Improved Nonparametric Empirical Bayes Estimation using Transfer Learning

Gourab Mukherjee

University of Southern California

28th November 2022

Monday Colloquium: Statistics and Mathematics Unit

Indian Statistical Institute, Kolkata

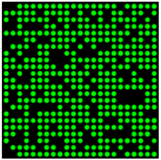
Joint work with: Jiajun Luo (LinkedIn), Trambak Banerjee (Kansas State)
Wenguang Sun (Zhejiang University).

Outline

- Benefits of integrating Auxiliary Data and Side-Information
- Decision theoretic support for some of these integrative methods
- Adaptive Sparse Estimation & 2nd Order Minimax theory
- Non-parametric EB estimation & Kernelized Stein's Discrepancy

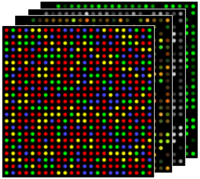
Side Information, Auxiliary Data & Information Pooling

- Technological advancements have allowed collection of vast amounts of data that are properly archived and publicly available.



Primary data for inference

+



Auxiliary data from heterogenous data sources, prior knowledge, etc

Data Enrichment, Market Research, Transfer learning;

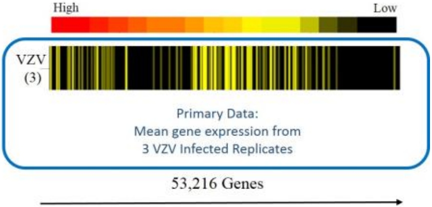
- Incorporating information from these auxiliary data can increase estimation efficiency.

- However, pooling information from these disparate data sources (some of which can be completely useless) is not easy.

- We discuss disciplined frameworks to integrate side information from these auxiliary data

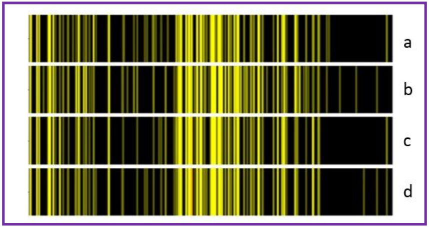
An Example – Estimating Gene Expression

- Observe expression level Y_i of n = 53,216 genes infected with VZV virus by RNA Sequencing
- Goal is to estimate the true expression level θ_i of these n genes under VZV infection.



Primary Data:
Mean gene expression from
3 VZV Infected Replicates

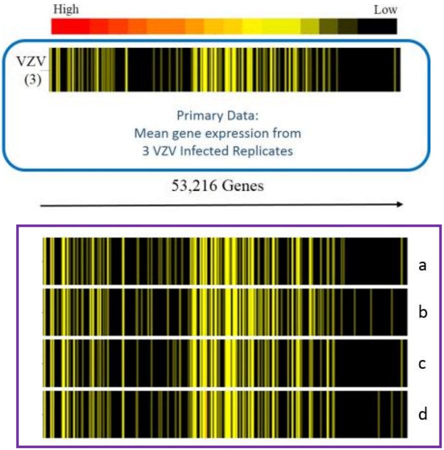
Auxiliary Information - expression levels for the same n genes corresponding to 4 different experimental conditions;



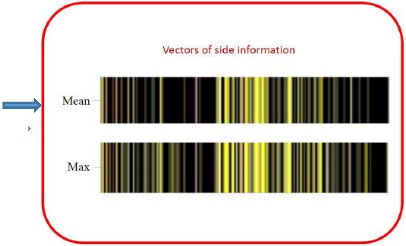
Data: Ann Arvin lab, Stanford medical school, Sen et. al., JVI, 2018

An Example – Estimating Gene Expression

- Observe expression level Y_i of n = 53,216 genes infected with VZV virus by RNA Sequencing
- Goal is to estimate the true expression level θ_i of these n genes under VZV infection.



- high unexpressed gene proportion in standard RNA sequencing Kit in these virology experiments
- auxiliary data contains information about the support of expressed genes;



Auxiliary Information - expression levels for the same n genes corresponding to 4 different experimental conditions; we can extract vector of useful information from them;

Multivariate Normal Mean Estimation

- Gaussian sequence Model. Observe a vector $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ with

$$y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$$

The standard deviation $\{\sigma_i : 1 \leq i \leq n\}$ are known.

- **Goal:** Estimate the multivariate mean vector $\boldsymbol{\theta} = (\theta_i, \ldots, \theta_n)$.

