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

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

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

Supplementary Material

In Sections S.1-S.7 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. We also provide an additional numerical experiment, details regarding the real data example of Section 5 and an additional real data example in Section S.8.

S1 Examples for integrative Tweedie

- **Example 1:** Suppose (Y_i, S_i) are conditionally independent given $(\mu_{i,y}, \mu_{i,s})$. We begin by considering a scenario 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 $\epsilon_i \sim N(0, \sigma^2)$, and $\epsilon_i' \sim N(0, \sigma^2)$. 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:

$$\hat{\mu}_i^{op} = \frac{\frac{1}{2}\sigma^2\mu_0 + \tau^2 Z_i}{\frac{1}{2}\sigma^2 + \tau^2} = \frac{\sigma^2\mu_0 + \tau^2(Y_i + S_i)}{\sigma^2 + 2\tau^2}. \tag{S1.1}$$

The conditional distribution of Y_i given S_i is $Y_i|S_i \sim \mathcal{N}\left(\frac{\sigma^2\mu_0+\tau^2 S_i}{\tau^2+\sigma^2}, \frac{\sigma^2(2\tau^2+\sigma^2)}{\tau^2+\sigma^2}\right)$. 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 (S1.1). It is important to note that if we perturbate the model slightly, say by adding η_i to S_i: $S_i = \mu_i + \eta_i + \epsilon_i'$, or letting $S_i = f(\mu_i) + \epsilon_i'$, then averaging Y and S via (S1.1) may result in poor estimates. However, integrative Tweedie provides a robust data combination approach that consistently reduces the estimation risk (Proposition 2).

- **Example 2:** Consider a scenario where S_i is a group indicator with two equal-sized groups ($S = 1$ and $S = 2$). The primary data follows $Y_i | S_i = k \sim (1 - \pi_k)N(0, 1) + \pi_k N(\mu_i, 1)$, with $\pi_1 = 0.01$, $\pi_2 = 0.4$, and $\mu_i \sim N(2, 1)$. We compare the performance of two oracle Bayes rules, namely $\delta_i^\pi(Y_i, S_i)$ and $\delta_i^\pi(Y_i)$. Calculations show that $\big[B(\delta_i^\pi(Y_i)) - B\{\delta_i^\pi(Y_i, S_i)\}\big]/B\{\delta_i^\pi(Y_i)\} = 0.216$, indicating that incorporating S_i can significantly reduce the risk. Despite the considerable difference between the distributions of Y_i and S_i (continuous vs. binary), integrative Tweedie remains highly effective in reducing estimation risk by leveraging the grouping structure encoded in S_i.

S2 Proof of Proposition 1

The idea of the proof follows from Brown (1971); we provide it here for completeness.

First, note that $\nabla_y \log f(y|s) = \nabla_y \log f(y, s)$ as

$$\nabla_y \log f(y|s) = \nabla_y\{\log f(y, s) - \log f(s)\} = \nabla_y \log f(y, s).$$

Next, from equations (1.1) and (2.2), $\boldsymbol{S}_i$ and Y_i are independent given θ_i, and so $f(y|\theta, \boldsymbol{s}) = f(y|\theta)$ for all θ and $\boldsymbol{s}$. Therefore, noting that $f(y, \boldsymbol{s}) = \int f(y, \boldsymbol{s}|\theta) dh_\theta(\theta)$ and $f(y, \boldsymbol{s}|\theta) = f(y|\theta)f(\boldsymbol{s}|\theta)$, 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.

S3 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 (S3.4). The lower bound follows from (S3.6), whose proof is quite convoluted and is presented separately in Lemma 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 (2.4) 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 $\{\boldsymbol{x}_1 : \|\boldsymbol{x}_1\|_2 \leq 2\gamma \log n\}$:

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 \leq 2\gamma \log n\right\}\right] \leq 2\gamma \log(n)\, \Delta^{(1)}_{\lambda,n}. \tag{S3.2}$$

On the set $\{\|\boldsymbol{x}_1\|_2 > 2\gamma \log n\}$, again using the aforementioned property of score estimates from (2.4) we note that

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 > 2\gamma \log n\right\}\right] \lesssim \mathbb{E}\left[\|\boldsymbol{x}_1\|_2^2 I\{\|\boldsymbol{x}_1\|_2 > 2\gamma \log n\}\right], \tag{S3.3}$$

