Submitted to the Annals of Statistics

TRANSFER LEARNING FOR EMPIRICAL BAYES ESTIMATION: A NONPARAMETRIC INTEGRATIVE TWEEDIE APPROACH

BY JIAJUN LUO[1], GOURAB MUKHERJEE[1] AND WENGUANG SUN[1]

[1] *University of Southern California*
jiajunlu@usc.edu; gourab@usc.edu; wenguans@marshall.usc.edu

We consider compound estimation of normal means with auxiliary data collected from related source domains. The empirical Bayes framework provides an elegant interface to pool information across different samples and construct efficient shrinkage estimators. We propose a nonparametric integrative Tweedie (NIT) approach to transferring structural knowledge encoded in the auxiliary data from related source domains to assist the simultaneous estimation of multiple parameters in the target domain. Our transfer learning algorithm uses convex optimization tools to directly estimate the gradient of the log-density through an embedding in the reproducing kernel Hilbert space (RKHS), which is induced by the Stein's discrepancy metric. Most popular structural constraints can be easily incorporated into our estimation framework. We characterize the asymptotic L_p risk of NIT by first rigorously analyzing its connections to the RKHS risk, and second establishing the rate at which NIT converges to the oracle estimator. The improvements in the estimation risk and the deteriorations in the learning rate are precisely tabulated as the dimension of side information increases. The numerical performance of NIT and its superiority over existing methods are illustrated through the analysis of both simulated and real data.

1. Introduction.

In a broad class of integrative inference problems such as meta analysis, replicability analysis, multi-task learning and multi-view data analysis, an essential task is to combine information from multiple sources to make valid and informative decisions. Consider a compound estimation problem where $\boldsymbol{Y} = (Y_i : 1 \leq i \leq n)$ is a vector of summary statistics in the target domain obeying

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

We assume that σ^2 is known. The goal is to estimate a high-dimensional parameter $\boldsymbol{\theta} = \mathbb{E}(\boldsymbol{Y}) = \{\theta_i : 1 \leq i \leq n\}$. Suppose we also collect K auxiliary sequences $\boldsymbol{S}^{(k)} = \{S_i^{(k)} : 1 \leq i \leq n\}$, $1 \leq k \leq K$, from related source domains. 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.

Transfer learning for large-scale estimation aims to extract and transfer structural knowledge encoded in auxiliary data $\boldsymbol{S}$ to assist the simultaneous estimation of multiple parameters in the target domain. The new setup poses several new challenges in the data aggregation process. First, conventional meta-analytical methods, which often involve estimating an *overall effect* by constructing weighted estimators to combine data across several subpopulations, would become problematic when the source and target distributions differ. A key principle in our methodological development is that the inaccuracy of auxiliary information or mismatch of the target and source distributions should not lead to negative learning. Second, existing

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

meta-analytical methods, which construct weighted estimators for only one or a few parameters, can be highly inefficient in large-scale estimation problems. When thousands of parameters are estimated simultaneously, useful structural knowledge, which is often encoded in auxiliary data sources, is highly informative but has been underexploited in conventional analyses. Finally, most transfer learning theories have focused on classification algorithms. We aim to develop a new theoretical framework to gain understandings of the benefits and caveats of transfer learning for shrinkage estimation.

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 $\{Y_i : 1 \le i \le n\}$ from n independent experiments. Let $\boldsymbol{\delta} = (\delta_i : 1 \le i \le n)$ be a decision rule. Classical ideas such as the compound decision theory [32], empirical Bayes (EB) methods [33] and James-Stein shrinkage estimator [38], as well as the more recent multiple testing methodologies [13, 39], showed that the joint structure of all observations is highly informative, and can be exploited to construct more efficient classification, estimation and multiple testing procedures. For example, the submimimax rule in [32] showed that the disparity in the proportions of positive and negative signals can be incorporated into inference to reduce the mis-classification rate, and the adaptive z-value procedure in [39] showed that the asymmetry in the shape of the alternative distribution can be utilized to construct more powerful false discovery rate (FDR, [3]) procedures.

In light of auxiliary data, the inference units become unequal. This heterogeneity provides new structural knowledge that can be further utilized to improve the efficiency of existing methods. The idea is to first learn the finer structure of the high-dimensional object through auxiliary data and then apply the new structural knowledge to the target domain. For example, in genomics research, prior data and domain knowledge may be used to define prioritized subset of genes. [34] 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 an important topic that has received much recent attention; see [26, 9, 27, 16, 31] for a partial list of references. These works show that the power of existing FDR methods can be substantially improved by utilizing auxiliary data to place differential weights or to set varied thresholds on corresponding test statistics. Similar ideas have been adopted by a few recent works on shrinkage estimation. For example, [41] and [1] propose to incorporate the side information into inference by first creating groups and 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 [19, 7, 12, 23, 18, 35]. Under the nonparametric empirical Bayes framework, the formula is particularly appealing for large-scale estimation problems for it is simple to implement, removes the selection bias [12] and enjoys frequentist's optimality properties asymptotically [19, 7].

This article develops a nonparametric integrative Tweedie (NIT) method to extract and incorporate useful structural knowledge from both primary and auxiliary data. NIT has several merits under the transfer learning setup. First, NIT allows the target and source distributions to differ, which effectively avoids negative learning. Second, NIT provides a general framework for incorporating various types of structural information and can effectively handle multivariate covariates. Finally, in contrast with the linear EB shrinkage estimators [42, 40, 20, 46] that are only optimal under parametric Gaussian priors, NIT belongs to the class of generalized empirical Bayes (GEB) estimators, which enjoy asymptotic and minimax optimality properties for a wide class of models [45, 7].

1.3. *Main ideas of our approach.* The EB implementation of Tweeide's formula involves two steps: first estimating the marginal distribution and then predicting the unknown using a plug-in rule. We describe some important developments in the literature. [45] showed that a truncated GEB estimator, which is based on a Fourier infinite-order smoothing kernel, asymptotically achieves both the Bayes and minimax risks. The GMLEB approach by [19] implements Tweedie's formula by estimating the unknown prior distribution via the Kiefer-Wolfwitz estimator. GMLEB is approximately minimax and universally reduces the estimation risk. [7] employed a nonparametric EB estimator via Gaussian kernels and showed that the estimator achieves asymptotic optimality for both dense and sparse means. Empirically GMLEB outperforms the kernel method by [7]. However, GMLEB is computationally intensive and may not be suitable for data–intensive applications. The connection between compound estimation and convex optimization was pioneered by [23], who cast GMLEB estimation as a convex program. The algorithm is fast and compares favorably to competing methods. However, the nonparametric GEB approach to compound estimation with covariates has not been pursued in the literature.

We show that the EB implementation of NIT essentially boils down to the estimation of the log-gradient of the conditional distribution of Y given $\boldsymbol{S}$. Through a carefully designed reproducing kernel Hilbert space (RKHS) representation of Stein's discrepancy, we recast compound estimation as a convex optimization problem, where the optimal shrinkage factor is found by searching among all feasible score embeddings in the RKHS. The algorithm is computationally fast and scalable, and enjoys superior performance empirically. The kernelized optimization framework provides a rigorous and powerful mathematical interface for theoretical analysis. By appealing to 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 the dimension of covariates on the rate of convergence.

1.4. *Our contributions.* **(1). Methodological contributions.** First, NIT provides an assumption-lean framework for assimilating auxiliary data from multiple sources. Existing works [21, 11, 24] require that the conditional mean function must be specified in the form of $m(\boldsymbol{S}_i) = E(\theta_i|\boldsymbol{S}_i) = \boldsymbol{S}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ are regression coefficients that are usually unknown. [17] derive the minimax rates of convergence for a class of functions $m(\boldsymbol{S}_i)$ but their work is confined to linear shrinkage estimators. By contrast, NIT does not require the specification of any functional relationship and its asymptotic optimality holds for a wider class of prior distributions. Second, NIT is capable of incorporating various types of side information and handling multivariate covariates. By contrast, [41, 1] only focus on the variance or sparsity structure, and both methods can only handle univariate covariates. Third, NIT is fast and scalable, produces stable estimates, and provides a flexible tool for incorporating various structural constraints. Finally, NIT eliminates the needs for defining groups, which avoids information loss in discretization as encountered in [41, 1].

(2). Theoretical contributions. First, we establish the rates at which NIT converges to the oracle integrative Tweedie estimator; this explicit characterization of the improvements in estimation risk not only reveals how much benefits the transfer learning algorithm can provide, but also justifies the claim that NIT avoids negative learning asymptotically. Second, our theory precisely tabulates the deteriorations in the learning rates as the dimension of side information increases. This gives caveats on utilizing high-dimensional auxiliary data. Finally, we develop new analytical tools to formalize the theoretical properties of the optimization framework using kernelized Stein's discrepancy (KSD). The KSD approach has been applied in a host of recent statistical and machine learning problems [28, 10, 43, 29, 30, 2]. The success of KSD based methods critically depends on the conjecture that a lower risk in RKHS norm would translate to a lower L_p risk. However, a general isometry theory between the RKHS and L_p risks does not exist. Our work provides the first rigorous analysis that establishes this isometry (in the context of compound estimation); the probability tools therein can be of independent interest for decision theorists.

1.5. *Organization of the Paper.* The article is organized as follows. In Section 2, we discuss the empirical Bayes estimation framework, NIT estimator and computational algorithms. Section 3 studies the theoretical properties of the NIT estimator. Sections 4 and 5 investigate the performance of NIT using simulated and real data respectively. We conclude the article with a discussion of some open problems. Additional technical details and proofs are provided in the Supplementary Material.

2. Methodology.

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

The transfer learning setup may be conceptualized via a hierarchical model. We assume that the primary and auxiliary data are related through a latent vector $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)^T$:

$$
\begin{aligned}
\theta_i &= g_\theta(\xi_i, \eta_{y,i}), \quad 1 \le i \le n, \\
s_i^{(j)} &= g_j^s(\xi_i, \tilde{\eta}_{j,i}), \quad 1 \le j \le K,
\end{aligned} \tag{2}
$$

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 provides a general framework that can be utilized to incorporate both continuous and discrete auxiliary data into inference.

This section first introduces an oracle rule that optimally borrows information from $\boldsymbol{S}$, then discusses a data-driven rule that emulates the oracle rule.

2.1. *Learning via integrative Tweedie.* Consider an oracle with access to the joint density $f(y|\boldsymbol{s})$. We study the optimal rule that minimizes the Bayes risk. The integrative Tweedie's formula, given by the next proposition, generalizes Tweedie's formula ([33, 12]) from the classical setup to the transfer learning setup.

PROPOSITION 1 (Integrative Tweedie). *Consider the hierarchical model (1) and (2). Let $f(y|\boldsymbol{s})$ be the conditional density of Y given $\boldsymbol{S}$. The 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*

$$
\delta^\pi(y,\boldsymbol{s}) = y + \sigma^2\nabla_y \log f(y|\boldsymbol{s}). \tag{3}
$$

Integrative Tweedie is simple and intuitive, nonetheless it provides a general and flexible framework for transfer learning. First, existing works on shrinkage estimation with side information require that the form of the conditional mean function $m(\boldsymbol{S}_i) = E(Y_i|\boldsymbol{S}_i)$ must be pre-specified [21, 11, 24]. By contrast, integrative Tweedie incorporates the side information via a general function $f(y|\boldsymbol{s})$, which eliminates the need to pre-specify a fixed relationship between Y and $\boldsymbol{S}$. Next, the effectiveness of existing transfer learning algorithms critically depends on the similarity between the target and source distributions, and may suffer from negative learning when the target and source distributions differ. By contrast, integrative Tweedie effectively avoids negative learning; the auxiliary data $\boldsymbol{S}$ are only utilized through $f(y|\boldsymbol{s})$ to assist inference by providing the structural knowledge of Y.

To illustrate the key advantages of the new transfer learning machinery, we present two toy examples, which respectively show that: (a) If the two distributions match perfectly, then integrative Tweedie reduces to an intuitive data averaging strategy; (b) If the two distributions differ, then integrative Tweedie is still effective in reducing the risk.

EXAMPLE 1. Suppose (Y_i, S_i) are conditionally independent given $(\mu_{i,y}, \mu_{i,s})$. We begin by considering a case where S_i is an independent copy of Y_i:

$$\mu_i \sim N(\mu_0, \tau^2), \quad Y_i = \mu_i + \epsilon_i, \quad S_i = \mu_i + \epsilon_i',$$

where $\epsilon_i \sim N(0, \sigma^2)$, $\epsilon_i' \sim N(0, \sigma^2)$. Intuitively, the optimal Bayes estimator is to use $Z_i = (Y_i + S_i)/2 \sim N(\mu_i, \sigma^2/2)$ as the new data point:

$$\hat{\mu}_i^{op} = \frac{\frac{1}{2}\sigma^2\mu_0 + \tau^2 Z_i}{\frac{1}{2}\sigma^2 + \tau^2} = \frac{\sigma^2\mu_0 + \tau^2(Y_i + S_i)}{\sigma^2 + 2\tau^2}. \tag{4}$$

For this bivariate normal model, the conditional distribution of Y_i given S_i is $Y_i|S_i \sim N\left(\sigma^2\mu_0 + \tau^2 S_i/(\tau^2+\sigma^2), \sigma^2(2\tau^2+\sigma^2)/(\tau^2+\sigma^2)\right)$. It follows that $l'(Y_i|S_i) = \sigma^{-2}(2\tau^2 + \sigma^2)^{-1}(\tau^2 S_i + \sigma^2\mu_0) - Y_i(\tau^2+\sigma^2)$. We obtain $\delta_i^\pi = (\sigma^2 + 2\tau^2)^{-1}\{\sigma^2\mu_0 + \tau^2(Y_i + S_i)\}$, recovering the optimal estimator (4). It is important to note that if we slightly alter the model, say perturbating S_i by η_i: $S_i = \mu_i + \eta_i + \epsilon_i'$, or letting $S_i = f(\mu_i) + \epsilon_i'$, then averaging Y and S via (4) may lead to negative learning. However, integrative Tweedie provides a robust data combination approach that always leads to a reduction in estimation risk (Proposition 2).