Find estimator $\hat{\boldsymbol{\delta}}(\boldsymbol{y})$ of $\boldsymbol{\theta}$ that minimizes the mean square error

$$\mathcal{L}_n^2(\hat{\boldsymbol{\delta}}, \boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} (\theta_i - \hat{\delta}_i)^2.$$

Problem Formulation: Auxiliary Data

Primary sequence

$\boldsymbol{y} = (y_1, \cdots, y_n)^T$ and $y_i|\theta_i \overset{\text{ind.}}{\sim} \mathcal{N}(\theta_i, \sigma_i^2)$

- Auxiliary sequences $\boldsymbol{S} = (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n)^T$

  K auxiliary sequences $\boldsymbol{s}_i = (s_{i,1}, \cdots, s_{i,K})^T$ each of dimesnion n

- $\boldsymbol{Y}$ and $\boldsymbol{S}$ related. So, we consider estimators $\hat{\delta}(\boldsymbol{Y}, \boldsymbol{S})$ of $\boldsymbol{\theta}$ that uses both $\boldsymbol{Y}$ and $\boldsymbol{S}$.
- Model the relation between $\boldsymbol{Y}$ and $\boldsymbol{S}$ by a highly flexible multi-stage hierarchical model with shared and unshared parameters.

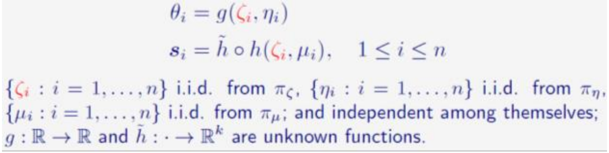
$$\theta_i = g(\zeta_i, \eta_i)$$
$$\boldsymbol{s}_i = \tilde{h} \circ h(\zeta_i, \mu_i), \quad 1 \leq i \leq n$$

$\{\zeta_i : i = 1, \ldots, n\}$ i.i.d. from π_ζ, $\{\eta_i : i = 1, \ldots, n\}$ i.i.d. from π_η, $\{\mu_i : i = 1, \ldots, n\}$ i.i.d. from π_μ; and independent among themselves; $g : \mathbb{R} \to \mathbb{R}$ and $\tilde{h} : \cdot \to \mathbb{R}^k$ are unknown functions.

Problem Formulation: Auxiliary Data

A flexible hierarchical framework

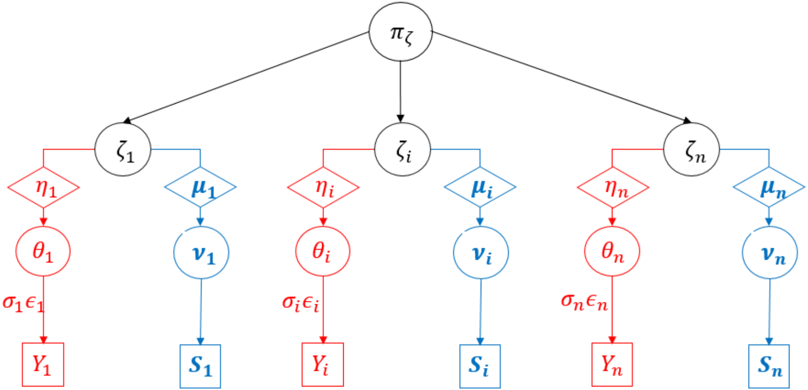
- Y – Primary data vector
- $\boldsymbol{S}$ – observed auxiliary variables



$$\theta_i = g(\zeta_i, \eta_i)$$

$$\boldsymbol{s}_i = \tilde{h} \circ h(\zeta_i, \mu_i), \quad 1 \le i \le n$$

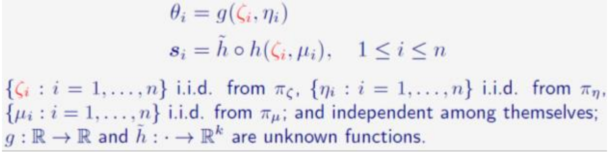
$\{\zeta_i : i = 1, \ldots, n\}$ i.i.d. from π_ζ, $\{\eta_i : i = 1, \ldots, n\}$ i.i.d. from π_η, $\{\mu_i : i = 1, \ldots, n\}$ i.i.d. from π_μ; and independent among themselves; $g : \mathbb{R} \to \mathbb{R}$ and $\tilde{h} : \cdot \to \mathbb{R}^k$ are unknown functions.