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 (S3.3) is bounded by $O(n^{-1})$. Combining (S3.2) and (S3.3) we have the following upper bound on $\Delta^{(2)}_{\lambda,n}$:

$$\Delta^{(2)}_{\lambda,n} \lesssim \log(n)\, \Delta^{(1)}_{\lambda,n} + n^{-1}\ . \tag{S3.4}$$

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:

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

It can be shown that

$$\Delta^{(1)}_{\lambda,n} \lesssim \sqrt{\{\log(n)\}^{K+1}\, \bar{\Delta}_{\lambda,n}} + n^{-1} \text{ as } n \to \infty. \tag{S3.5}$$

Proof of (S3.5). 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

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

On the tail $\{\boldsymbol{x}_1 : \|\boldsymbol{x}_1\|_2 > 2\gamma\log n\}$ using the same argument as (S3.3), we have

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

(S3.5) follows by combining the above two displays.

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

Lemma 1. *For any $\lambda > 0$, we have*

$$\bar{\Delta}_{\lambda,n} \lesssim \lambda^{-(K+1)}\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + \lambda^2\log n + \lambda(\log n)^{K+3}\Delta^{(2)}_{\lambda,n}\ . \tag{S3.6}$$

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 (S3.4), (S3.5) and (S3.6). Then, using $\lambda \asymp n^{-\frac{1}{K+2}}$ and the fact that $\Delta^{(1)}_{\lambda,n}$ is bounded, we arrive at

$$\Delta^{(1)}_{\lambda,n} \lesssim \sqrt{\{\log(n)\}^{K+1}\left\{n^{\frac{K+1}{K+2}}\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + n^{-\frac{2}{K+2}}\log(n) + n^{-\frac{1}{K+2}}(\log n)^{K+4}\,\Delta^{(1)}_{\lambda,n}\right\}}. \tag{S3.7}$$

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

$$\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] \le \frac{\mathbb{E}\left\{\mathfrak{h}_f(\boldsymbol{x}_1)\right\}^2 - \mathbb{E}\left\{\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1)\right\}^2}{n} \tag{S3.8}$$

Using the similar argument as (S3.4), the numerator in above can be further upper bounded by $2\gamma\,\Delta^{(1)}_{\lambda,n+1} + O(n^{-1})$. Substituting this in (S3.7), 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 (S3.7) and (S3.8):

$$A_n \leq C\,(\log n)^{K+1} n^{-\frac{1}{2K+4}} \sqrt{A_{n+1}}, \tag{S3.9}$$

where C is a constant independent of n.

Applying (S3.9) recursively m times we have:

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

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

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

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

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

This completes the proof of Theorem 1.

S3.1 Proofs of results used in the proof of Theorem 1

Proposition 1. *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*

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

Proof of Proposition 1. By the construction of the $\hat{\boldsymbol{h}}_{\lambda,n}$, we have

$$\widehat{\mathcal{S}}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n}] \leq \widehat{\mathcal{S}}_\lambda[\boldsymbol{\mathfrak{h}}_f]. \tag{S3.10}$$

Taking the expectation on the both sides of equation (S3.10), we get

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

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 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

$$I^{(1)}_{\epsilon;\lambda} \leq \mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + O\left(n^{-0.5\log n}\right). \tag{S3.11}$$

The proof of (S3.11) 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

$$I^{(2)}_{\epsilon;\lambda} \leq I^{(1)}_{\epsilon;\lambda} + O(\epsilon^{K+3}). \tag{S3.12}$$

The proof of (S3.12) is quite involved and is presented afterwards. Finally, we establish the following bound which along with (S3.11) and (S3.12) complete the proof of the lemma:

$$\bar{\Delta}_{\lambda,n} \lesssim \lambda^{-K-1} I^{(2)}_{\epsilon;\lambda} + \lambda^2 (\log n)^{K+3} + \lambda \Delta^{(2)}_{\lambda,n} \log n. \tag{S3.13}$$

Proof of (S3.11). 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 kernel $K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) = e^{-\frac{1}{2\lambda^2}\|\boldsymbol{x}_1 - \boldsymbol{x}_{n+1}\|^2}$ 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, (S3.11) follows.