EXAMPLE 2. Suppose that S_i is a group indicator. The two groups ($S = 1$ and $S = 2$) have equal sample sizes. The primary data obey $Y_i|S_i = k \sim (1-\pi_k)N(0,1) + \pi_k N(\mu_i, 1)$, with $\pi_1 = 0.01$, $\pi_2 = 0.4$ and $\mu_i \sim N(2, 1)$. Consider two oracle Bayes rules $\delta_i^\pi(Y_i, S_i)$ and $\delta_i^\pi(Y_i)$. Some calculations yield

$$\left[B(\delta_i^\pi(Y_i)) - B\{\delta_i^\pi(Y_i, S_i)\}\right]/B\{\delta_i^\pi(Y_i)\} = 0.216,$$

which shows that incorporating S_i can reduce the risk by as much as 21.6%. It is important to note that the distributions of Y_i and S_i are very different (continuous vs. binary), making it difficult for conventional machine learning algorithms to pool information across the target and source domains. Integrative Tweedie is still highly effective in reducing estimation risk by exploiting the grouping structure encoded in S_i.

2.2. *Nonparametric estimation via convex programming.* This section develops a data-driven procedure to emulate the oracle rule. Denote the collection of all data by $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$, where for $i = 1, \ldots, n$, we have $x_{1i} = y_i$ and the summary statistics collected from the kth source domain constitute $x_{ki} = s_i^{(k-1)}$ for $k = 2, \ldots, K+1$. Our goal is to estimate the shrinkage factor

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

Let $K_\lambda(\boldsymbol{x}, \boldsymbol{x}')$ be a kernel function that is integrally strictly positive definite; 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:

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

For a fixed λ, let $\hat{\boldsymbol{h}}_{\lambda,n}$ be the solution to the following quadratic program:

$$\operatorname*{arg\,min}_{\boldsymbol{h} \in \boldsymbol{V}_n} \quad \boldsymbol{h}^T \boldsymbol{K}_\lambda \boldsymbol{h} + 2\boldsymbol{h}^T \nabla \boldsymbol{K}_\lambda \mathbf{1}, \tag{5}$$

where $\boldsymbol{V}_n$ is a convex subset of $\mathbb{R}^n$. We give two remarks.

REMARK 1. *Convex constraints such as linearity and monotonicity, detailed in Section 2.3, are fundamental to the compound decision problem [23]. The constraints imposed via $\boldsymbol{V}_n$ can improve the stability and efficiency of the shrinkage estimator.*

REMARK 2. *The convex program* (5) *is motivated by the kernelized Stein's discrepancy (KSD; [28, 10, 2]), a useful tool in our theoretical analysis formally defined in Section 3. Roughly speaking, the KSD measures how far a given* $\boldsymbol{h}$ *is from the true score* $\boldsymbol{h}_f$*. The KSD is always non-negative, and is equal to 0 if and only if* $\boldsymbol{h} = \boldsymbol{h}_f$*. Hence, solving the convex program* (5) *is equivalent to finding a shrinkage estimator (assisted by side information) that makes the estimation risk as small as possible.*

Proposition 1 motivates us to consider the following class of nonparametric integrative Tweedie (NIT) estimators

$$\left\{\hat{\boldsymbol{\delta}}_\lambda^{\text{IT}} : \lambda \in (0, \infty)\right\}, \qquad \text{where } \hat{\boldsymbol{\delta}}_\lambda^{\text{IT}} = \boldsymbol{y} + \sigma_y^2 \hat{\boldsymbol{h}}_{\lambda,n}. \tag{6}$$

In Section 3, we show that as $n \to \infty$ there exist choices of λ such that $\hat{\boldsymbol{h}}_{\lambda,n}$ can estimate $\boldsymbol{h}_f$ with negligible errors. It follows that the resulting estimator (6) is asymptotically optimal

The NIT estimator (6) marks a clear departure from existing GMLEB methods [19, 23, 15], which cannot be easily extended to handle multivariate auxiliary data. Moreover, NIT has several additional advantages over existing nonparametric GEB methods in both theory and computation. First, in comparison with the GMLEB method [19], the convex program (5) is fast and scalable. Second, [7] proposed to estimate the score function by the ratio $\hat{f}^{(1)}/\hat{f}$, where $\hat{f}$ is a kernel density estimate and $\hat{f}^{(1)}$ is its derivative. By contrast, our direct optimization approach avoids computing ratios and produces more stable and accurate estimates. Third, our convex program can be fine-tuned by selecting a suitable λ. This leads to improved numerical performance and enables a disciplined theoretical analysis compared to the proposal in [23]. Finally, the criterion in (5) can be rigorously analyzed to establish new rates of transfer learning (Sec. 3.1) that are unknown in the literature.

2.3. *Computational details.*

This section provides the following details in computation: (a) how to impose convex constraints; (b) how to construct kernel functions to handle multivariate and possibly correlated covariates; and (c) how to choose the bandwidth λ.

1. Structural constraints. We illustrate how to design appropriate constraints to enforce unbiasedness and monotonicity conditions. The unbiasedness can be achieved by setting $\sum_{i=1}^n h_i = 0$, with corresponding linear constraints given by $\mathbf{1}^T \boldsymbol{h} = 0$. The monotonicity constraints proposed in [23] are highly effective for improving the estimation accuracy. These constraints can be included as $M\boldsymbol{h} \preceq \boldsymbol{a}$. For ease of presentation, assume that $y_1 \le y_2 \le \cdots \le y_n$. To impose the monotonicty constraints $\sigma^2 h_{i-1} - \sigma^2 h_i \le y_i - y_{i-1}$ for all i, we choose M as the following upper triangular matrix $M_{ij} = \sigma^2(I\{i = j\} - I\{i = j-1\})$, and let $\boldsymbol{a}^T = (y_2 - y_1, y_3 - y_2, \cdots, y_n - y_{n-1})$.

2. Kernel functions. The kernel function needs to be carefully constructed to deal with various complications in the multivariate setting, where the auxiliary sequences $\boldsymbol{S}^k$, $1 \le k \le M$, may be correlated and have different 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 $\Sigma_{\boldsymbol{x}}$ is the sample covariance matrix. The RBF kernel is $K_\lambda(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-0.5\lambda^2 \|\boldsymbol{x} - \boldsymbol{x}'\|_{\Sigma_{\boldsymbol{x}}}^2\}$, where λ is the bandwidth whose choice is discussed next. Compared to the Euclidean distance that treats each coordinate equally, Mahalanobis distance is more suitable for combining data collected from heterogeneous sources for it is unitless, scale-invariant and takes into account the correlation in the data. When auxiliary data contain both continuous and categorical variables, we propose to use the generalized Mahalanobis distance [25]. We illustrate the methodology for mixed types of variables in the numerical studies in Section 4.1 (Setting 4), but only pursue theory for the case where both $\boldsymbol{Y}$ and $\boldsymbol{S}$ are continuous.

3. Modified cross-validation (MCV). Implementing the NIT estimator requires selecting the tuning parameters λ. Following [8], we propose to determine λ via modified cross validation (MCV). Let $\eta_i \sim \mathcal{N}(0,\sigma^2)$, $1 \leq i \leq n$ be *i.i.d.* noise variables independent of $\boldsymbol{y}$ and $\boldsymbol{s}_1,\cdots,\boldsymbol{s}_K$. Define $U_i = y_i + \alpha\eta_i$ and $V_i = y_i - \eta_i/\alpha$. Then, U_i and V_i are independent. The MCV uses $\boldsymbol{U} = \{U_1,\cdots,U_n\}$ for constructing estimators $\delta_\lambda^{\mathsf{IT}}(\boldsymbol{U},\boldsymbol{S})$, while using $V_1,\cdots,V_n$ for validation. Consider the validation loss

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

For a small α, the data-driven bandwidth is chosen by minimizing $\hat{L}_n(\lambda,\alpha)$, i.e., $\hat{\lambda} = \arg\min_{\lambda\in\Lambda}\hat{L}_n(\lambda,\alpha)$ for any $\Lambda \subset \mathbb{R}^+$ and the proposed NIT estimator is $\{y_i + \sigma_y^2\hat{\boldsymbol{h}}_{\hat{\lambda},n}(i) : 1 \leq i \leq n\}$. Proposition 3 in Section 3.6 shows that asymptotically the validation loss is close to the true loss, justifying the aforementioned algorithm.

3. Theory.

This section develops large-sample theory for the data-driven NIT estimator. We assume that $\boldsymbol{X}_i = (Y_i, S_{i1},\cdots,S_{ik})^T$, $i = 1,\cdots,n$, are i.i.d. samples from a continuous multivariate density f on $\mathbb{R}^{K+1}$. Denote $\mathbf{X} = (\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)^T$ the $n\times(K+1)$ matrix of observations and $\boldsymbol{X}^{(k)}$ be the kth column of $\mathbf{X}$. Let f be a density function defined on $\mathbb{R}^{K+1}$ and $\mathfrak{h}_f(\boldsymbol{x}) = \nabla_1 \log f(x_1|\boldsymbol{x}_{-1})$ the conditional score function corresponding to the first coordinate, where $\boldsymbol{x}_{-1} = \{x_j : 2 \leq j \leq K+1\}$. It follows from the definition that this conditional score is same as 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 that we aim to estimate $\boldsymbol{\theta} = \mathbb{E}\left\{\boldsymbol{X}^{(1)}\right\}$. The score function $\mathfrak{h}_f(\boldsymbol{x})$, by Proposition 1, plays the key role in the oracle NIT estimator. We next show how $\mathfrak{h}_f(\boldsymbol{x})$ can be well estimated by the solution to (5).

3.1. *Kernelized Stein's discrepancy.*

We first introduce the kernelized Stein's discrepancy. 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 denote two conditional univariate distributions (conditional on a K-dimensional covariate) on $\mathbb{R}^{K+1}$. Denote p and q the respective conditional densities and $\mathfrak{h}_p$ and $\mathfrak{h}_q$ the score functions. The kernelized Stein's discripancy (KSD; [28, 10, 2]) between P and Q is the kernel weighted distance between $\mathfrak{h}_p$ and $\mathfrak{h}_q$:

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

The KSD is closely connected to the maximum mean discrepancy (MMD, [14]). Minimizing KSD is a popular technique in several satistical applications. It has been recently applied to the development of new goodness-of-fit tests [28, 10, 43], bayesian inference procedures [29, 30] and simultaneous estimation methods [2]. Compared to the MMD, the KSD is particularly suitable for empirical Bayes estimation because it can be directly constructed based on the score functions, which by Proposition 1 yield the optimal shrinkage factors. It can be shown that the KSD satisfies

$$D_\lambda(P,Q) \geq 0 \text{ and } D_\lambda(P,Q) = 0 \text{ if and only if } P = Q.$$

The direct evaluation of $D_\lambda(P,Q)$ is difficult. Next we discuss an alternative representation of the KSD that can be easily evaluated empirically. Specifically, for any functional $\mathfrak{h} : \mathbb{R}^{K+1}\to\mathbb{R}$, define the following quadratic functional $\kappa_\lambda[\mathfrak{h}]$ over $\mathbb{R}^{K+1}\times\mathbb{R}^{K+1}$:

$$\kappa_\lambda[\mathfrak{h}](\boldsymbol{u},\boldsymbol{v}) = K_\lambda(\boldsymbol{u},\boldsymbol{v})\mathfrak{h}(\boldsymbol{u})\mathfrak{h}(\boldsymbol{v}) + \nabla_{\boldsymbol{v}}K_\lambda(\boldsymbol{u},\boldsymbol{v})\,\mathfrak{h}(\boldsymbol{u}) + \nabla_{\boldsymbol{u}}K_\lambda(\boldsymbol{u},\boldsymbol{v})\,\mathfrak{h}(\boldsymbol{v}) + \nabla_{\boldsymbol{u},\boldsymbol{v}}K_\lambda(\boldsymbol{u},\boldsymbol{v})\,. \tag{7}$$

The alternative representation of the KSD, given in [28, 10], uses the quadratic function in (7) and only involves the score function of q:

$$D_\lambda(P,Q) = \mathbb{E}_{(\boldsymbol{u},\boldsymbol{v})\overset{\text{i.i.d.}}{\sim}P}\{\kappa_\lambda[\mathfrak{h}_q](\boldsymbol{u},\boldsymbol{v})\}. \tag{8}$$

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

$$\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}) = \boldsymbol{h}^T \boldsymbol{K}_\lambda \boldsymbol{h} + 2\boldsymbol{h}^T \nabla \boldsymbol{K}_\lambda \mathbf{1} + \mathbf{1}^T \nabla^2 \boldsymbol{K}_\lambda \mathbf{1}, \tag{9}$$

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 the bandwith λ and $\nabla \boldsymbol{K}_\lambda$ is defined in Section 2.2. It is easy to see that the convex program (5) is equivalent to minimizing $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h})$ over $\boldsymbol{h}$[1]. Now, note that when $\boldsymbol{h}$ equals $\boldsymbol{h}_q = \{\mathfrak{h}_q(\boldsymbol{x}_1), \dots, \mathfrak{h}_q(\boldsymbol{x}_n)\}$, (9) is just the empirical version of the KSD defined in (8):

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

where $\hat{F}_n$ is the empirical distribution function.