Problem Formulation: Auxiliary Data

A flexible hierarchical framework

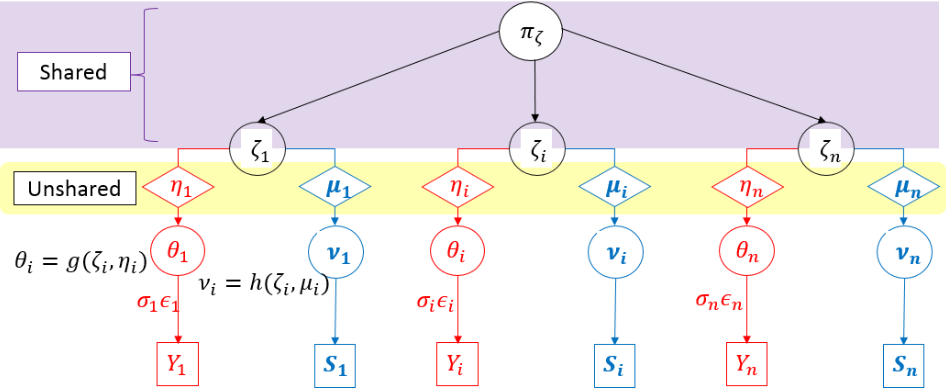
- Y – Primary data vector
- ***S*** – observed auxiliary variables



$$\theta_i = g(\zeta_i, \eta_i)$$

$$\boldsymbol{s}_i = \tilde{h} \circ h(\zeta_i, \mu_i), \quad 1 \le i \le n$$

$\{\zeta_i : i = 1, \ldots, n\}$ i.i.d. from π_ζ, $\{\eta_i : i = 1, \ldots, n\}$ i.i.d. from π_η, $\{\mu_i : i = 1, \ldots, n\}$ i.i.d. from π_μ; and independent among themselves; $g : \mathbb{R} \to \mathbb{R}$ and $\tilde{h} : \cdot \to \mathbb{R}^k$ are unknown functions.



$\theta_i = g(\zeta_i, \eta_i)$

$\nu_i = h(\zeta_i, \mu_i)$

Adaptive Sparse Estimation using Side Information & 2^{nd} order Minimax Risk

ASUS: Adaptive Sparse Estimation using Side Information

- In the virology example, we see that different groups of genes have different levels of sparsity in their expression profiles.

- In practice we do not know these groups. But, we might have auxiliary data that contains information on these groups.

Key Idea:

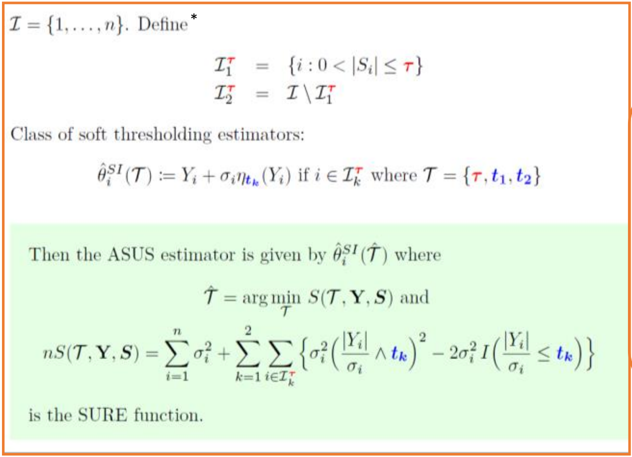
(a) Group genes using side-information; this might help in tacking the heterogeneity in sparsity levels

(b) Adaptively threshold within each group for optimal sparse estimation

Criterion: Minimize Stein's Unbiased Risk Estimate (SURE) to simultaneously select threshold + grouping hyper-parameters.

*for simplicity we consider only one vector of auxiliary data & two groups below

ASUS



$\mathcal{I} = \{1, \ldots, n\}$. Define*

$$\begin{aligned} \mathcal{I}_1^{\tau} &= \{i : 0 < |S_i| \le \tau\} \\ \mathcal{I}_2^{\tau} &= \mathcal{I} \setminus \mathcal{I}_1^{\tau} \end{aligned}$$

Class of soft thresholding estimators:

$$\hat{\theta}_i^{SI}(\mathcal{T}) := Y_i + \sigma_i \eta_{t_k}(Y_i) \text{ if } i \in \mathcal{I}_k^{\tau} \text{ where } \mathcal{T} = \{\tau, t_1, t_2\}$$