Proof of (S3.12). 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

$$\Big|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_{n+1}) - \hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_{n+1}](\boldsymbol{x}_1)\Big| \leq L_{n,\epsilon}\, \epsilon. \tag{S3.14}$$

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

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

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

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

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 $K_\lambda^2(\boldsymbol{x}_1, \boldsymbol{x}_{n+1}) \leq K_\lambda(\boldsymbol{x}_1, \boldsymbol{x}_{n+1})$ and then we have

$$I_{\epsilon;\lambda}^{(2)} \lesssim I_{\epsilon;\lambda}^{(1)} + \epsilon\sqrt{\epsilon^{K+1}\Delta_{\epsilon;\lambda}^2}.$$

This completes the proof of (S3.12).

Proof of (S3.13)**.** We introduce an intermediate quantity:

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

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:

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

Combined with (S3.14), we get the difference between $I_{\epsilon;\lambda}^{(3)}$ and $I_{\epsilon;\lambda}^{(2)}$ is bounded by

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

which implies that

$$I_{\epsilon;\lambda}^{(3)} \lesssim I_{\epsilon;\lambda}^{(2)} + \epsilon^{K+3}. \tag{S3.15}$$

Next we introduce another intermediate quantity

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

which is close to $I_{\epsilon;\lambda}^{(3)}$. 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

$$\lambda^{K+1}\,\bar{\Delta}_{\lambda,n}(f) \lesssim I^{(4)}_{\epsilon;\lambda}. \tag{S3.16}$$

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

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

Combined with (S3.15) and (S3.16), the result (S3.13) follows.

S4 Proof of Lemma 1

We follow the notions in Section S3. The convergence rate of $\Delta^{(2)}_{\lambda,n}$ is achieved by extending the results of $\Delta^{(1)}_{\lambda,n}$ in Section S3. Recall that (S3.4) 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 S3. Combining these two, we obtain the result stated in this lemma.

S5 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 $I_Y \geq I_{Y|\boldsymbol{S}}$.

S6 Proof of Theorem 2

The proof of this theorem follows along the similar lines of the proof for Theorem 1. Denote $\beta = \frac{1}{(K+2)(K+3+2\delta)}$. In this case we entertain the possibility that the joint density f can be a heavier tailed density. We concentrate on set $\left\{\|\boldsymbol{x}_1\|_2 \leq n^\beta\right\}$ 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 $O\left(\|\boldsymbol{x}_1\|\right)$, it follows that $|\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) -$

$\mathfrak{h}_f(\boldsymbol{x}_1)| = O\left(\|\boldsymbol{x}_1\|\right)$. Then applying the Cauchy-Schwarz inequality, we get

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 \le n^\beta\right\}\right] \lesssim \left\{n^{(K+3)\beta}\, \bar{\Delta}_{\lambda,n}\right\}^{1/2}. \tag{S6.17}$$

Next, we consider the situation when $||\boldsymbol{x}_i||_2$ is large. Using assumption 2, it follows

$$\mathbb{E}\left[\left(\hat{\boldsymbol{h}}_\lambda[\boldsymbol{X}_n](\boldsymbol{x}_1) - \mathfrak{h}_f(\boldsymbol{x}_1)\right)^2 I\left\{\|\boldsymbol{x}_1\|_2 > n^\beta\right\}\right] \lesssim n^{-\delta\beta}\ . \tag{S6.18}$$

Combining (S6.17) and (S6.18) gives the bound on $\Delta^{(2)}_{\lambda,n}$ as

$$\Delta^{(2)}_{\lambda,n} \lesssim \left\{n^{(K+3)\beta}\, \bar{\Delta}_{\lambda,n}\right\}^{1/2} + n^{-\delta\beta}\ . \tag{S6.19}$$

Now, recall (S3.6) in Lemma 1 upper bounds $\bar{\Delta}_{\lambda,n}$ by a function of $\Delta^{(2)}_{\lambda,n}$. Using (S3.6) and (S6.19), we get

$$\Delta^{(2)}_{\lambda,n} \lesssim n^{(K+3)\beta/2}\left\{\lambda^{-(K+1)}\mathcal{S}_\lambda[\hat{\boldsymbol{h}}_{\lambda,n+1}] + \lambda^2 \log n + \lambda(\log n)^{K+3}\Delta^{(2)}_{\lambda,n}\right\}^{1/2} + n^{-\delta\beta}\ . \tag{S6.20}$$

Note that, $\Delta^{(2)}_{\lambda,n}$ is bounded and so, Proposition 1 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 $\beta = \frac{2}{(K+2)(K+3+\delta)}$ in (S6.20) to obtain

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

which completes the proof of Theorem 2.

