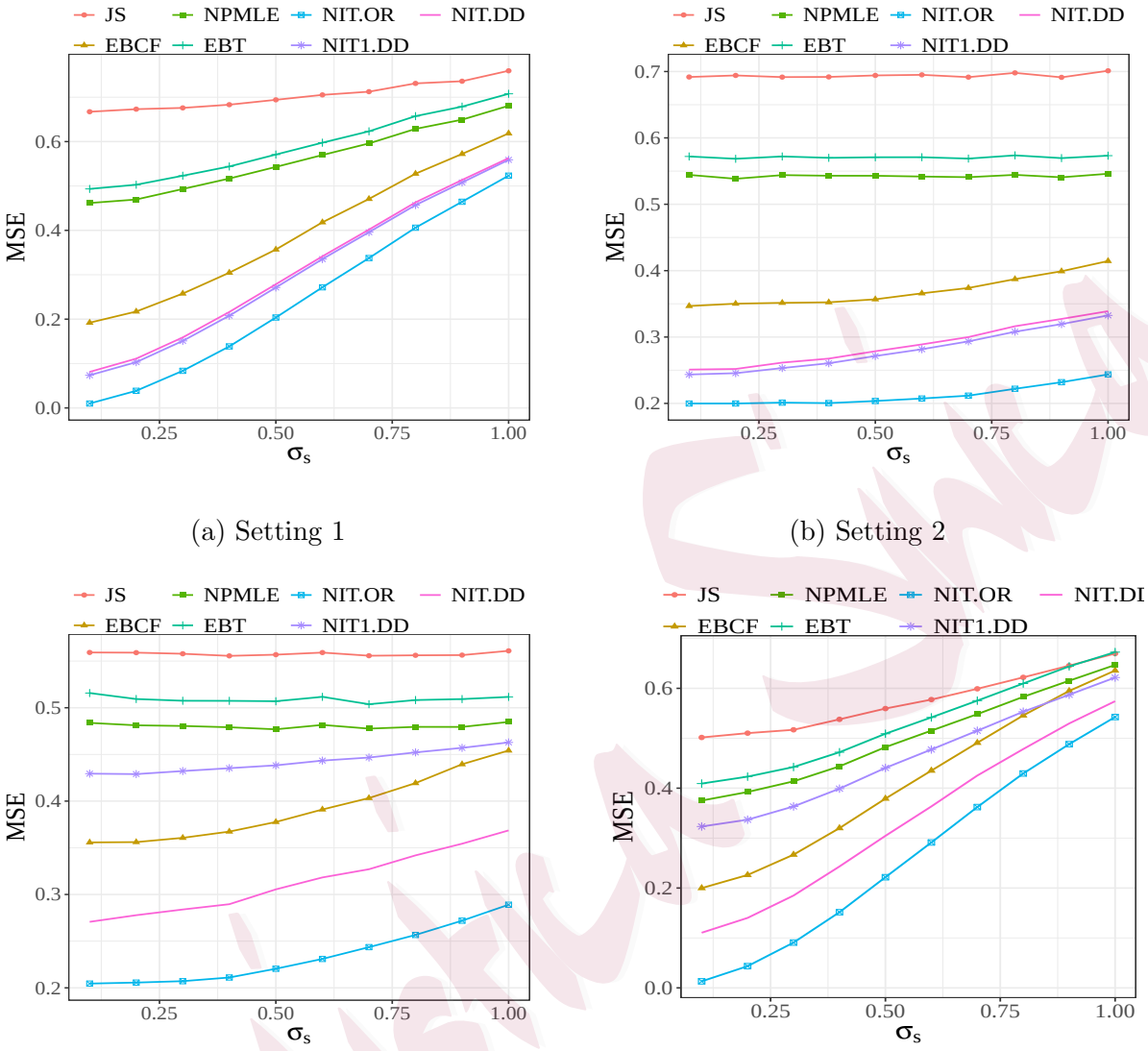
$$\theta_i^Y \sim \mathcal{N}(\eta_i^1, \sigma^2); \quad \theta_i^j \sim \mathcal{N}(\eta_i^1, \sigma^2), j = 1, 2; \quad \theta_i^j \sim \mathcal{N}(\eta_i^2, \sigma^2), j = 3, 4.$$

We fix $\sigma = 0.5$ and vary σ_s from 0.1 to 1.

Setting 4: the data are generated using the same models as in Setting 3 except that we fix $\sigma_s = 0.5$ and vary σ from 0.1 to 1.