Then the ASUS estimator is given by $\hat{\theta}_i^{SI}(\hat{\mathcal{T}})$ where

$$\hat{\mathcal{T}} = \arg\min_{\mathcal{T}} S(\mathcal{T}, \mathbf{Y}, \boldsymbol{S}) \text{ and}$$

$$nS(\mathcal{T}, \mathbf{Y}, \boldsymbol{S}) = \sum_{i=1}^{n} \sigma_i^2 + \sum_{k=1}^{2} \sum_{i \in \mathcal{I}_k^{\tau}} \left\{ \sigma_i^2 \left(\frac{|Y_i|}{\sigma_i} \wedge t_k \right)^2 - 2\sigma_i^2 I \left(\frac{|Y_i|}{\sigma_i} \le t_k \right) \right\}$$

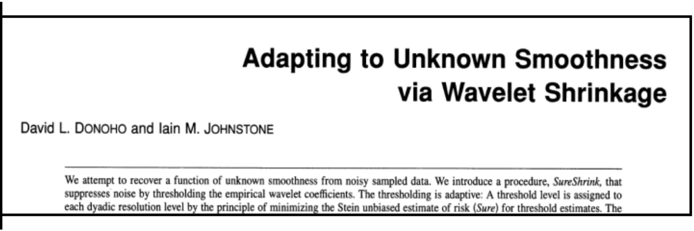
is the SURE function.

ASUS
Algorithm
=
SureShrink
+
Grouping
Hyper-parameters

SureShrink

Algorithm:

JASA, 1995



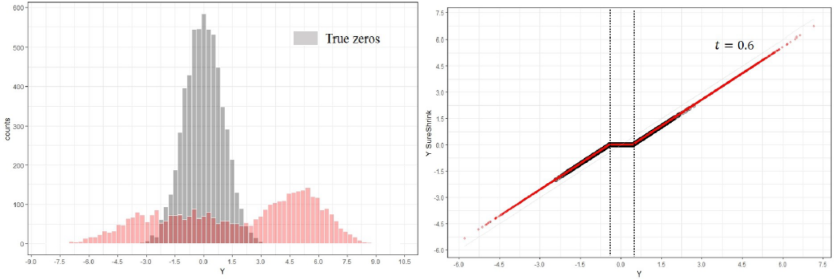
Adapting to Unknown Smoothness via Wavelet Shrinkage

David L. Donoho and Iain M. Johnstone

We attempt to recover a function of unknown smoothness from noisy sampled data. We introduce a procedure, *SureShrink*, that suppresses noise by thresholding the empirical wavelet coefficients. The thresholding is adaptive: A threshold level is assigned to each dyadic resolution level by the principle of minimizing the Stein unbiased estimate of risk (*Sure*) for threshold estimates. The

ASUS Working Principle

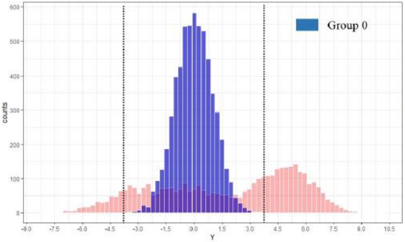
$Y_i = \theta_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 1)$ and $\boldsymbol{\theta}$ is a sparse vector.

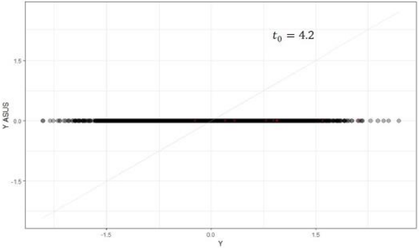


SureShrink estimator [Donoho and Johnstone 1994] soft thresholds the observations at threshold $t = 0.6$ resulting in an MSE of 0.338

ASUS Working Principle

Suppose we have side information available in another variable $\boldsymbol{S}$ that holds sparsity information about $\boldsymbol{\theta}$.

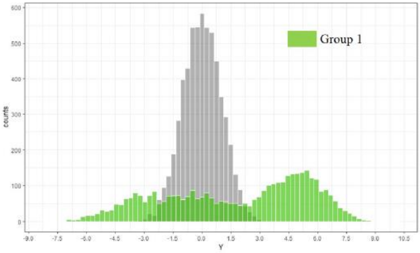


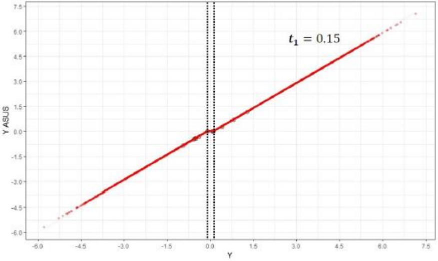


ASUS soft thresholds the observations in Group 0 at $t_0 = 4.2$.