S7 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 $x_{\max} - x_{\min} \leq C_1$ and $\eta_{\max} - \eta_{\min} \leq C_2$ 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 (1.1) 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 (2.4) 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 (2.4) 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.

S8 Further details on simulation and Real Data Illustrations

S8.1 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 $\mu_{2i} = \mathbb{E}(X_{2i})$, $1 \leq i \leq n$. Suppose we are interested in estimating the differences $\boldsymbol{\theta} = \{\mu_{1i} - \mu_{2i} : 1 \leq i \leq n\}$. The primary statistic is given by $\boldsymbol{Y} = \{X_{1i} - X_{2i} : 1 \leq i \leq n\}$. However, it is argued in Cai et al. (2019) 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 $\boldsymbol{\eta} = \{\mu_{1i} + \mu_{2i} : 1 \leq i \leq n\}$. Our idea is to construct an auxiliary sequence $\boldsymbol{S} = \{X_{1i} + X_{2i} : 1 \leq i \leq n\}$ and incorporate $\boldsymbol{S}$ into inference to improve the efficiency.

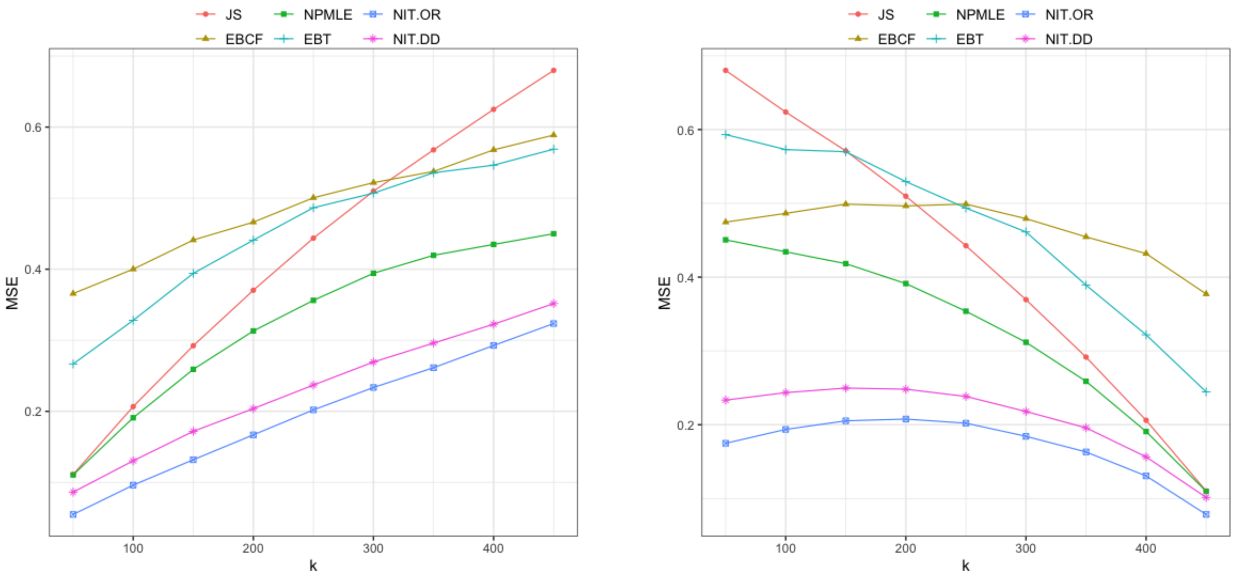
To illustrate the effectiveness of the integrative estimation strategy, we simulate data according to the following two settings and obtain primary and auxiliary data as $\boldsymbol{Y} = \{X_{1i} - X_{2i} : 1 \leq i \leq n$ and $\boldsymbol{S} = \{X_{1i} + X_{2i} : 1 \leq i \leq n\}$.

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

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



(a) Setting 1

(b) Setting 2

Figure 1: Two-sample inference of sparse means.

to investigate the impact of sparsity level on the efficiency of different methods.

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 1. 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.

(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.

S8.2 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 2 (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 2 present the different function and process types that are impacted.