We are now ready to explain the heuristic idea behind the optimization criterion (5). As $n \to \infty$, $\hat{F}_n \to F$ and one can show that $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}_q) \to \mathcal{S}_\lambda(\mathfrak{h}_q) := D_\lambda(F, Q)$. Moreover, $D_\lambda(F, Q) = 0$ iff $F = Q$. Thus, if we could have minimized $S_\lambda(\mathfrak{h}_q)$ over the class $\mathcal{H} = \{\mathfrak{h}_q = \nabla_1 \log q(\boldsymbol{x}) : q \text{ is any density on } \mathbb{R}^{K+1}\}$[2], then the minimum would be achieved at the true score function $\mathfrak{h}_f$ and the minimum value would be 0. However, $\mathcal{S}_\lambda(\mathfrak{h}_q)$ involves the unknown true distribution F, which makes such a direct minimization impossible. Alternatively, we minimize the corresponding sample based criterion $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h}_q)$ in (9) (or equivalently, (5)). 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. The convergence rates of the estimates are established next.

3.2. *Score estimation under the L_p loss.* The next two subsections formulate a rigorous theoretical framework, in the context of compound estimation, to derive the convergence rates of the proposed estimator.

The criterion (5) involves minimizing the V-statistic $\widehat{\mathcal{S}}_{\lambda,n}(\boldsymbol{h})$. Using standard asymptotics results for V-statistics [37], 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 that $\hat{\boldsymbol{h}}_{\lambda,n}$, the solution to (5), satisfies:

$$n^{-2} \sum_{i,j} K_\lambda(\boldsymbol{x}_i, \boldsymbol{x}_j) \left\{\hat{\boldsymbol{h}}_{\lambda,n}(i) - \mathfrak{h}_f(\boldsymbol{x}_i)\right\} \left\{\hat{\boldsymbol{h}}_{\lambda,n}(j) - \mathfrak{h}_f(\boldsymbol{x}_j)\right\} = O_P(n^{-1}) \tag{10}$$

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

While (10) 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 (10) 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 $d_\lambda \le C_1 \ell_2$, where C_1 is a constant. However, the inequality in the other direction does not always hold.

[1] This can be easily seen because the extra term $\mathbf{1}^T \nabla^2 \boldsymbol{K}_\lambda \mathbf{1}$ does not involve $\boldsymbol{h}$.

[2] 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 (5) produce efficient estimates.

We aim to show that $\ell_2 \le C_2\, d_\lambda$ 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.

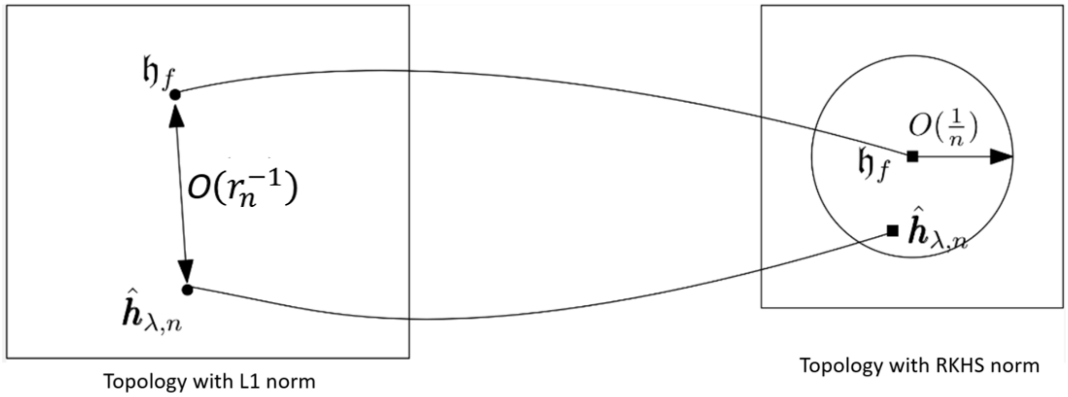


Fig 1: Schematic illustrating the relation between the RKHS and ℓ_1 risks of $\hat{\boldsymbol{h}}_{\lambda,n}$. The (approximate) isometry can be still established. However, the error rate is increased from n^{-1} to r_n^{-1} due to inversion.

In Sections 3.3 and 3.5, we show that as $n \to \infty$ then $\ell_2(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f) \le c_{\lambda,n}\, d_\lambda(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f)\{1 + o_P(1)\}$, for some $c_{\lambda,n}$ that depends on λ and n only. This result, coupled with the convergence in the RKHS norm (10), produces a tolerance bound on the L_p risk of $\hat{\boldsymbol{h}}_{\lambda,n}$, which is 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 (10) 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. [2] and Section 5.1 of Liu, Lee and Jordan [28]. 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 (10) 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.

Our main result is concerned with the ℓ_1 risk of the solution from (5). 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 $c_1 a_n \le b_n \le c_2 b_n$ for all large n and some constants $c_2 \ge c_1 > 0$.

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

$$r_n \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |\hat{\boldsymbol{h}}_{\lambda,n}(i) - \mathfrak{h}_f(\boldsymbol{x}_i)| \right) \to 0 \ \text{ in } L_1, \tag{11}$$

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

REMARK 3. *It follows immediately from Theorem 1 that the deviations between our proposed estimate and the oracle estimator in (3) obeys:*

$$r_n \left(\frac{1}{n} \sum_{i=1}^{n} \left| \hat{\boldsymbol{\delta}}_{\lambda}^{\mathsf{IT}}(i) - \delta_i^{\pi}(y_i|\boldsymbol{s}_i) \right| \right) \to 0 \text{ in } L_1 \text{ as } n \to \infty. \tag{12}$$

Under the classical setting with no auxiliary data ($K = 0$), we achieve the traditional $\sqrt{n}-$rate as established in Jiang and Zhang [19]. In this context, the rates are sharp.

We can see that, barring the poly-log terms, the convergence rate $r_n = n^{1/(K+2)}$ decreases as K increases. We provide further discussions on this attribute, 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 (10) 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:

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

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 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 2 below shows that this difference is asymptotically negligible.

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

$$\mathcal{L}_n^2(\hat{\boldsymbol{\delta}}_\lambda^{\mathsf{IT}}, \boldsymbol{\theta}) - \mathcal{L}_n^2(\boldsymbol{\delta}^\pi, \boldsymbol{\theta}) = o_p(r_n^{-1}) \text{ as } n \to \infty. \tag{13}$$

Combining (12) and (13), 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 [6] (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.

PROPOSITION 2. *Consider hierarchical model (1)–(2). 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*

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

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

In Theorem 1 and Lemma 2, the rate of convergence r_n decreases as K increases. In light of Proposition 2, this means that although theoretically we never lose by adding more columns of auxiliary data (even if they are non-informative), there is still a tradeoff under our estimation framework. The increase of K leads to a widened gap between the oracle and data-driven rules, which may offset the benefits of incorporating more side information. The following numerical example illustrates two aspects of the phenomenon.

Consider the hierarchical model (1)–(2). 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 $K \geq 5$.

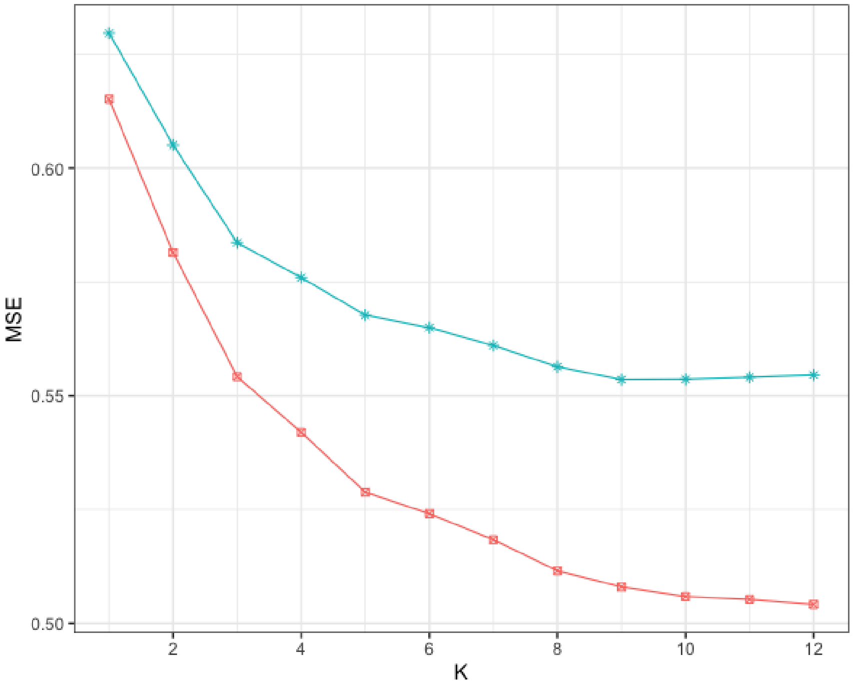


Fig 2: Mean squared error of our proposed method (in magenta) is plotted along with the oracle risk (in sky blue) as the number of auxiliary variable (K) increases. The MSE of the oracle procedure always decreases but the MSE of the data-driven NIT procedure stops decreasing as $K \geq 9$.

3.5. *Convergence rates for heavy-tail densities.* We extend the results in Section 3.3 to a wider class of prior distributions. To rule out cases where $||h_f||_2$ is negligible (such as uniform prior), we consider the following mild assumption on the prior that rules out densities with tail behaviors heavier than Cauchy.

Assumption 2: The class of priors $\pi(\theta)$ satisfy: $\theta^2\pi(\theta)$ is bounded for all θ.

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 3. *Under Assumption 2, with $\lambda \asymp n^{-1/(K+2)}$ and $r_n = n^{1/(3(K+2))}$, we have*

$$r_n \cdot \left(\frac{1}{n}\sum_{i=1}^{n} |\hat{\boldsymbol{h}}_{\lambda,n}(i) - \mathfrak{h}_f(\boldsymbol{x}_i)|\right) \to 0 \ \ in\ L_1\ as\ n \to \infty.$$

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 $n \to \infty$.

3.6. *Consistency of the MCV criterion.* In Sections 3.3 and 3.5, we have established asymptotic risk proporties of our proposed method as bandwidth $\lambda \to 0$. For finite sample sizes, it is important to select the "best" 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 (5) *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 4.2); (c) *multiple* auxiliary sequences that share a common structure with the primary sequence (Section 4.3). We conduct simulations to compare the following methods:

- James-Stein estimator (JS).
- The empirical Bayes Tweedie (EBT) estimator implemented using kernel smoothing as described in [7].
- Non-parametric maximum likelihood estimator (NPMLE), which implements Tweedie's formula using the convex optimization approach [23]. The method is implemented by the R-package "REBayes" in [22].
- Empirical Bayes with cross-fitting (EBCF) by [17].
- The oracle NIT procedure (3) with known $f(y|\boldsymbol{s})$ (NIT.OR).
- The data-driven NIT procedure (6) 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. The MSE of OR is provided as the optimal benchmark for assessing the efficiency of various methods.

In the implementation of NIT.DD, we utilize the generalized Mahalanobis distance, discussed in Section 2.3, to compute the RBF kernel with bandwidth λ. A data-driven choice of the tuning parameter λ is obtained by first solving optimization problems 5 over a grid of λ values and then computing the corresponding modified cross-validation (MCV) loss. We choose the λ with minimum MCV loss as the data-driven bandwidth.

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

$$\zeta_i \sim \mathcal{N}(\xi_i, \sigma^2), \quad S_i \sim \mathcal{N}(\zeta_i, \sigma_s^2).$$

The above data generating mechanism is a special case of the hierarchical model (2). 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.

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 $\boldsymbol{\xi} = (\xi_i : 1 \le i \le n)$ 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: $\theta_i \sim \mathcal{N}(2\xi_i, 0.25), \quad y_i \sim \mathcal{N}(\theta_i, 1)$. 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 data simulated according to the above models and compute the MSEs using 100 replications. The simulation results for Settings 1-4 are displayed in Figure 3. We summarize some important patterns of the plots and provide interpretations.

(a). The integrative methods (NIT.DD, EBCF) outperform univariate methods (JS, NPMLE, EBT) that do not utilize side information in most settings. NIT.DD uniformly dominates EBCF. The efficiency gain is substantial in many settings.
(b). Settings 1-2 shows that the efficiency gain of the integrative methods decreases when σ and σ_s increase (e.g. the auxiliary data become less informative or noisier).
(c). Setting 3 shows that the sample size has big impacts on integrative empirical Bayes estimation. A large sample size is essential for effectively incorporating the side information. The EBCF may underperform univariate methods when n is small.
(d). The gap between NIT.OR and NIT.DD narrows when n increases.
(e). Setting 4 shows that the side information can be highly informative even when the types of primary and auxiliary data do not match.

4.2. *Simulation 2: Integrative estimation in two-sample inference of sparse means.* This section considers compound estimation in two-sample inference. Let X_{1i} and X_{2i} be two Gaussian random variables. Denote $\mu_{1i} = \mathbb{E}(X_{1i})$ and $\mu_{2i} = \mathbb{E}(X_{2i})$, $1 \le i \le n$. Suppose we are interested in estimating the differences $\boldsymbol{\theta} = \{\mu_{1i} - \mu_{2i} : 1 \le i \le n\}$. The primary statistic is given by $\boldsymbol{Y} = \{X_{1i} - X_{2i} : 1 \le i \le n\}$. However, it is argued in [9] that the primary statistic $\boldsymbol{Y}$ is *not* a sufficient statistic. Consider the case where both $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are individually sparse. Then an important fact is that the union support $\mathcal{U} = \{i : \mu_{1i} \neq 0 \text{ or } \mu_{2i} \neq 0\}$ is also sparse. The intuition is that the sparsity structure of $\boldsymbol{\theta}$ is captured by an auxiliary parameter $\boldsymbol{\eta} = \{\mu_{1i} + \mu_{2i} : 1 \le i \le n\}$. Our idea is to construct an auxiliary sequence $\boldsymbol{S} = \{X_{1i} + X_{2i} : 1 \le i \le n\}$ and incorporate $\boldsymbol{S}$ into inference to improve the efficiency[3].