We apply different methods to simulated data and summarize the results in Figure 4. Our observations are as follows. First, the integrative methods (NIT.DD, NIT.OR, EBCF, NIT1.DD) outperform the univariate methods (JS, NPMLE, EBT), with the efficiency gain being more pronounced when σ and σ_s are small. Second, NIT.DD dominates EBCF, and the gap between the performances of NIT.OR and NIT.DD widens with higher-dimensional estimation problems involving multiple auxiliary sequences. Third, in Settings 1-2, NIT1.DD is more efficient than NIT.DD as there is no loss in data reduction and fewer sequences are utilized in estimation. Finally, in Settings 3-4, the average $\bar{\boldsymbol{S}}$ does not provide an effective way to combine the information in auxiliary data. Improper data reduction leads to substantial information loss, such that NIT1.DD still outperforms univariate methods but is

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063



(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Figure 4: Integrative estimation with multiple auxiliary sequences.

much worse than EBCF and NIT.DD. Overall, our simulation results suggest that reducing the dimension of auxiliary data can be potentially beneficial, but there can be significant information loss if the data reduction step is carried out improperly. It would be of interest to develop principled methods for data reduction to extract structural information from a large number of auxiliary sequences.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

In Section S8.1 of the supplement we present an additional simulation study involving integrative estimation in two-sample inference of sparse means.

5. Application: Integrative Nonparametric estimation of Gene Expressions

We consider the data set in Sen et al. (2018) that measures gene expression levels from cells that are without interferon alpha (INFA) protein and have been infected with varicella-zoster virus (VZV). VZV is known to cause chickenpox and shingles in humans (Zerboni et al., 2014). INFA helps in host defense against VZV but is often regulated in the presence of virus. Thus, it is important to estimate the gene expressions in infected cells without INFA. Let $\boldsymbol{\theta}$ be the true unknown vector of mean gene expression values that need to be estimated. Further details about the dataset is provided in Section S8.2 of the Supplement.

The data had gene expression measurements from two independent experiments studying VZV infected cells without INFA. We use one vector, denoted $\boldsymbol{Y}$, to construct the estimates and the other, denoted $\tilde{\boldsymbol{Y}}$, for validation. To estimate $\boldsymbol{\theta}$, alongside the primary data $\boldsymbol{Y}$, we also consider auxiliary information: $\boldsymbol{S}_{\mathsf{U}}$ which are corresponding gene expression values from uninfected cells, and Figure 5 shows the heatmap of the primary, the auxiliary and the validation sequences. We implemented the following estimators (a) the modified James-Stein (JS) following Xie et al. (2012), (b) Non-parametric Tweedie estimator without auxiliary information, (c) Empirical Bayes with cross-fitting (EBCF) by Ignatiadis and Wager (2019) and the Non-parametric Integrated Tweedie (NIT) with auxiliary information: (d) with $\boldsymbol{S}_{\mathsf{U}}$ only, (e) with $\boldsymbol{S}_{\mathsf{I}}$ only (f) using both auxiliary sequences. The mean square prediction errors of the above estimates were computed with respect to the validation vector $\tilde{\boldsymbol{Y}}$.

Table 1 reports the percentage gain acheived over the naive unshrunken estimator that

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

uses $\boldsymbol{Y}$ to estimate $\boldsymbol{\theta}$. It shows that non-parametric shrinkage produces an additional 0.6% gain over parametric JS and using auxiliary information via NIT yields a further 5.2% gain. In particular, NIT method outperforms EBCF, which also leverage side information from both S_{U} and S_{I}, by 1.7% gain. Panel B of Figure 5 shows the differences between the Tweedie and NIT estimates. The differences are more pronounced in the left tails where Tweedie estimator is seen to overestimate the levels compared to NIT. The JS and NIT effective size estimates disagree by more than 50% at 28 genes (which are listed in the top panel of Figure 2 in the Supplement). These genes impact 35 biological processes and 12 molecular functions in human cells (see bottom two panels of Figure 2 in the Supplement); this implies that important inferential gains can be made by using auxiliary information via our proposed NIT estimator.

Table 1: % gain in prediction errors by different estimators over the naive unshrunken estimator of gene expressions of INFA regulated infected cells.

Methods	James-Stein	Tweedie	EBCF using $\boldsymbol{S}_{\mathsf{U}}$ & $\boldsymbol{S}_{\mathsf{I}}$	NIT using $\boldsymbol{S}_{\mathsf{U}}$	NIT using $\boldsymbol{S}_{\mathsf{I}}$	NIT using $\boldsymbol{S}_{\mathsf{U}}$ & $\boldsymbol{S}_{\mathsf{I}}$
% Gain	3.5	4.1	7.6	6.9	7.5	9.3
MSE	2.014	2.001	1.927	1.930	1.951	1.895

In Section S8.3 of the supplement we present an additional real data application that involves leveraging auxiliary information to predict monthly sales of common grocery items across stores.