S8.3 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 as shown in Figure 4. The data is extracted from Bronnenberg et al. (2008), which has been widely studied for inventory management and consumer preference analyses; see also Bronnenberg et al. (2012) 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 $t \in \{1, \ldots, 12\}$. 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 $t = 1, \ldots, 6$ for estimating store demand variabilities $\hat{\sigma}_i^2, i = 1, \ldots, n$. For $t = 7, \ldots, 12$, 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 consider the modified James-Stein (JS) estimator of Xie et al. (2012):

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

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 3 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 S8.4.

In Table 1, 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 $t = 7, \ldots, 12$. 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 nonparametric Tweedie. Also, as demonstrated in Table 3, 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 1: 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

S8.4 Additional details on the monthly sales data example

Here, we consider the monthly sales at the store level for 3 additional commodities: milk, deodorant and hotdog. The distribution of 866 store across different US states is shown in Figure 4 and Table 2 shows the correlation between the different products.

In Table 3, 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 3 reports the average performance of these methods over the six successive trails. In Figure 5, 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.

Table 2: 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 3: Month-wise % 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

Bibliography

Bronnenberg, B. J., J.-P. H. Dubé, and M. Gentzkow (2012). The evolution of brand preferences: Evidence from consumer migration. *American Economic Review 102*(6), 2472–2508.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008). Database paper—the iri marketing data set. *Marketing science 27*(4), 745–748.

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

Cai, T. T., W. Sun, and W. Wang (2019). CARS: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *J. Roy. Statist. Soc. B 81*, 187–234.

Johnstone, I. M. (2011). Gaussian estimation: Sequence and wavelet models. *Manuscript, December*.

Mi, H., A. Muruganujan, and P. D. Thomas (2012). 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.