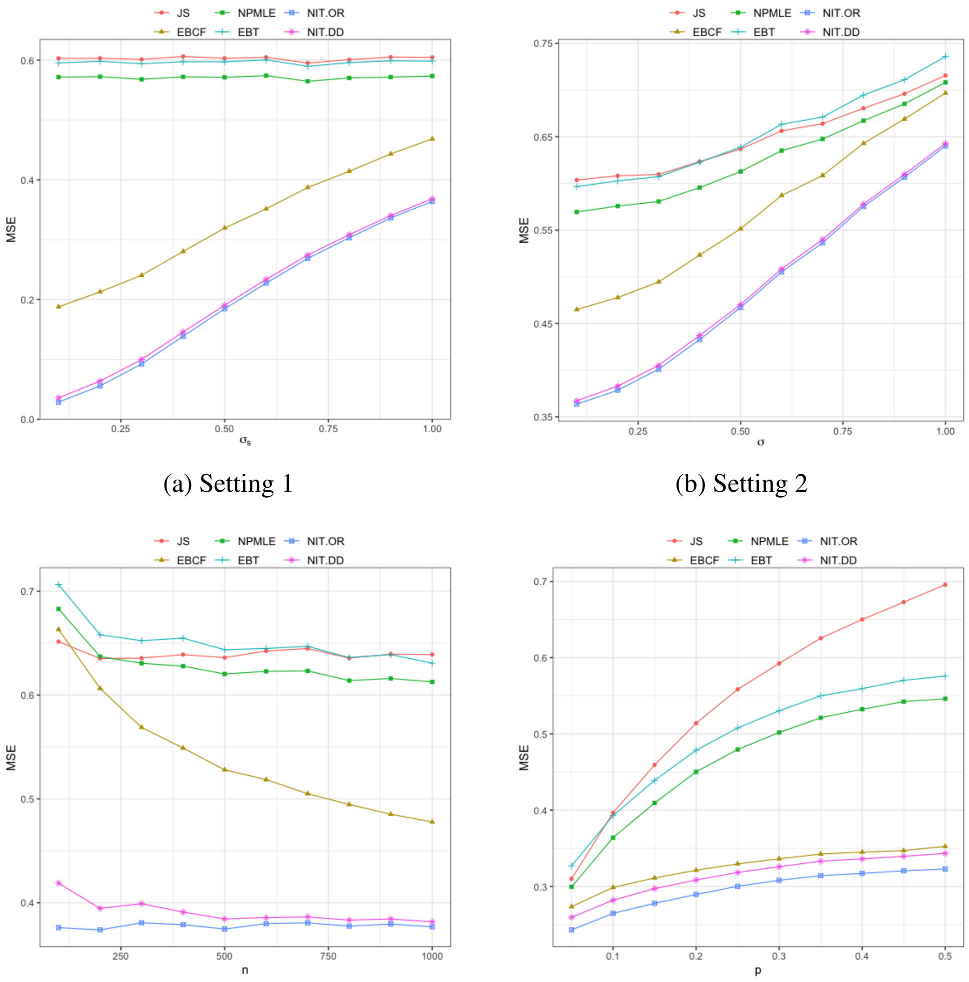
To illustrate the effectiveness of the integrative estimation strategy, we simulate data according to the following two settings and obtain primary and auxiliary data as $\boldsymbol{Y} = \{X_{1i} - X_{2i} : 1 \le i \le n$ and $\boldsymbol{S} = \{X_{1i} + X_{2i} : 1 \le i \le n\}$.

Setting 1: X_{1i} and X_{2i} are generated from $X_{1i} \sim \mathcal{N}(\mu_{1i}, 1)$ and $X_{2i} \sim \mathcal{N}(\mu_{2i}, 1)$, where

$$\begin{aligned} \boldsymbol{\mu}_1[1:k] &= 2.5, & \boldsymbol{\mu}_2[1:k] &= 1 \\ \boldsymbol{\mu}_1[k+1:2k] &= 1, & \boldsymbol{\mu}_2[k+1:2k] &= 1 \\ \boldsymbol{\mu}_1[2k+1:n] &= 0, & \boldsymbol{\mu}_2[2k+1:n] &= 0 \end{aligned}$$

The sparsity level of $\boldsymbol{\theta}$ is controlled by k. We fix $n = 1000$ and vary k from 50 to 450 to investigate the impact of sparsity level on the efficiency of different methods.

[3] It can be shown that $\{(X_{1i} - X_{2i}, X_{1i} + X_{2i}) : 1 \le i \le n\}$ is minimal sufficient and retains all information about $\boldsymbol{\theta}$.



(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Fig 3: Simulation results for one given auxiliary sequence.

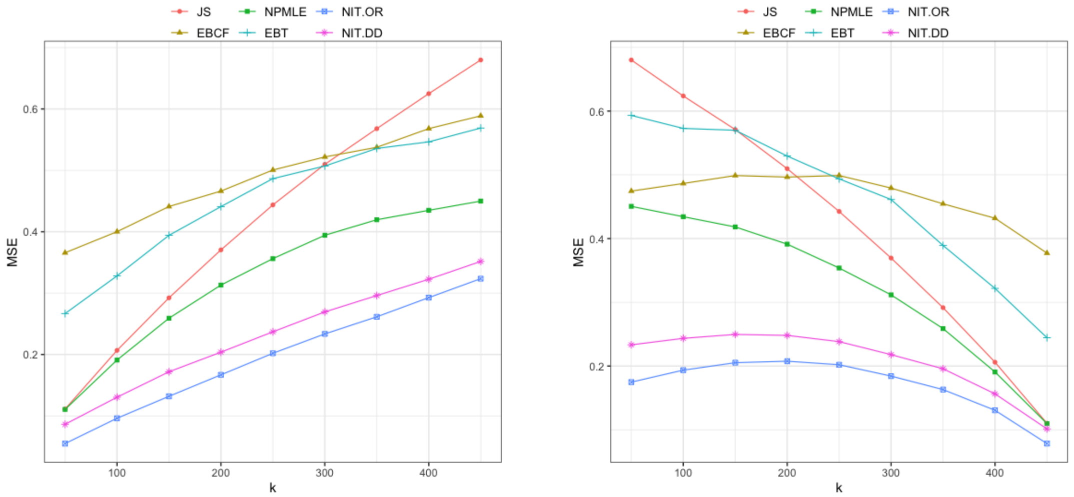
Setting 2: X_{1i} and X_{2i} are generated from $X_{1i} \sim \mathcal{N}(\mu_{1i}, 1)$ and $X_{2i} \sim \mathcal{N}(\mu_{2i}, 1)$, where

$$\begin{aligned} \boldsymbol{\mu}_1[1:k] &= 1, & \boldsymbol{\mu}_2[1:k] &= 1 \\ \boldsymbol{\mu}_1[k+1:500] &= 2.5, & \boldsymbol{\mu}_2[k+1:500] &= 1 \\ \boldsymbol{\mu}_1[501:n] &= 0, & \boldsymbol{\mu}_2[501:n] &= 0 \end{aligned}$$

The primary parameter $\boldsymbol{\theta}$ becomes more sparse when k increases. We fix $n = 1000$ and vary k from 50 to 450 to investigate the efficiency gain of NIT.

We apply different methods to simulated data and calculate the MSEs using 100 replications. The simulation results are displayed in Figure 4. The following can be observed.

(a). The side information provided by the auxiliary sequence can be highly informative for reducing the estimation risk. Our proposed methods (NIT.DD, NIT.OR) have smaller MSEs than competing methods (EBCF, JS, NPMLE, EBT). The efficiency gain over univariate methods (JS, EBT, NPMLE) is more pronounced when signals become more sparse.



(a) Setting 1

(b) Setting 2

Fig 4: Two-sample inference of sparse means.

(b). EBCF is dominated by NIT, and can be inferior to univariate shrinkage methods.

(c). The class of linear estimators is inefficient under the sparse setting. For example, the NPMLE method dominates the JS estimator, and the efficiency gain increases when the signals become more sparse.

4.3. *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:

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

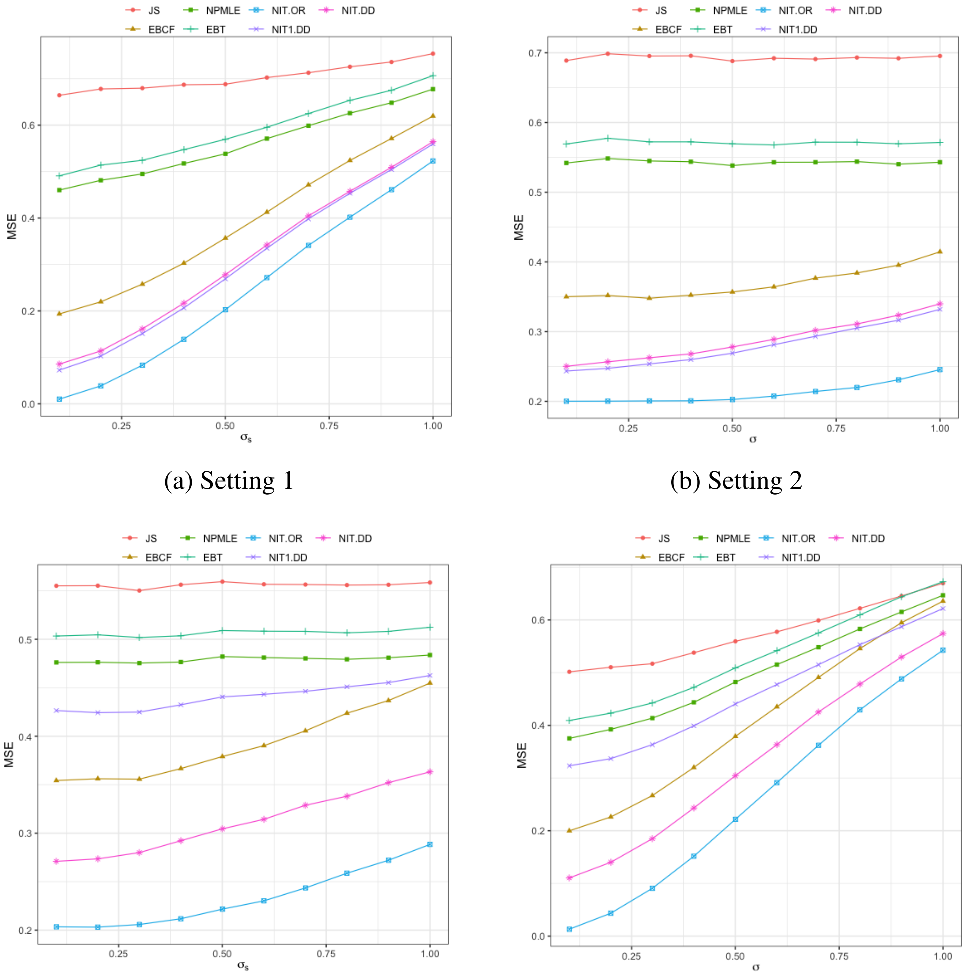
There can be various ways to incorporate auxiliary data from multiple sources. We consider, in addition to NIT.DD that utilizes all sequences, an alternative strategy that involves firstly constructing a new auxiliary sequence $\bar{\boldsymbol{S}} = \frac{1}{4}(\boldsymbol{S}^1 + \boldsymbol{S}^2 + \boldsymbol{S}^3 + \boldsymbol{S}^4)$ 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:

$$Y_i = \theta_i^Y + \epsilon_i^Y, \quad S_i^j = \theta_i^j + \epsilon_i^j, \tag{14}$$

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.



(a) Setting 1

(b) Setting 2

(c) Setting 3

(d) Setting 4

Fig 5: Integrative estimation with multiple auxiliary sequences.

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 (14). However, we now allow $\boldsymbol{\theta}^j$ to have different structures across j. Specifically, let

$$\boldsymbol{\eta}^1[1:500] = \boldsymbol{\eta}[1:500], \quad \boldsymbol{\eta}^1[501:n] = 0, \quad \boldsymbol{\eta}^2[1:500] = 0, \quad \boldsymbol{\eta}^2[501:n] = \boldsymbol{\eta}[501:n].$$

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 summarized the results in Figure 3. We make the following remarks.

(a). The univariate methods (JS, NPMLE, EBT) are dominated by the integrative methods (NIT.DD, NIT.OR, EBCF, NIT1.DD). The efficiency gain is more pronounced when σ and σ_s are small.

(b). EBCF is dominated by NIT.DD. Compared to the setting with one auxiliary sequence, the gap between the performances of NIT.OR and NIT.DD has widened because of the increased complexity of the estimation problem in higher dimensions.

(c). 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.

(d). In Settings 3-4, the average $\bar{\boldsymbol{S}}$ does not provide an effective way to combine the information in auxiliary data. Since the last two sequences are not useful, such a data reduction step leads to substantial information loss. NIT1.DD still outperforms univariate methods but is much worse than EBCF and NIT.DD.

The simulation results show that it is potentially beneficial to reduce the dimension of auxiliary data. However, 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 for extracting structural information from a large number of auxiliary sequences.

5. Applications. This section compares NIT and its competitors on gene expression data and monthly sales data.

5.1. *Integrative Non-parametric estimation of Gene Expressions.* We consider the data set in [36] 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 [44]. 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 S.7.1 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 6 shows the heatmap of the primary, the auxiliary and the validation sequences. We implemented the following estimators (a) the modified James-Stein (JS) following [42], (b) Non-parametric Tweedie estimator without auxiliary information, (c) Empirical Bayes with cross-fitting (EBCF) by [17] and the Non-parametric Integated Tweedie (NIT) with auxiliary infomation: (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 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. Figure 6B 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 supplement fig S.1). These genes impact 35 biological processes and 12 molecular functions in human cells (see supplement fig S.1 bottom two panels); this implies that important inferential gains can be made by using auxiliary information via our proposed NIT estimator.