6. Discussion

The NIT procedure introduces a powerful framework for leveraging useful structural knowledge from auxiliary data to facilitate the estimation of a high-dimensional parameter. It

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

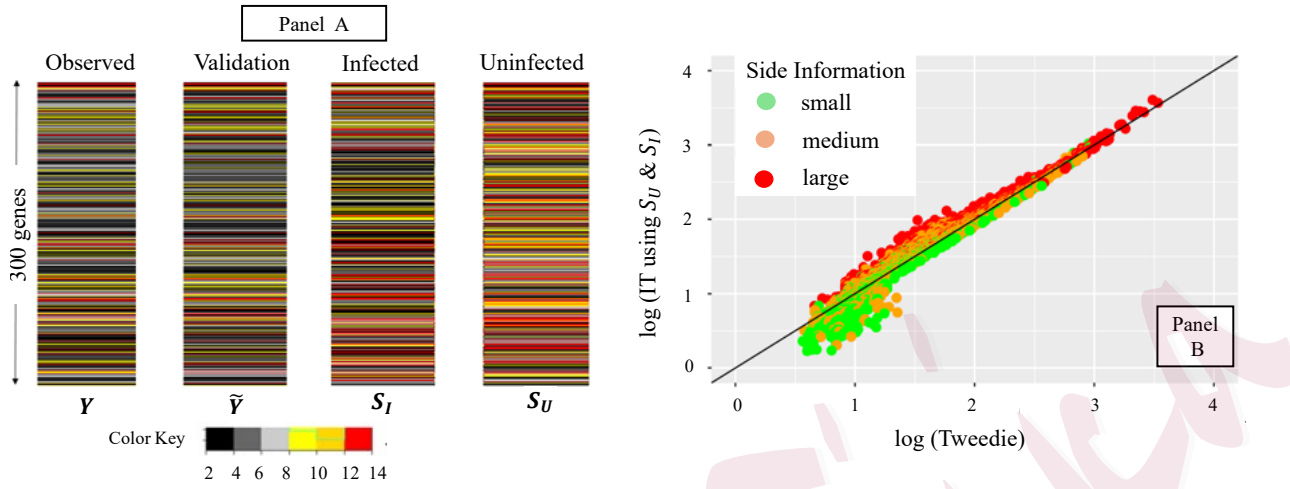


Figure 5: Panel A: Heatmaps of the gene expression datasets showing the four expression vectors corresponding to the observed, validation and auxiliary sequences. Panel B: scatterplot of the effect size estimates of gene expressions based on Tweedie and NIT (using both S_{U} and S_{I}). Magnitude of the auxiliary variables used in the NIT estimate is reflected by different colors.

builds upon classical empirical Bayes ideas and significantly expands upon them, allowing for the handling of multivariate auxiliary data. The framework is highly versatile as it does not impose any distributional assumptions on auxiliary data, which can be categorical, numerical, or of mixed types.

When the noise variance σ^2 is known in Equation (1.1), our theoretical analysis quantifies the reduction in estimation errors and deterioration in learning rates as the dimension of $\boldsymbol{S}$ increases. This suggests that when faced with a large number of variables as potential choices for auxiliary data, it may be beneficial to conduct data reduction prior to applying the NIT estimator. However, our simulation results in Section 4.2 demonstrate that improper data reduction can lead to significant information loss. This highlights the need for

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

future research in three directions: (a) extending the theoretical analysis considered here with a consistent estimator of σ^2, (b) investigating the trade-off between achievable error limits of the oracle rule and the decreased convergence rate of the data-driven rule as K increases, and (c) developing principled structure-preserving dimension reduction methods under the integrative estimation framework to extract useful structural information from a large number of auxiliary sequences.

Supplementary Material

The online Supplement provides the proofs of all results stated in the main paper, an additional numerical experiment, further details regarding the real data example of Section 5 and an additional real data example.

Acknowledgements

The authors thank the Editor, Associate Editor, and three anonymous referees for their thoughtful and constructive feedback, which has substantially improved the quality and presentation of this article.

References

Banerjee, T., L. J. Fu, G. M. James, G. Mukherjee, and W. Sun (2024). Nonparametric empirical bayes estimation on heterogeneous data. *arXiv preprint arXiv:2002.12586*.

Banerjee, T., Q. Liu, G. Mukherjee, and W. Sun (2021). A general framework for empirical bayes estimation in discrete linear exponential family. *Journal of Machine Learning Research 22*(67), 1–46.

Banerjee, T., G. Mukherjee, and D. Paul (2021). Improved shrinkage prediction under a spiked covariance structure. *Journal of machine learning research 22*(180), 1–40.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

Banerjee, T., G. Mukherjee, and W. Sun (2020). Adaptive sparse estimation with side information. *Journal of the American Statistical Association 115*, 2053–2067.

Banerjee, T. and P. Sharma (2025). Nonparametric empirical bayes prediction in mixed models. *Statistics and Computing 35*(5), 145.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological* ***57***, 289–300.