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

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

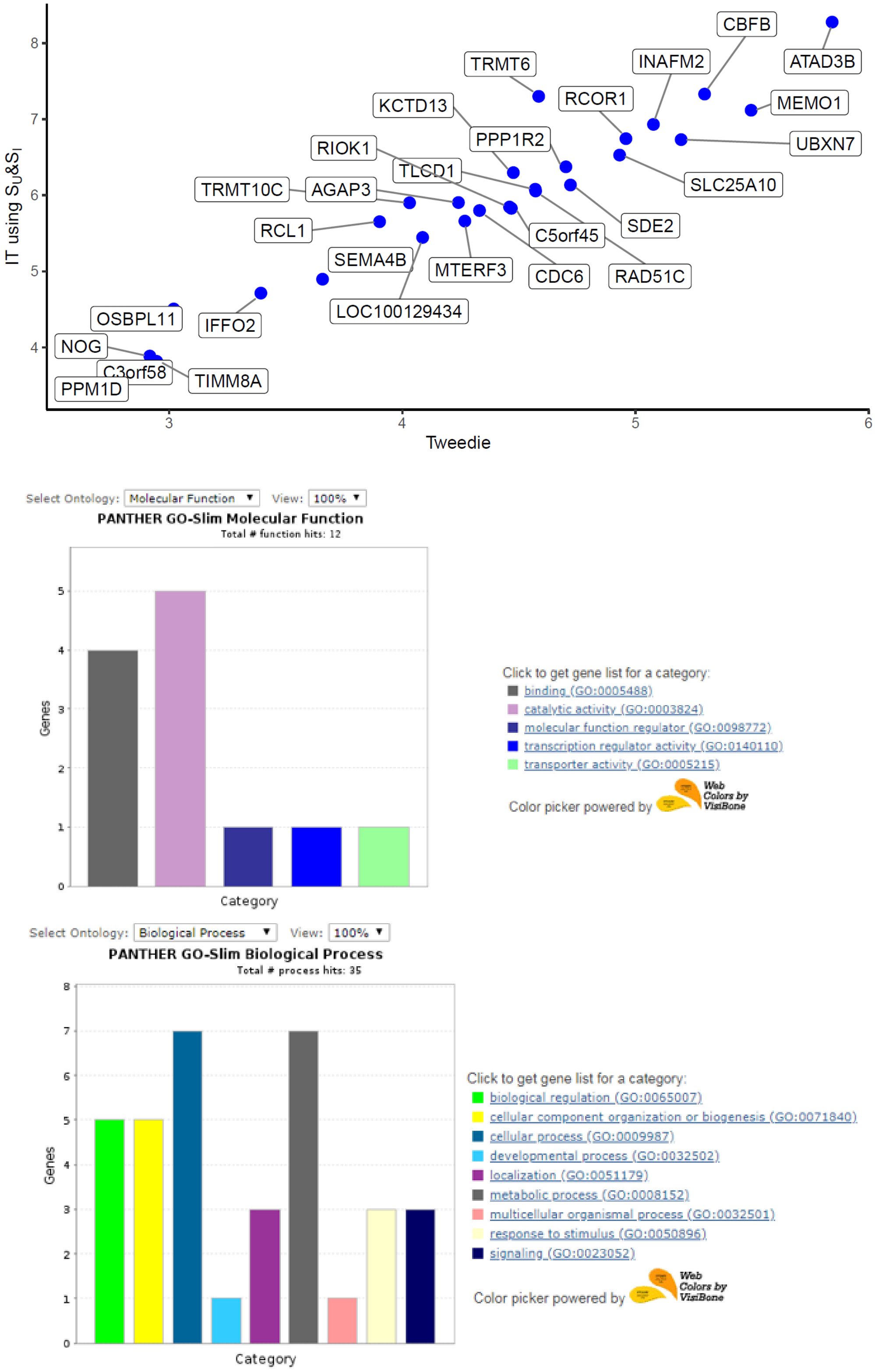


Figure 2: Top panel: Scatterplot and names of genes where Tweedie and Integrated Tweedie effect size estimates disagreed by more than 50%. The other panels show the different molecular function types and biological processes that are impacted by these genes.

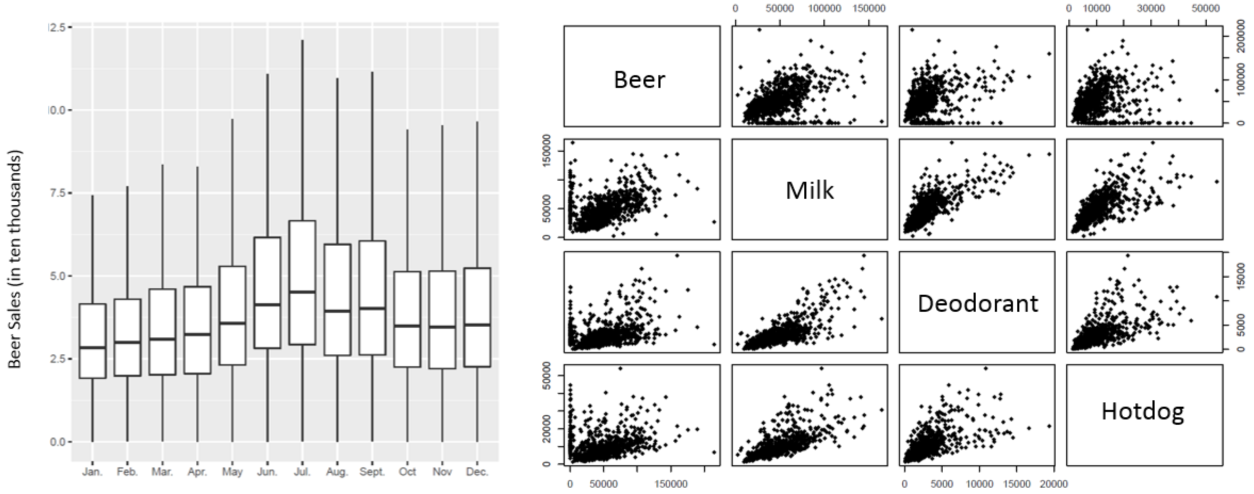


Figure 3: Distribution of monthly sales of beer across stores (on left) and the pairwise distribution of joint sales of different products in the month of July (in right).