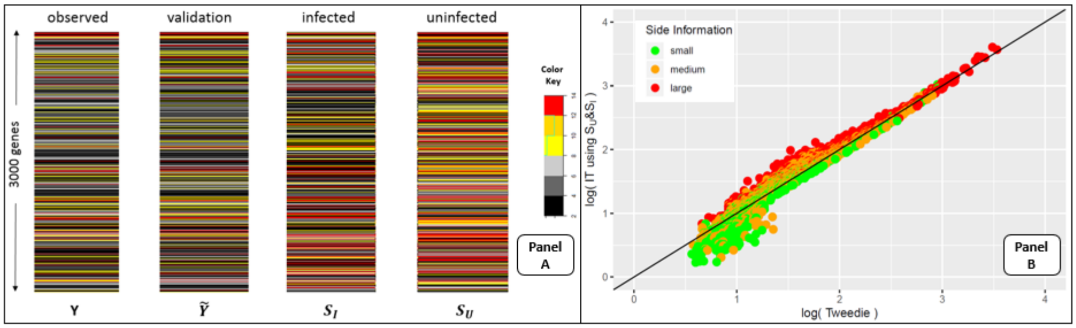


Fig 6: Panel A: Heatmaps of the gene expression datasets showing the four expression vectors corresponding to the observed, validation and auxiliary sequences. Panel B: scatterplot of the effect size estimates of gene expressions based on Tweedie and NIT (using both S_{U} and S_{I}). Magnitude of the auxiliary variables used in the NIT estimate is reflected by different colors.

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

5.2. *Leveraging auxiliary information in predicting monthly sales.* We consider the total monthly sales of beers across $n = 866$ stores of a retail grocery chain. These stores are spread across different states in the USA (see Figure S.2 in the Supplement). The data is extracted from [5], which has been widely studied for inventory management and consumer preference analyses; see also [4] and the references therein.

Let $\boldsymbol{Y}^t$ be the n dimensional vector denoting the monthly sales of beer across the n stores in month $t \in \{1, \dots, 12\}$. For inventory planning, it is economically important to estimate future demand. In this context, we consider estimating the monthly demand vector (across stores) for month t using the previous month's sales $\boldsymbol{Y}^{t-1}$. We use the first six months $t = 1, \dots, 6$ for estimating store demand variabilities $\hat{\sigma}_i^2, i = 1, \dots, n$. For $t = 7, \dots, 12$, using estimators based on month t's sales, we calculate their demand prediction error for month $t+1$ by using its monthly sale data for validation. Among the estimators, we introspect the modified James-Stein (JS) estimator of Xie, Kou and Brown [42]:

$$\hat{\boldsymbol{\theta}}_i^{t+1}[\mathsf{JS}] = \widehat{\mathsf{JS}}_i^t + \left[1 - \frac{n-3}{\sum_i \hat{\sigma}_i^{-2}(Y_i^t - \widehat{\mathsf{JS}}_i^t)^2}\right]_+ (Y_i^t - \widehat{\mathsf{JS}}_i^t) \text{ where } \widehat{\mathsf{JS}}_i^t = \frac{\sum_{i=1}^n \sigma_i^{-2} Y_i^t}{\sum_{i=1}^n \sigma_i^{-2}},$$

as well as the Tweedie (T) estimator $\hat{\boldsymbol{\theta}}_i^{t+1}[\mathsf{T}] = Y_i^t + \hat{\sigma}_i \hat{h}_i$ where $\hat{h}_i$ are estimates of $\nabla_1 \log f(\hat{\sigma}_i^{-1} Y_i^t)$ based on the marginal density of standardized sales. We also consider the sales of three other products: milk, deodorant and hotdog from these stores. They are not directly related to the sale of beers but they might contain possibly useful information regarding consumer preferences to beers particularly as they share zip-code and other store specific responses. We use them as auxiliary sequences in our NIT methodology. Figure 7 shows the distribution of beer sales (across stores) for different months and the pairwise distribution of the sales of different products. Further details about the dataset is provided in Section S.7.2 of the Supplement.

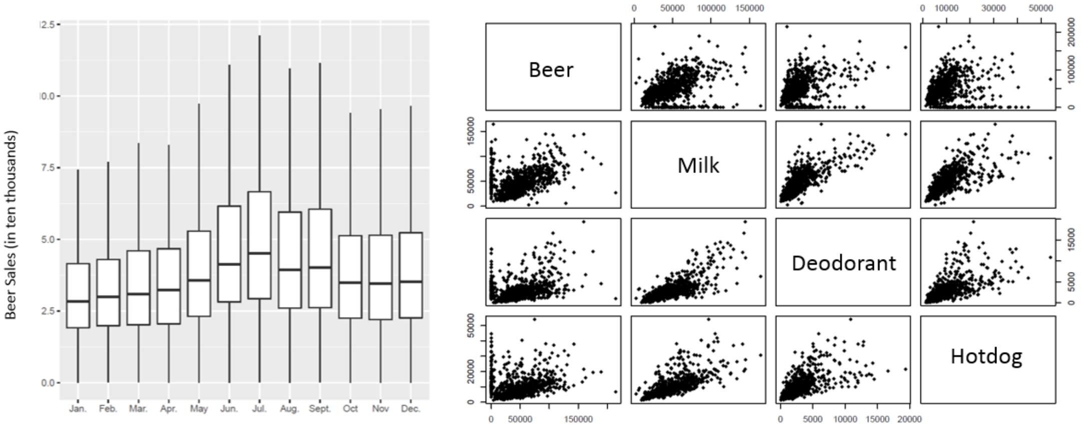


Fig 7: Distribution of monthly sales of beer across stores (on left) and the pairwise distribution of joint sales of different products in the month of July (in right).

In Table 2, we report the average % gain in predictive error by the James-Stein (JS), Tweedie (T) and integrative Tweedie (IT) estimators (using different combinations of auxiliary sequences) over the naive estimator $\hat{\delta}^{t+1,\text{naive}} = \boldsymbol{Y}^t$ for the demand prediction problem at $t = 7, \ldots, 12$. The detailed month-wise gains and the loss function are provided in the supplement. Using auxiliary variables via our proposed NIT framework yields significant additional gains over non-integrative methods. However, the improvement slackens as an increasing number of auxiliary sequences are incorporated. It is to be noted that the demand data set is highly complex and heterogeneous and $n = 866$ may not be adequately large for conducting successful non-parametric estimation. Hence suitably anchored parametric JS estimator produces better prediction than non-parmatric Tweedie. Also, as demonstrated in Table S.2 of the Supplement, there are months where shrinkage estimation methods do not yield positive gains. Nonetheless, the NIT estimator produces significant advantages over competing methods. It produces on average 7.7% gain over unshrunken methods and attains an additional 3.7% gain over non-parametric shrinkage methods.

TABLE 2

Average % gains over the naive unshrunken estimator for monthly beer sales prediction

JS	Tweedie	IT-Milk	IT-Deodorant	IT-Hotdog	IT-M&D	IT-M&H	IT-D&H	IT-M&D&H
5.7	4.0	6.0	7.1	6.8	6.1	6.6	7.5	7.7

6. Discussion. NIT, inspired by classical empirical Bayes ideas, provides a new framework for transferring useful structural knowledge from related sources domains to assist the estimation of a high-dimensional parameter in the target domain. The framework avoids negative learning because no distributional assumptions are imposed on auxiliary data $\boldsymbol{S}$, which are allowed to be categorical, numerical or of mixed type. The auxiliary data are only used to provide structural knowledge of the high-dimensional parameter in the target domain.

Our theory tabulates the reductions in estimation errors and deteriorations in the learning rates as the dimension of $\boldsymbol{S}$ increases. This indicates that if we have a large number of variables as potential choices for auxiliary data, it would be beneficial to first conduct data reduction before applying the NIT estimator. The loss of information resulted from the data reduction process can possibly be compensated by the increased precision in the optimization process. However, our simulation results in Section 4.3 show that there can be significant information loss if the data reduction step is carried out improperly. Our findings suggest two directions for future research: (a) the investigation of the tradeoff, as K increases, between the achievable error limit of the oracle rule and the decreased convergence rate of the data-driven rule, and (b) the development of principled structure-preserving dimension reduction methods under the transfer learning framework for extracting useful structural information from a large number of auxiliary sequences.

REFERENCES

[1] BANERJEE, T., MUKHERJEE, G. and SUN, W. (2020). Adaptive sparse estimation with side information. *Journal of the American Statistical Association* **115** 2053-2067.

[2] BANERJEE, T., LIU, Q., MUKHERJEE, G. and SUN, W. (2021). A General Framework for Empirical Bayes Estimation in Discrete Linear Exponential Family. *Journal of Machine Learning Research* **22** 1-46.

[3] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* **57** 289–300. MR1325392 (96d:62143)

[4] BRONNENBERG, B. J., DUBÉ, J.-P. H. and GENTZKOW, M. (2012). The evolution of brand preferences: Evidence from consumer migration. *American Economic Review* **102** 2472–2508.

[5] BRONNENBERG, B. J., KRUGER, M. W. and MELA, C. F. (2008). Database paper—The IRI marketing data set. *Marketing science* **27** 745–748.

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

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

[8] BROWN, L. D., GREENSHTEIN, E. and RITOV, Y. (2013). The Poisson compound decision problem revisited. *Journal of the American Statistical Association* **108** 741–749.

[9] CAI, T. T., SUN, W. and WANG, W. (2019). CARS: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *J. Roy. Statist. Soc. B* **81** 187–234.

[10] CHWIALKOWSKI, K., STRATHMANN, H. and GRETTON, A. (2016). A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings.

[11] COHEN, N., GREENSHTEIN, E. and RITOV, Y. (2013). Empirical Bayes in the presence of explanatory variables. *Statistica Sinica* 333–357.

[12] EFRON, B. (2011). Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association* **106** 1602–1614.

[13] EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

[14] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13** 723–773.

[15] GU, J. and KOENKER, R. (2017). Unobserved heterogeneity in income dynamics: An empirical Bayes perspective. *Journal of Business & Economic Statistics* **35** 1–16.

[16] IGNATIADIS, N. and HUBER, W. (2020). Covariate powered cross-weighted multiple testing. *arXiv preprint arXiv:1701.05179*.

[17] IGNATIADIS, N. and WAGER, S. (2019). Covariate-Powered Empirical Bayes Estimation. In *Advances in Neural Information Processing Systems* 9617–9629.

[18] IGNATIADIS, N., SAHA, S., SUN, D. L. and MURALIDHARAN, O. (2019). Empirical Bayes mean estimation with nonparametric errors via order statistic regression. *arXiv preprint arXiv:1911.05970*.

[19] JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics* **37** 1647–1684.

[20] JING, B.-Y., LI, Z., PAN, G. and ZHOU, W. (2016). On sure-type double shrinkage estimation. *Journal of the American Statistical Association* **111** 1696–1704.

[21] KE, T., JIN, J. and FAN, J. (2014). Covariance assisted screening and estimation. *Annals of statistics* **42** 2202-2242.

[22] KOENKER, R. and GU, J. (2017). REBayes: An R Package for Empirical Bayes Mixture Methods. *Journal of Statistical Software* **82** 1–26.

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

[24] KOU, S. and YANG, J. J. (2017). Optimal shrinkage estimation in heteroscedastic hierarchical linear models. In *Big and Complex Data Analysis* 249–284. Springer.

[25] KRUSIŃSKA, E. (1987). A valuation of state of object based on weighted Mahalanobis distance. *Pattern Recognition* **20** 413–418.

[26] LEI, L. and FITHIAN, W. (2018). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 649–679.

[27] LI, A. and BARBER, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81** 45–74.

[28] LIU, Q., LEE, J. and JORDAN, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. *International conference on machine learning* 276–284.

[29] LIU, Q. and WANG, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems* 2378–2386.

[30] OATES, C. J., GIROLAMI, M. and CHOPIN, N. (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 695–718.

[31] REN, Z. and CANDÈS, E. (2020). Knockoffs with side information. *arXiv preprint arXiv:2001.07835*.

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

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

[34] ROEDER, K. and WASSERMAN, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics* **24** 398.

[35] SAHA, S. and GUNTUBOYINA, A. (2020). On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Annals of Statistics* **48** 738–762.

[36] SEN, N., SUNG, P., PANDA, A. and ARVIN, A. M. (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** e01151–18.

[37] SERFLING, R. J. (2009). *Approximation Theorems of Mathematical Statistics*. *Wiley Series in Probability and Statistics*. Wiley.

[38] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution Technical Report, STANFORD UNIVERSITY STANFORD United States.

[39] SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

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

[41] WEINSTEIN, A., MA, Z., BROWN, L. D. and ZHANG, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *Journal of the American Statistical Association* **113** 698–710.

[42] XIE, X., KOU, S. and BROWN, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association* **107** 1465–1479.

[43] YANG, J., LIU, Q., RAO, V. and NEVILLE, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning* 5557–5566.

[44] ZERBONI, L., SEN, N., OLIVER, S. L. and ARVIN, A. M. (2014). Molecular mechanisms of varicella zoster virus pathogenesis. *Nature reviews microbiology* **12** 197–210.

[45] ZHANG, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statistica Sinica* **7** 181–193.

[46] ZHANG, X. and BHATTACHARYA, A. (2017). Empirical bayes, sure and sparse normal mean models. *arXiv preprint arXiv:1702.05195*.

Supplementary Material for "Transfer Learning for Empirical Bayes Estimation: a Nonparametric Integrative Tweedie Approach"

Jiajun Luo, Gourab Mukherjee and Wenguang Sun

University of Southern California

In Sections S.1-S.6 of this supplement we present the proofs of all results stated in the main paper. The proofs are presented in the order the results appear in the main paper. The equations and results that only appear in the supplement but not in the main paper are prefixed by S. We also provide further details regarding the real data examples Section S.7.

S.1 Proof of Proposition 1

The idea of the proof follows from Brown [1971]; we provide it here for compeleteness.