ASUS Working Principle

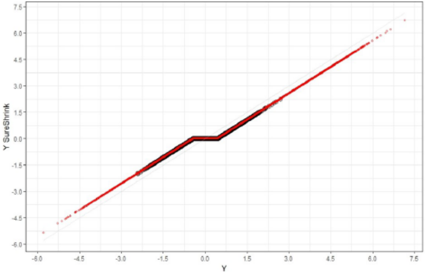
Suppose we have side information available in another variable $\boldsymbol{S}$ that holds sparsity information about $\boldsymbol{\theta}$.

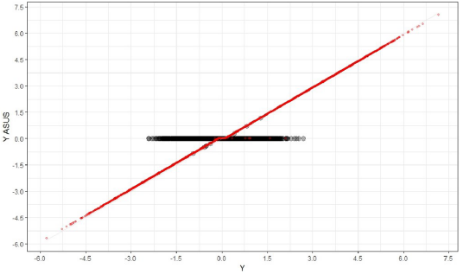




ASUS soft thresholds the observations in Group 1 at $t_1 = 0.15$**.**

ASUS Working Principle





MSE reduction is almost 40%

ASUS & Efficient Information Pooling

We need to develop a disciplined system of efficient information pooling.

Background: (Hypothesis Testing involving sparse means)

Cai, Sun, Wang, JRSS B 2019 showed that in satellite imaging data improved two-sample tests can be constructed by carefully using auxiliary data:

- That encode sparsity information
- Can be integrated via a conditionally independence based inferential framework.

For sparse estimation to show that ASUS is efficient we need to prove:

- It is **adaptive** to the strength of side information and **robust** to it's non-usefulness.
- if the side information is imperfect then the estimator is not too far away from state-of-the-art sparse estimators **built on using no side information.**

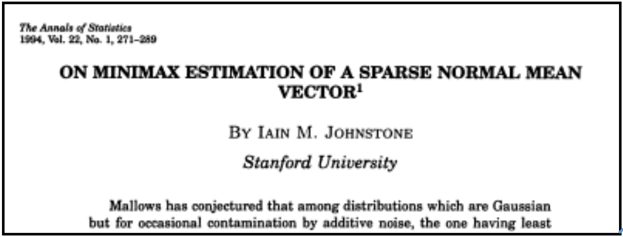
ASUS & Efficient Information Pooling

To establish

(a) usefulness of the side information: characterize the conditions under which methodologies (thresholding rules) ignoring side information are suboptimal and finding out their sub-optimality

(b) Asymptotic optimality & Robustness of ASUS

worst-case risk analysis was conducted (Sec. 3 of Banerjee, M. , Sun, JASA 2019).



The Annals of Statistics
1994, Vol. 22, No. 1, 271–289

ON MINIMAX ESTIMATION OF A SPARSE NORMAL MEAN VECTOR[1]

By Iain M. Johnstone

Stanford University

Mallows has conjectured that among distributions which are Gaussian but for occasional contamination by additive noise, the one having least

However, as Bickel (AoS, 1983) notes, the first order approximation is rather crude and not practically useful.

- First order minimax optimality results involving threshold rules were inadequate for this analysis.
- Elegant higher order minimax risk characterizations from Johnstone, AoS, 1994 were used.
- Saw first-hand the power of minimax decision theoretic results for disciplined development of contemporary data pooling problems.

Non-parametric EB estimation using Side Information: Kernelized Stein Discrepancy

Compound Decision & Tweedie's Formula

- Impose a higher level prior strcuture on the unknown mean:

$$(\theta_1, \ldots, \theta_n) \overset{i.i.d.}{\sim} \pi_\theta, \quad \pi_\theta : \text{unknown prior.}$$

- The compound Bayes risk of $\hat{\boldsymbol{\delta}}$ under mean squared error loss is

$$B(\hat{\boldsymbol{\delta}}) = \mathbb{E}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{\theta}} \mathcal{L}_n^2(\hat{\boldsymbol{\delta}}(\boldsymbol{y}), \boldsymbol{\theta})$$

- **Tweedie's formula:** The **optimal Bayes estimator** of $\boldsymbol{\theta}$

$$\hat{\delta}_i^{\mathsf{T}} = \mathbb{E}(\theta_i | y_i) = y_i + \sigma_i^2 \frac{\boldsymbol{d}}{\boldsymbol{dy}} \mathbf{log}\, \boldsymbol{f}(y_i)$$

  involves the **score function**. [Brown'71]

  $f(y)$ is the unknown marginal density of y

  When $\sigma_i = 1$ for all i: $f = \pi_\theta * \phi$.