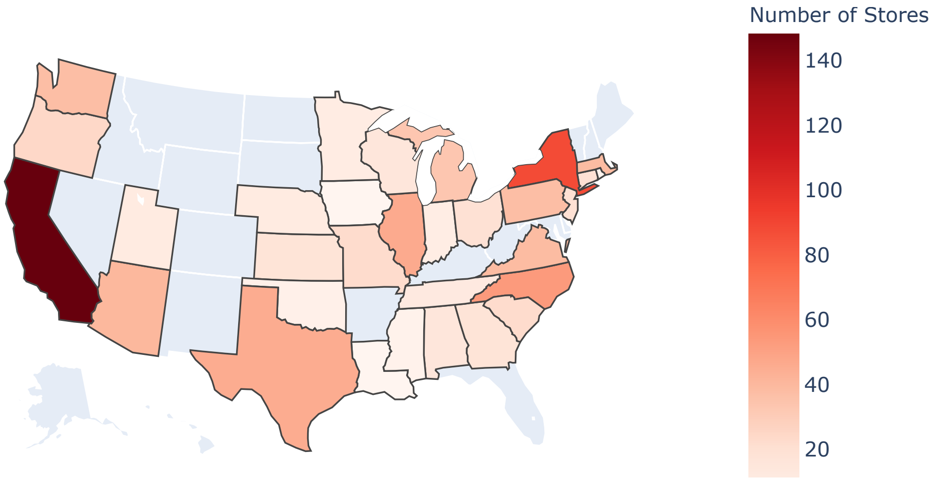


Figure 4: Distribution of the 866 stores across different states in USA.

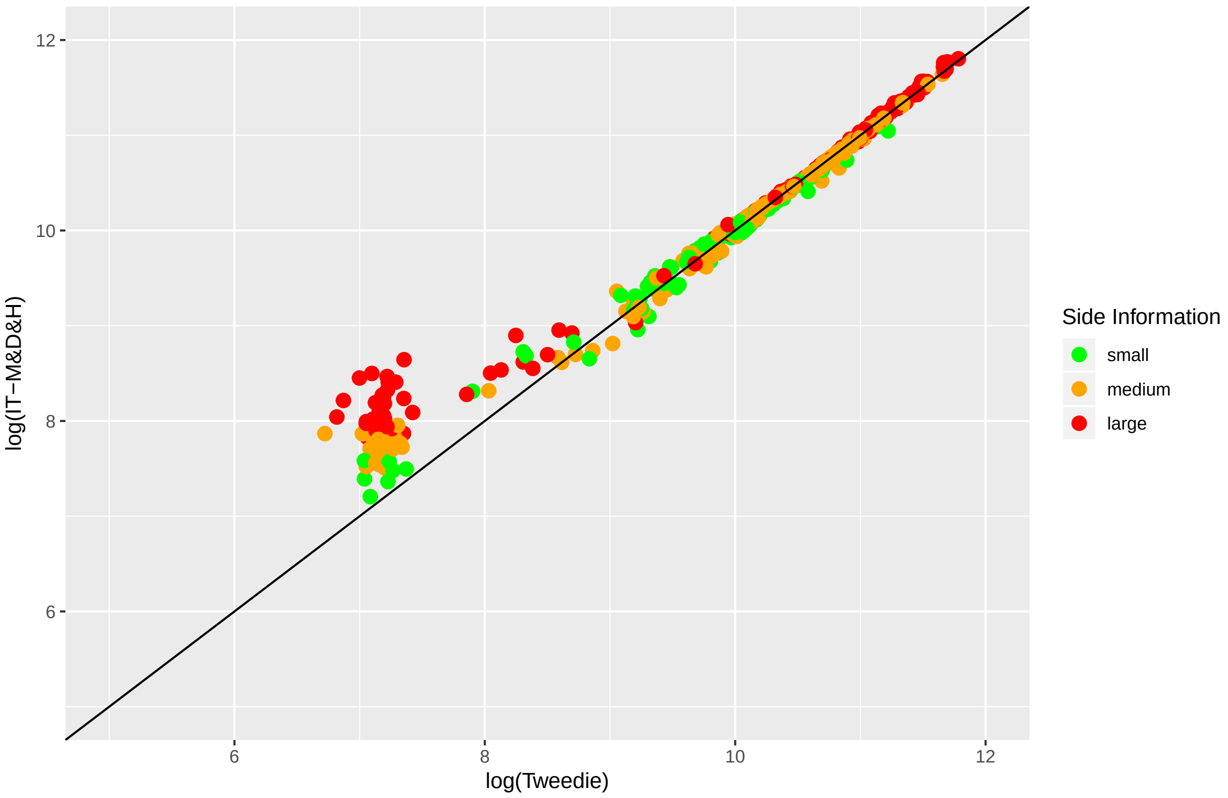


Figure 5: Scatterplot of the logarithm of beer demand estimates in the month of August. The magnitudes of the corresponding auxiliary variables used in the IT estimate are reflected in the different colors. We can see that the most significant differences between Tweedie and integrative Tweedie are observed in the left-tails. This shows the region where the side information is most helpful.