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

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

Cai, T. T., W. Sun, and W. Wang (2019). CARS: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology 81*, 187–234.

Chwialkowski, K., H. Strathmann, and A. Gretton (2016). A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48 of *Proceedings of Machine Learning Research*, New York, New York, USA, pp. 2606–2615. PMLR.

Cohen, N., E. Greenshtein, and Y. Ritov (2013). Empirical bayes in the presence of explanatory variables. *Statistica Sinica 23*(1), 333–357.

Dou, Z., S. Kotekal, Z. Xu, and H. H. Zhou (2024). From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*.

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

Efron, B. (2016). Empirical bayes deconvolution estimates. *Biometrika 103*(1), 1–20.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc. 96*, 1151–1160.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *The Journal of Machine Learning Research 13*(1), 723–773.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

Gu, J. and R. Koenker (2023). Invidious comparisons: Ranking and selection as compound decisions. *Econometrica 91*(1), 1–41.

Ignatiadis, N. and W. Huber (2021). Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(4), 720–751.

Ignatiadis, N., S. Saha, D. L. Sun, and O. Muralidharan (2023). Empirical bayes mean estimation with nonparametric errors via order statistic regression. *Journal of the American Statistical Association 118*(542), 987–999.

Ignatiadis, N. and S. Wager (2019). Covariate-powered empirical bayes estimation. In *Advances in Neural Information Processing Systems*, pp. 9617–9629. Curran Associates, Inc.

Jana, S., Y. Polyanskiy, A. Z. Teh, and Y. Wu (2023). Empirical bayes via erm and rademacher complexities: the poisson model. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5199–5235. PMLR.

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

Jiang, W. and C.-H. Zhang (2010). Empirical bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, Volume 6, pp. 263–274. Institute of Mathematical Statistics.

Jitkrittum, W., H. Kanagawa, and B. Schölkopf (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pp. 221–230. PMLR.

Ke, T., J. Jin, and J. Fan (2014). Covariance assisted screening and estimation. *Annals of statistics 42*(6), 2202–2242.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics 27*(1), 887–906.

Kim, Y., W. Wang, P. Carbonetto, and M. Stephens (2022). A flexible empirical bayes approach to multiple linear regression and connections with penalized regression. *arXiv preprint arXiv:2208.10910*.

Koenker, R. and J. Gu (2017a). REBayes: An R package for empirical bayes mixture methods. *Journal of Statistical Software 82*(8), 1–26.

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

Krusińska, E. (1987). A valuation of state of object based on weighted mahalanobis distance. *Pattern Recognition 20*(4), 413–418.

Lei, L. and W. Fithian (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 649–679.

Li, A. and R. F. Barber (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(1), 45–74.

Liu, Q., J. Lee, and M. Jordan (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR.

Liu, Q. and D. Wang (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, Volume 29, pp. 2378–2386. Curran Associates, Inc.

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

Polyanskiy, Y. and Y. Wu (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*.

Ren, Z. and E. Candès (2020). Knockoffs with side information. *arXiv preprint arXiv:2001.07835*.

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

Roeder, K. and L. Wasserman (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics 24*(4), 398.

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

Sen, N., P. Sung, A. Panda, and A. M. Arvin (2018). Distinctive roles for type i and type ii interferons and interferon regulatory factors in the host cell defense against varicella-zoster virus. *Journal of virology 92*(21), e01151–18.

Serfling, R. (2009). *Approximation Theorems of Mathematical Statistics*. Wiley.

Shen, Y. and Y. Wu (2022). Empirical bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)? *arXiv preprint arXiv:2211.12692*.

Soloff, J. A., A. Guntuboyina, and B. Sen (2021). Multivariate, heteroscedastic empirical bayes via nonparametric

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

maximum likelihood. *arXiv preprint arXiv:2109.03466*.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, STANFORD UNIVERSITY STANFORD United States.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* ***102***, 901–912.

Wand, M. and M. Jones (1995). *Kernel Smoothing*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC.

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 113*(522), 698–710.

Wibisono, A., Y. Wu, and K. Y. Yang (2024). Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*.

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

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

Zerboni, L., N. Sen, S. L. Oliver, and A. M. Arvin (2014). Molecular mechanisms of varicella zoster virus pathogenesis. *Nature reviews microbiology 12*(3), 197–210.

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

Zhang, K., C. H. Yin, F. Liang, and J. Liu (2024). Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*.

Zhang, Y., Y. Cui, B. Sen, and K.-C. Toh (2022). On efficient and scalable computation of the nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2208.07514*.

Jiajun Luo - University of Southern California

E-mail: jiajunluo121@gmail.com

Trambak Banerjee - University of Kansas

E-mail: trambak@ku.edu

Gourab Mukherjee - University of Southern California

E-mail: gourab@usc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0063

Wenguang Sun - Zhejiang University

E-mail: wgsun@zju.edu.cn