Noting that $f(y, \boldsymbol{s}) = \int f(y, \boldsymbol{s}|\theta) dh_\theta(\theta)$ and $f(y, \boldsymbol{s}|\theta) = f(y|\boldsymbol{s}, \theta) f(\boldsymbol{s}|\theta)$, expand the partial derivative of $f(y, \boldsymbol{s})$:

$$\begin{aligned}\nabla_y f(y, \boldsymbol{s}) &= \sigma^{-2}\Big(\int \theta f(y|\boldsymbol{s}, \theta) f(\boldsymbol{s}|\theta) dh_\theta(\theta) - y \int f(y|\boldsymbol{s}, \theta) f(\boldsymbol{s}|\theta) dh_\theta(\theta)\Big) \\ &= \sigma^{-2}\Big(\int \theta f(y, \boldsymbol{s}|\theta) dh_\theta(\theta) - y f(y, \boldsymbol{s})\Big)\end{aligned}$$

Then, left-multiplying by σ^2 and dividing by $f(y, \boldsymbol{s})$ on both sides, it follows that

$$\sigma^2 \frac{\nabla_y f(y, \boldsymbol{s})}{f(y, \boldsymbol{s})} = \frac{\int \theta f(y, \boldsymbol{s}|\theta) dh_\theta(\theta)}{f(y, \boldsymbol{s}|\theta)} - y$$

Under square error loss, the posterior mean minimizes the Bayes risk. And so, the Bayes estimator is given by

$$\mathbb{E}(\theta|y, \boldsymbol{s}) = \frac{\int \theta f(y, \boldsymbol{s}|\theta) dh_\theta(\theta)}{f(y, \boldsymbol{s})} = y + \sigma^2 \frac{\nabla_y f(y, \boldsymbol{s})}{f(y, \boldsymbol{s})} \ ,$$

where, the second equality follows from the above two displays.

S.2 Proof of Theorem 1

First note that the expected value of the concerned ℓ_p distance

$$\ell_p(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f) = n^{-1} \sum_{i=1}^{n} |\hat{\boldsymbol{h}}_{\lambda,n}(i) - \mathfrak{h}_f(\boldsymbol{x}_i)|^p$$

is given by $\Delta^{(p)}_{\lambda,n}(f) = \mathbb{E}_{\boldsymbol{X}}\{\ell_p(\hat{\boldsymbol{h}}_{\lambda,n}, \mathfrak{h}_f)\}$ where, the expected value is over $\boldsymbol{X}_n = (\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_n)$ where $\boldsymbol{x}_i$s are i.i.d. from f. Thus,

$$\Delta^{(p)}_{\lambda,n}(f) = \mathbb{E}\, |\hat{\boldsymbol{h}}_{\lambda,n}(1) - \mathfrak{h}_f(\boldsymbol{x}_1)|^p = \mathbb{E}|\hat{\boldsymbol{h}}_{\lambda}[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)|^p \, .$$

For notational ease, we would often keep the dependence on f in $\Delta^{(p)}_{\lambda,n}(f)$ implicit. The proof involves upper and lower bounding $\Delta^{(2)}_{\lambda,n}$ by the functionals involving $\Delta^{(1)}_{\lambda,n}$. The upper bound is provided below in (S.3). The lower bound follows from (S.5), whose proof is quite convoluted and is presented separately in Lemma S.2.1.

As the marginal density of the $\boldsymbol{\theta}$ is the convolution with a Gaussian distribution, it follows that there exists some constant $C \geq 0$ such that

$$|\mathfrak{h}_f(\boldsymbol{x}_1)|/\|\boldsymbol{x}_1\|_2 \leq C \text{ for all large } ||\boldsymbol{x}_1||_2.$$

and $|\hat{\boldsymbol{h}}_{\lambda}[\boldsymbol{X}_n](\boldsymbol{x}_1)| = O(\|\boldsymbol{x}_1\|_2)$. With out loss of generality we include such constraints on $\boldsymbol{h}$ in the convex program to solve (5) and so, $|\hat{\boldsymbol{h}}_{\lambda}[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)|$ is also bounded by $O(\|\boldsymbol{x}_1\|_2)$.

Using this property of the score estimates, we have the following bound for all $\boldsymbol{x}_1$ satisfying $\{\boldsymbol{x}_1 : \|\boldsymbol{x}_1\|_2 \leq 2\gamma \log n\}$:

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_{\lambda}[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 \leq 2\gamma \log n\right\}\right] \leq 2\gamma \log(n)\, \Delta^{(1)}_{\lambda,n}. \tag{S.1}$$

On the set $\{\|\boldsymbol{x}_1\|_2 > 2\gamma \log n\}$, again using the aforementioned property of score estimates from (5) we note that

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_{\lambda}[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 > 2\gamma \log n\right\}\right] \lesssim \mathbb{E}\left[\|\boldsymbol{x}_1\|_2^2 I\{\|\boldsymbol{x}_1\|_2 > 2\gamma \log n\}\right] \;, \tag{S.2}$$

where, for any two sequences a_n, b_n, we use the notation $a_n \lesssim b_n$ to denote $a_n/b_n = O(1)$ as $n \to \infty$.

Now, as $\boldsymbol{x}_1$ satisfies assumption 1, the right hand side (S.2) is bounded by $O(n^{-1})$. Combining (S.1) and (S.2) we have the following upper bound on $\Delta^{(2)}_{\lambda,n}$:

$$\Delta^{(2)}_{\lambda,n} \lesssim \log(n)\, \Delta^{(1)}_{\lambda,n} + n^{-1} \, . \tag{S.3}$$

For the lower bound on $\Delta^{(2)}_{\lambda,n}$ consider the following intermediate quantity which is related to the KSD norm d_λ on the score functions:

$$\bar{\Delta}_{\lambda,n}(f) = \mathbb{E}\big\{\big(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\big)^2 f(\boldsymbol{x}_1)\big\}\ .$$

It can be shown that

$$\Delta^{(1)}_{\lambda,n} \lesssim \sqrt{\{\log(n)\}^{K+1}\,\bar{\Delta}_{\lambda,n}} + n^{-1} \text{ as } n \to \infty. \tag{S.4}$$

Proof of (S.4)**.** Restricting $\boldsymbol{x}_1$ on set $\{\boldsymbol{x}_1 : \|\boldsymbol{x}_1\|_2 \leq 2\gamma \log n\}$ and using Cauchy-Schwarz inequality, we get

$$\mathbb{E}\left[\big(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\big)^2 I\left\{\|\boldsymbol{x}_1\|_2 \leq 2\gamma \log n\right\}\right] \leq \left[C_{K,\gamma}\left\{\log(n)\right\}^{K+1}\bar{\Delta}_{\lambda,n}(f)\right]^{\frac{1}{2}}.$$

On the tail $\{\boldsymbol{x}_1 : \|\boldsymbol{x}_1\|_2 > 2\gamma \log n\}$ using the same argument as (S.2), we have

$$\mathbb{E}\left[\left|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right| I\left\{\|\boldsymbol{x}_1\|_2 > 2\gamma \log n\right\}\right] = O(n^{-1}).$$

(S.4) follows by combining the above two displays.

The following result lower bounds $\Delta^{(2)}_{\lambda,n}$ using $\bar{\Delta}_{\lambda,n}$.

Lemma S.2.1. *For any $\lambda > 0$, we have*

$$\bar{\Delta}_{\lambda,n} \lesssim \lambda^{-(K+1)}\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + \lambda^2 \log n + \lambda(\log n)^{K+3}\Delta^{(2)}_{\lambda,n}\ . \tag{S.5}$$

The proof of the above lemma is intricate and is presented at the end of this section.

Now, for the proof of theorem 1, we combine (S.3), (S.4) and (S.5). Then, using $\lambda \asymp n^{-\frac{1}{K+2}}$ and the fact that $\Delta^{(1)}_{\lambda,n}$ is bounded, we arrive at

$$\Delta^{(1)}_{\lambda,n} \lesssim \sqrt{\{\log(n)\}^{K+1}\left\{n^{\frac{K+1}{K+2}}\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + n^{-\frac{2}{K+2}}\log(n) + n^{-\frac{1}{K+2}}(\log n)^{K+4}\,\Delta^{(1)}_{\lambda,n}\right\}}. \tag{S.6}$$

Proportion S.2.2, which is stated and proved at the end of this proof, provides the following upper bound on $\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}]$:

$$\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] \leq \frac{\mathbb{E}\left\{\mathfrak{h}_f(\boldsymbol{x}_1)\right\}^2 - \mathbb{E}\left\{\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1)\right\}^2}{n} \tag{S.7}$$

Using the similar argument as (S.3), the numerator in above can be further upper bounded by $2\gamma\,\Delta^{(1)}_{\lambda,n+1}+O(n^{-1})$. Substituting this in (S.6), we arrive at an inequality only involving quantities $\Delta^{(1)}_{\lambda,n}$ and $\Delta^{(1)}_{\lambda,n+1}$. Now, noting that $\lambda \asymp n^{-\frac{1}{K+2}}$ and $\Delta^{(1)}_{\lambda,n}$ is bounded, it easily follows that $\Delta^{(1)}_{\lambda,n} \to 0$ as $n \to \infty$.

Establishing the rate of convergence of $\Delta^{(1)}_{\lambda,n}$ needs further calculations. For that purpose consider $A_n = \max\left\{\Delta^{(1)}_{\lambda,n},\, 2\,n^{-\frac{1}{K+2}}(\log n)^{2K+5}\right\}$. For all large n, the following inequality can be derived from (S.6) and (S.7):

$$A_n \le C\,(\log n)^{K+1} n^{-\frac{1}{2K+4}}\sqrt{A_{n+1}}, \tag{S.8}$$

where C is a constant independent of n.

Applying (S.8) recursively m times we have:

$$A_n \le \left(C(\log n)^{K+1} n^{-\frac{1}{2K+4}}\right)^{1+\cdots+\frac{1}{2^m}} A_{n+m+1}^{\frac{1}{2^{m+1}}}.$$

Note that $A_n < 1$ for all large n. This implies that for any $m > 0$,

$$A_n \le \left(C(\log n)^{K+1} n^{-\frac{1}{2K+4}}\right)^{1+\cdots+\frac{1}{2^m}}.$$

Finally, let $m \to \infty$, we proved that $A_n \le C(\log n)^{2K+2} n^{-\frac{1}{K+2}}$, which implies

$$\Delta^{(1)}_{\lambda,n} \lesssim (\log n)^{2K+2} n^{-\frac{1}{K+2}}\ .$$

This completes the proof of Theorem 1.

S.2.1 Proofs of results used in the proof of Theorem 1

Proposition S.2.2. *Let $K_\lambda(\cdot,\cdot)$ be RBF kernel with bandwidth parameter $\lambda \in \Lambda$ and Λ is a compact set of $\mathbb{R}^+$ bounded from zero. Then we have*

$$\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n}] \le \frac{\mathbb{E}\left\{\mathfrak{h}_f(\boldsymbol{x}_1)\right\}^2 - \mathbb{E}\left\{\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1)\right\}^2}{n-1}.$$

Proof of Proposition S.2.2. By the construction of the $\hat{\boldsymbol{h}}_{\lambda,n}$, we have

$$\widehat{\mathcal{S}}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n}] \le \widehat{\mathcal{S}}_\lambda[\mathfrak{h}_f]. \tag{S.9}$$

Taking the expectation on the both sides of equation (S.9), we get

$$\frac{n^2-n}{n^2}\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n}] + \frac{n}{n^2}\Big(\mathbb{E}\left\{\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1)\right\}^2 + \frac{1}{\lambda}\Big) \leq \frac{n^2-n}{n^2}\mathcal{S}_\lambda[\mathfrak{h}_f] + \frac{n}{n^2}\Big(\mathbb{E}\left\{\mathfrak{h}_f(\boldsymbol{x}_1)\right\}^2 + \frac{1}{\lambda}\Big).$$

Notice that $\mathcal{S}_\lambda[\mathfrak{h}_f] = 0$ and then the above inequality implies

$$\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n}] \leq \frac{\mathbb{E}\left\{\mathfrak{h}_f(\boldsymbol{x}_1)\right\}^2 - \mathbb{E}\left\{\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1)\right\}^2}{n-1},$$

which completes the proof.

Proof of Lemma S.2.1.

First we assume there are $n+1$ i.i.d. samples, $\boldsymbol{X}_{n+1} = (\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_{n+1})$ where $\boldsymbol{x}_i$s are i.i.d. from f. Note that the definition of $\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}]$ is equivalent to the following definition:

$$\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] = \mathbb{E}\left[D_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1})\right],$$

where the KSD is given by

$$D_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) = K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1})\big(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\big)\big(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_{n+1}) - \mathfrak{h}_f(\boldsymbol{x}_{n+1})\big).$$

We consider the situation when $\boldsymbol{x}_{n+1}$ is in the ϵ-neighboor of $\boldsymbol{x}_1$. For a fixed $\epsilon > 0$, denote

$$I^{(1)}_{\epsilon;\lambda} := \mathbb{E}\Big[D_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}\Big].$$

When $\epsilon = \lambda \log n$, we have

$$I^{(1)}_{\epsilon;\lambda} \leq \mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + O\left(n^{-0.5\log n}\right). \tag{S.10}$$

The proof of (S.10) is non-trivial. To avoid disrupting the flow of arguments here, its proof is not presented immediately but is provided at the end of this subsection.

Denote the following intermediate quantity $I^{(2)}_{\epsilon;\lambda}$ which is close to $\bar{\Delta}_{\lambda,n}(f)$ as

$$I^{(2)}_{\epsilon;\lambda} := \mathbb{E}\left[K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1})\big(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\big)^2 I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}\right].$$

We use Cauchy Schwarz inequality and lipschitz continuity of score function to show $I^{(2)}_{\epsilon;\lambda}$ is bounded by a function of $I^{(1)}_{\epsilon;\lambda}$ as

$$I^{(2)}_{\epsilon;\lambda} \leq I^{(1)}_{\epsilon;\lambda} + O(\epsilon^{K+3}). \tag{S.11}$$

The proof of (S.11) is quite involved and is presented afterwards. Finally, we establish the following bound which along with (S.10) and (S.11) complete the proof of the lemma:

$$\bar{\Delta}_{\lambda,n} \lesssim \lambda^{-K-1} I^{(2)}_{\epsilon;\lambda} + \lambda^2 (\log n)^{K+3} + \lambda \Delta^{(2)}_{\lambda,n} \log n. \tag{S.12}$$

Proof of (S.10)**.** Note that the difference between $\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}]$ and $I^{(1)}_{\epsilon;\lambda}$ is

$$\mathbb{E}\Big[D_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| \geq \epsilon\}\Big].$$

If we use the Gaussian kernal $K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) = e^{-\frac{1}{2\lambda^2}\|\boldsymbol{x}_1 - \boldsymbol{x}_{n+1}\|^2}$ and set $\epsilon = \lambda \log n$, we have $K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| \geq \epsilon\}$ is always bounded by $n^{-0.5\log n}$, which implies the above difference is bounded by $\Delta^{(2)}_{\lambda,n+1} n^{-0.5\log n}$. Note that $\Delta^{(2)}_{\lambda,n+1}$ is bounded, (S.10) follows.

Proof of (S.11)**.** Note that the score function $\mathfrak{h}_f$ is L_f-Lipschitz continuous. If we assume for small ϵ, when $\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon$, we have $\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_{n+1})$ is $L_{n,\epsilon}$-Lipschitz continuous as

$$\left|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_{n+1}) - \hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1)\right| \leq L_{n,\epsilon}\, \epsilon. \tag{S.13}$$

where $L_{n,\epsilon}$ satisfies that $\mathbb{E}L^2_{n,\epsilon}$ is bounded. Then the difference between $I^{(2)}_{\epsilon;\lambda}$ and $I^{(1)}_{\epsilon;\lambda}$ is bounded by

$$\mathbb{E}\Big[\epsilon\,(L_f + L_{n,\epsilon}) K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) \left|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right| I\{\|\boldsymbol{x}_n - \boldsymbol{x}_1\| < \epsilon\}\Big]$$

Apply the Cauchy-Schwarz inequality and the square of above difference can be further bounded by

$$\epsilon\,\mathbb{E}\left[(L_f + L_{n,\epsilon})^2 I\{\|\boldsymbol{x}_n - \boldsymbol{x}_1\| < \epsilon\}\right] \mathbb{E}\left[K^2_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) \left|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right|^2 I\{\|\boldsymbol{x}_n - \boldsymbol{x}_1\| < \epsilon\}\right]$$

Note that $\mathbb{E}\left[(L_f + L_{n,\epsilon})^2 I\{\|\boldsymbol{x}_n - \boldsymbol{x}_1\| < \epsilon\}\right]$ is bounded by

$$C_f \frac{\pi^{(K+1)/2}}{\Gamma(\frac{K+1}{2} + 1)} \epsilon^{K+1} \mathbb{E}(L_f + L_{n,\epsilon})^2,$$

where $\Gamma(x)$ is the gamma function. Notice that $K^2_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) \leq K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1})$ and then we have

$$I^{(2)}_{\epsilon;\lambda} \lesssim I^{(1)}_{\epsilon;\lambda} + \epsilon\sqrt{\epsilon^{K+1}\Delta^2_{\epsilon;\lambda}}.$$

This completes the proof of (S.11).

Proof of (S.12)**.** We introduce an intermediate quantity:

$$I^{(3)}_{\epsilon;\lambda} = \mathbb{E}\left[K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1})\big(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\big)^2 I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}\right].$$

Assume that when n is large and $\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon$, we have $\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_{n+1})$ is $L_{n,\epsilon}$-Lipschitz continuous as:

$$\left|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_{n+1}) - \hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1)\right| \le L_{n,\epsilon}\, \epsilon$$

Combined with (S.13), we get the difference between $I^{(3)}_{\epsilon;\lambda}$ and $I^{(2)}_{\epsilon;\lambda}$ is bounded by

$$4\epsilon^2\, \mathbb{E}\left[L^2_{n,\epsilon} K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}\right],$$

which implies that

$$I^{(3)}_{\epsilon;\lambda} \lesssim I^{(2)}_{\epsilon;\lambda} + \epsilon^{K+3}. \tag{S.14}$$

Next we introduce another intermediate quantity

$$I^{(4)}_{\epsilon;\lambda} = \mathbb{E}\int f(\boldsymbol{x}_1) K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1})\big(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\big)^2 I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}\, d\boldsymbol{x}_{n+1},$$

which is close to $I^{(3)}_{\epsilon;\lambda}$. When $\epsilon = \lambda \log n$, we have the following term

$$\int K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}\, d\boldsymbol{x}_{n+1}$$

is lower bounded by

$$\lambda^{K+1}\int e^{-\frac{1}{2}\|\boldsymbol{x}_{n+1}\|^2} I\{\|\boldsymbol{x}_{n+1}\| < \log n\}\, d\boldsymbol{x}_{n+1},$$

which can be further lower bounded by $c\,\lambda^{K+1}$ for some constant c when n is large. This implies

$$\lambda^{K+1}\,\bar{\Delta}_{\lambda,n}(f) \lesssim I^{(4)}_{\epsilon;\lambda}. \tag{S.15}$$

Now it is enough to show $I^{(4)}_{\epsilon;\lambda} \lesssim I^{(3)}_{\epsilon;\lambda} + \lambda^{K+2}\Delta^{(2)}_{\lambda,n}\log n$.

Assume that f is L_f-Lipschitz continuous. The difference between $I^{(4)}_{\epsilon;\lambda}$ and $I^{(3)}_{\epsilon;\lambda}$ is bounded by

$$L_f\epsilon\,\Delta^{(2)}_{\lambda,n}\int \left[K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}\right]\, d\boldsymbol{x}_{n+1}.$$

Notice that $\int [K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) I\{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_1\| < \epsilon\}]\, d\boldsymbol{x}_{n+1}$ is bounded by $C\,\lambda^{K+1}$ for some constant C. This implies that

$$I^{(4)}_{\epsilon;\lambda} \lesssim \Delta^3_{\epsilon;\lambda} + \lambda^{K+2} I^{(3)}_{\lambda,n} \log n.$$

Combined with (S.14) and (S.15), the result (S.12) follows.

S.3 Proof of Lemma 2

We follow the notions in Section S.2. The convergence rate of $\Delta^{(2)}_{\lambda,n}$ is achieved by extending the results of $\Delta^{(1)}_{\lambda,n}$ in Section S.2. Recall that (S.3) shows

$$\Delta^{(2)}_{\lambda,n} \lesssim \log(n)\,\Delta^{(1)}_{\lambda,n} + n^{-1}\ ,$$

and we have proved $\Delta^{(1)}_{\lambda,n} \lesssim (\log n)^{2K+2} n^{-\frac{1}{K+2}}$ in Section S.2. Combining these two, we obtain the result stated in this lemma.

S.4 Proof of Proposition 2

Proposition 4.5 in Johnstone [2011] shows that $B_n(\boldsymbol{\delta}^\pi(\boldsymbol{y})) = \sigma^2 - \sigma^4 I_Y$. Following the same arguments, we have $B_n(\boldsymbol{\delta}^\pi(\boldsymbol{y}, \boldsymbol{S})) = \sigma^2 - \sigma^4 I(p_{y|\boldsymbol{s}})$. Then it follows

$$B_n(\boldsymbol{\delta}^\pi(\boldsymbol{y})) - B_n(\boldsymbol{\delta}^\pi(\boldsymbol{y}, \boldsymbol{S})) = \sigma^4(I_{Y|\boldsymbol{S}} - I_Y)$$

Next, we prove that $I_{Y|\boldsymbol{S}} - I_Y$ is non-negative. By the definition of $I_{Y|\boldsymbol{S}}$, we have the following decomposition:

$$I_{Y|\boldsymbol{S}} = \iint \Big(\frac{f(y)\nabla_y f(\boldsymbol{s}|y) + f(\boldsymbol{s}|y)\nabla_y f(y)}{f(y, \boldsymbol{s})}\Big)^2 f(y, \boldsymbol{s})\, dy\, d\boldsymbol{s}.$$

Then we break the square and it follows

$$I_{Y|\boldsymbol{S}} = \iint \Big(\frac{\nabla_y f(\boldsymbol{s}|y)}{f(\boldsymbol{s}|y)}\Big)^2 f(y, \boldsymbol{s})\, dy\, d\boldsymbol{s} + \iint \Big(\frac{\nabla_y f(y)}{f(y)}\Big)^2 f(y)\, dy + 2\iint \nabla_y f(\boldsymbol{s}|y) \nabla_y f(y)\, dy\, d\boldsymbol{s}.$$

Note that the second term of right hand side is always non-negative. Then we consider the last term and exchange the integration and partial derivative, we get

$$\iint \nabla_y f(\boldsymbol{s}|y)\nabla_y f(y)\, dy\, d\boldsymbol{s} = \int \nabla_y f(y) \nabla_y \Big(\int f(\boldsymbol{s}|y) d\boldsymbol{s}\Big) dy = 0$$

It follows that $I_Y \geq I_{Y|\boldsymbol{S}}$.

S.5 Proof of Theorem 3

The proof of this theorem follows along the similar lines of the proof for Theorem 1. Denote $\alpha = 1/(3(K+1)(K+2))$. In this case we entertain the possiblity that the joint density f can be a heavier tailed density. We concentrate on set $\{\|\boldsymbol{x}_1\|_2 \leq n^{\alpha}\}$ instead of the set $\{\|\boldsymbol{x}_1\|_2 \lesssim \log n\}$ analyzed in the proof of Theorem 1.

Noting that $\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1)$ and $\mathfrak{h}_f(\boldsymbol{x}_1)$ are both $O(1)$ it follows that $|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)| = O(1)$. Then applying the Cauchy-Schwarz inequality, we get

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 \leq n^{\alpha}\right\}\right] \lesssim \left\{n^{(K+1)\alpha}\,\bar{\Delta}_{\lambda,n}\right\}^{1/2}. \tag{S.16}$$

Next, we consider the situation when $||\boldsymbol{x}_i||_2$ is large. Using $|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)| = O(1)$ and Assumption 2, it follows

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 > n^{\alpha}\right\}\right] \lesssim n^{-\alpha(K+1)}\,. \tag{S.17}$$

Combining (S.16) and (S.17) gives the bound on $\Delta^{(2)}_{\lambda,n}$ as

$$\Delta^{(2)}_{\lambda,n} \lesssim \left\{n^{(K+1)\alpha}\,\bar{\Delta}_{\lambda,n}\right\}^{1/2} + n^{-\alpha(K+1)}\,. \tag{S.18}$$

Now, recall (S.5) in Lemma S.2.1 upper bounds $\bar{\Delta}_{\lambda,n}$ by a function of $\Delta^{(2)}_{\lambda,n}$. Using (S.5) and (S.18), we get

$$\Delta^{(2)}_{\lambda,n} \lesssim n^{(K+1)\alpha/2}\left\{\lambda^{-(K+1)}\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + \lambda^2\log n + \lambda(\log n)^{K+3}\Delta^{(2)}_{\lambda,n}\right\}^{1/2} + n^{-\alpha(K+1)}\,. \tag{S.19}$$

Note that, $\Delta^{(2)}_{\lambda,n}$ is bounded and so, Proposition S.2.2 implies $\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}]$ is bounded by $O(n^{-1})$. Finally, let $\lambda \asymp \Theta(n^{-\frac{1}{K+2}})$ and substitute $\alpha = \frac{1}{3(K+1)(K+2)}$ in (S.19) to obtain

$$\Delta^{(2)}_{\lambda,n} \lesssim n^{-\frac{1}{3(K+2)}}\,,$$

which completes the proof of Theorem 3.

S.6 Proof of Proposition 3

First note that

$$K_\lambda\left((u_i, s_i); (u_j, s_j)\right) / K_\lambda\left((x_i, s_i); (x_j, s_j)\right) = \exp\left\{-\frac{1}{2\lambda}(u_i - u_j)^2 + \frac{1}{2\lambda}(x_i - x_j)^2\right\}$$
$$= \exp\left\{-\frac{1}{2\lambda}\left[\alpha^2(\eta_i - \eta_j)^2 - 2\alpha(\eta_i - \eta_j)(x_i - x_j)\right]\right\} := I_1.$$

For any fixed n, we have $x_{\max} - x_{\min} \le C_1$ and $\eta_{\max} - \eta_{\min} \le C_2$ for some quantities C_1 and C_2. Then the above is bounded by

$$I_2 := \exp\left\{-2^{-1}\lambda^{-1}\alpha(C_2^2\alpha - 2C_1C_2)\right\}.$$

The above ratio for ∇K_λ equals $I_1(u_i - u_j)/(x_i - x_j)$, which is bounded in magnitude by $I_2\,(C_3 + C_2)/C_3$ where $C_3 = \min_{i \neq j} |x_i - x_j|$ and $C_3 > 0$ as the distribution of Y in (1) is continuous.

Now consider the estimators

$$\hat{\delta}^{\text{IT}}_{\lambda,i}(U, S) = u_i + \sigma^2(1 + \alpha^2)\hat{g}_i = y_i + \alpha\eta_i + \sigma^2(1 + \alpha^2)\hat{g}_i, \text{ and,}$$
$$\hat{\delta}^{\text{IT}}_{\lambda,i}(Y, S) = y_i + \sigma^2(1 + \alpha^2)\hat{h}_i\ ,$$

where, for an arbitrary fixed value of λ, $\hat{g}_i$ and $\hat{h}_i$ are solutions from (5) using (u, s) and (y, s) respectively. Note that,

$$\hat{L}_n(\lambda, \alpha) = \frac{1}{n}\sum_i \left(\hat{\delta}^{\text{IT}}_{\lambda,i}(U, S) - v_i\right)^2 - \sigma^2(1 + \alpha^{-2}).$$

Taking expectation and using the fact that V is conditionally independent of (U, S), we get,

$$\mathbb{E}\{\hat{L}_n(\lambda, \alpha)\} = \mathbb{E}\left[\frac{1}{n}\sum_i \left(\hat{\delta}^{\text{IT}}_{\lambda,i}(U, S) - \theta_i\right)^2\right].$$

For any fixed n,

$$D_i := \hat{\delta}^{\text{IT}}_{\lambda,i}(U, S) - \hat{\delta}^{\text{IT}}_{\lambda,i}(U, S) = \alpha\eta_i + \sigma^2\left[(1 + \alpha^2)\hat{g}_i - \hat{h}_i\right].$$

Now, if the optimization in (5) is strictly convex, then for any small α, there exists ϵ_α such that $\max_i |\hat{g}_i - \hat{h}_i| < \epsilon_\alpha$ and $\epsilon_\alpha \downarrow 0$ as $\alpha \downarrow 0$ and the result stated in this proposition follows.

S.7 Further details on the Real Data Illustrations

S.7.1 Gene Expressions Estimation example

The data considered in this analysis was collected in Sen et al. [2018] via RNA sequencing. The set of genes in the sequencing kit was same across all the experiments. The standard deviations of the expressions values corresponding to the different genes were estimated from related gene expression samples which contain replications under different experimental conditions. Pooling data across these experiments, unexpressed and lowly expressed genes were filtered out. The resultant data consist of around 30% of the genes. We consider the estimation of the mean expression levels of $n = 3000$ genes. The primary parameter $\boldsymbol{\theta}$ is estimated based on primary vector $\boldsymbol{Y}$ and two auxiliary sequences $\boldsymbol{S}_{\mathsf{U}}$ and $\boldsymbol{S}_{\mathsf{I}}$.

In Figure S.1 (top panel), we list the 28 genes for which the Tweedie and integrative Tweedie estimates disagree by more than 50%. According to PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System [Mi et al., 2012], those genes impact 12 molecular functions and 35 biological processes in human cells. The bottom two panels of Figure S.1 present the different function and process types that are impacted.

S.7.2 Predicting monthly sales data example

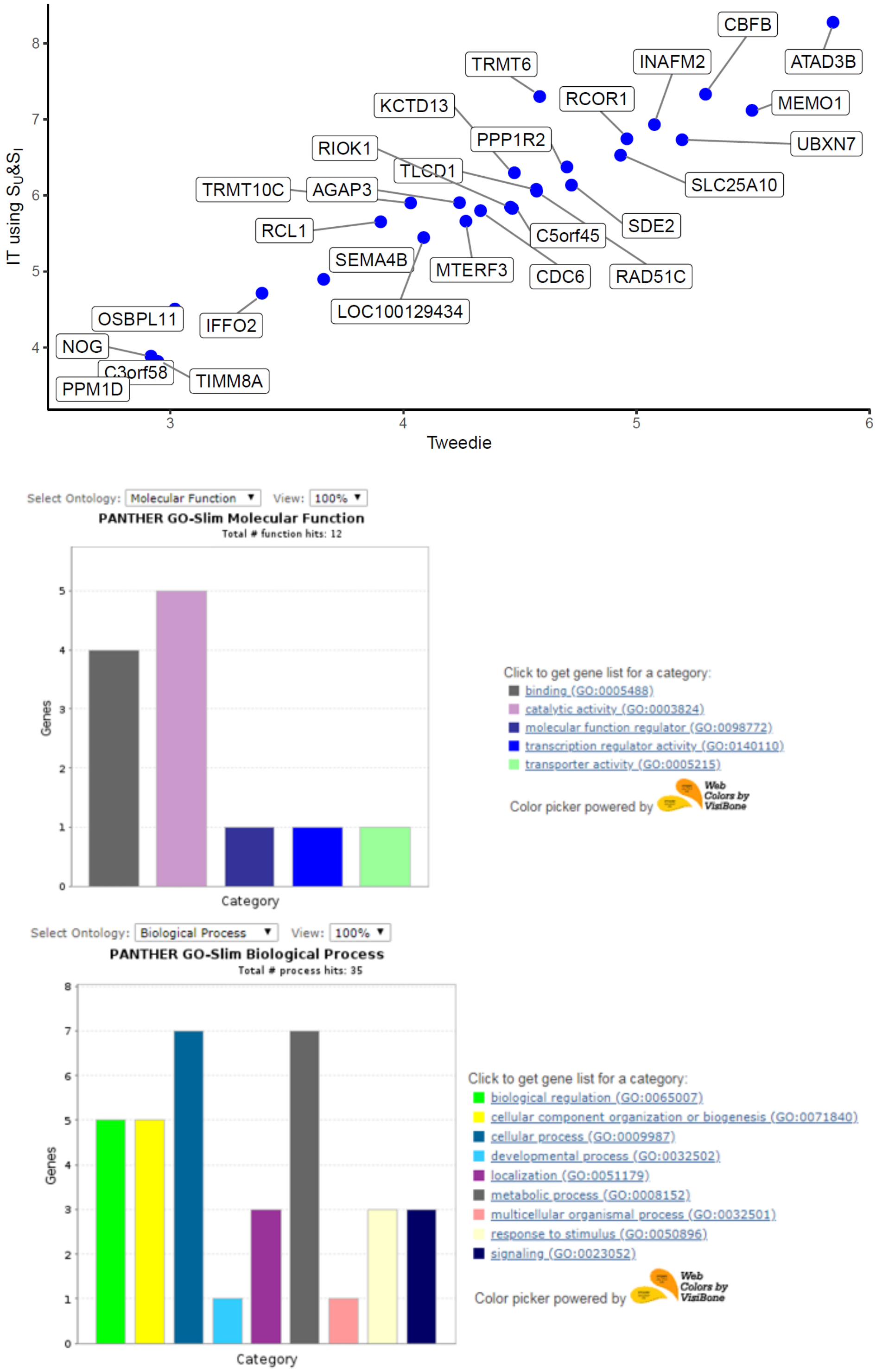
The data is extracted from Bronnenberg et al. [2008]. We consider the monthly sales at the store level for 4 different commodities: beer, milk, deodorant and hotdog. There are 866 stores. The distribution of store across different US states is shown in Figure S.2. Table S.1 shows the correlation between the different products.

In Table S.2, we report the average % gain in predictive error by the JS, T and IT estimators (using different combinations of auxiliary sequences) over the naive unshrunken estimator $\hat{\delta}^{t,\text{naive}} = \boldsymbol{Y}^{t-1}$ for the demand prediction problem at $t = 7, \ldots, 12$. For estimator $\hat{\boldsymbol{\delta}}$ we report,

$$\mathsf{Gain}_t(\hat{\boldsymbol{\delta}}) = \frac{\sum_{i=1}^n \hat{\sigma}_i^2(\hat{\delta}_i^t - \hat{y}_i^t)^2}{\sum_{i=1}^n \hat{\sigma}_i^2(\hat{\delta}^{t,\text{naive}} - \hat{y}_i^t)^2} \times 100\% \quad \text{for} \quad t = 7, \ldots, 12.$$

The last column in Table S.2 reports the average performance of these methods over the six successive trails. These average gains are reported in Table 2 of the main paper.

In Figure (S.3), we compare the prediction of monthly sales in August using Tweedie and IT-M&D&H. The magnitude of side-information is marked using different colors. We can see that the most significant differences between Tweedie and integrative Tweedie are observed in the left-tails.



Select Ontology: Molecular Function ▼ View: 100% ▼

Select Ontology: Biological Process ▼ View: 100% ▼

Figure S.1: Top panel: Scatterplot and names of genes where Tweedie and Integrated Tweedie effect-size estimates disagreed by more than 50%. The other panels show the different molecular function types and biological processes that are impacted by these genes.

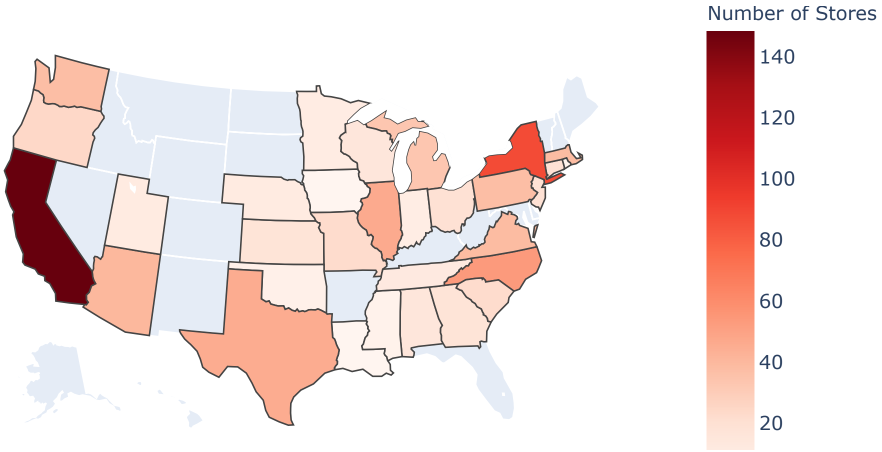


Figure S.2: Distribution of the 866 stores across different states in USA.

Table S.1: Correlation matrix of the monthly sales of different products.

Products		(1)	(2)	(3)	(4)
(1)	Beer	1.00			
(2)	Milk	0.33	1.00		
(3)	Deod	0.16	0.63	1.00	
(4)	Hotdog	0.84	0.38	0.19	1.00

Table S.2: Monthwise % gains for monthly beer sales prediction over the naive unshrunken estimator.

	July	**August**	**September**	**October**	**November**	**December**	**Average**
James-Stein	9.7	2.4	10.8	-2.7	-16.2	-3.7	5.7
Tweedie	7.5	7.5	9.6	-7.2	-22.6	-2.8	4
IT -Milk	11.7	5.2	9.4	-7.4	-8.8	-8.2	6
IT -Deo	11.3	5.1	10.7	-10.6	-13.7	3.7	7.1
IT -Hotdog	12.4	2.6	11.9	-3.2	-13.2	-6.5	6.8
IT-M&D	10.7	5.9	9.8	-7.4	-8.7	-7	6.1
IT-M&H	10.3	5.7	10.8	-4.3	-10.3	-4.8	6.6
IT-D&H	11.7	6.8	11	-8.2	-9.1	-0.6	7.5
IT-M&D&H	11.2	6.8	10.9	-8.1	-7.2	1.8	7.7

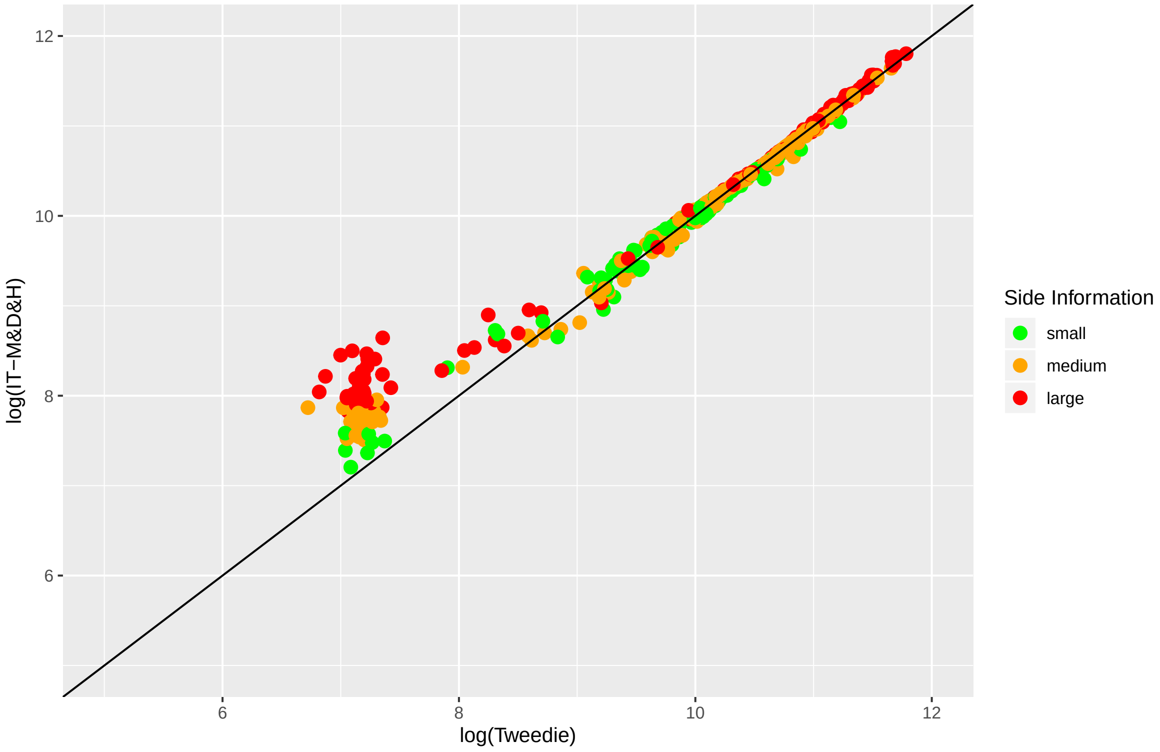


Figure S.3: Scatterplot of the logarithm of beer demand estimates in the month of August. The magnitudes of the corresponding auxiliary variables used in the IT estimate are reflected in the different colors. We can see that the most significant differences between Tweedie and integrative Tweedie are observed in the left-tails. This shows the region where the side information is most helpful.

References

Bart J Bronnenberg, Michael W Kruger, and Carl F Mela. Database paperthe iri marketing data set. *Marketing science*, 27(4):745–748, 2008.

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

Iain M Johnstone. Gaussian estimation: Sequence and wavelet models. *Manuscript, December*, 2011.

Huaiyu Mi, Anushya Muruganujan, and Paul D Thomas. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, 41(D1):D377–D386, 2012.

Nandini Sen, Phillip Sung, Arjun Panda, and Ann M Arvin. 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, 2018.