- Non-parametric Solutions: [Chapters 6, 7, 21: Efron & Hastie, 2016]

  f-modeling: Brown and Greenshtein'09, Koenker & Mizera'14, Guntoboyina & Saha'18;

  Deconvolution, modeling π_θ: Efron'09, '15, Jiang & Zhang'09.

Integrative Tweedie's Formula

- Mean square error: $\mathcal{L}_n^2(\hat{\boldsymbol{\delta}}(\boldsymbol{y},\boldsymbol{S}),\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}(\theta_i - \hat{\delta}_i)_2^2$
- The compound Bayes risk of $\boldsymbol{\delta}$ under mean square error loss is

$$B_n(\boldsymbol{\delta}) = \mathbb{E}_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{y},\boldsymbol{S}|\boldsymbol{\theta}}\mathcal{L}_n^2(\hat{\boldsymbol{\delta}}(\boldsymbol{y},\boldsymbol{S}),\boldsymbol{\theta})$$

Lemma: Integrated Tweedie's (IT) Formula

The optimal integrated Bayes estimator of $\boldsymbol{\theta}$ is

$$\hat{\delta}_i^{\mathsf{IT}} = \mathbb{E}[\theta_i|y_i,\boldsymbol{s}_i] = y_i + \sigma_i^2\frac{\partial}{\partial y}\log f(y_i,\boldsymbol{s}_i)$$

Recall: Tweedie's formula without auxiliary information is

$$\hat{\delta}_i^{\mathsf{T}} = \mathbb{E}(\theta_i|y_i) = y_i + \sigma_i^2\frac{d}{dy}\log f(y_i)$$

- Possible Efficiency gain: $B_n(\hat{\boldsymbol{\delta}}^{\mathsf{T}}) - B_n(\hat{\boldsymbol{\delta}}^{\mathsf{IT}})$

Possible Efficiency Gain from Auxiliary Data

Assume $\sigma_i = \sigma$ for all i. Then, the theoretically acheivable maximum efficiency gain due to incorporation of auxiliary variables is:

$$B_n(\hat{\boldsymbol{\delta}}^{\mathsf{T}}) - B_n(\hat{\boldsymbol{\delta}}^{\mathsf{IT}}) = \sigma^4\Big(I(f_{y,\boldsymbol{s}}) - I(f_y)\Big)$$

Moreover, $B_n(\hat{\boldsymbol{\delta}}^{\mathsf{T}}) - B_n(\hat{\boldsymbol{\delta}}^{\mathsf{IT}}) \geq 0$ and the equality is attained if and only if y and $\boldsymbol{s}$ are independent, .i.e, $f(y|\boldsymbol{s}) = f(y)$

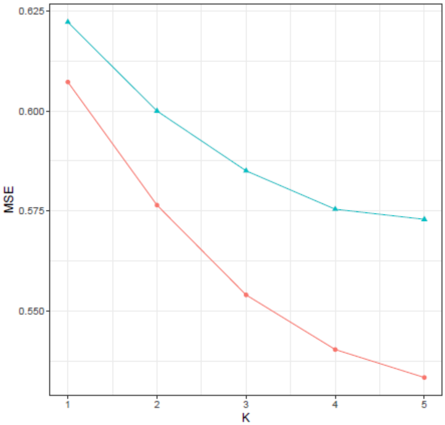
f_y and $f_{y,\boldsymbol{s}}$ denote the unknown marginal and joint density of y and $(y,\boldsymbol{s})$, then their Fisher information is:

$$I(f_y) = \int \big(\frac{d}{dy}\log f(y)\big)^2 f(y)dy \ \ \& \ \ I(f_{y|\boldsymbol{s}}) = \int \Big(\frac{\partial}{\partial y}\log f(y|\boldsymbol{s})\Big)^2 f(y,\boldsymbol{s})dyd\boldsymbol{s}$$

The result directly follows from
`Chapter 4.2, IMJ, Gaussian estimation: Sequence and wavelet models`

A Simulated Example

- The latent vector ζ is drawn i.i.d. from a two-point mixture model: $P(\zeta_i = 0) = P(\zeta_i = 2) = 0.5$
- Mean vectors: $\theta_i = \zeta_i + Z_{1,i}$ and $\nu_{k,i} = \zeta_i + Z_{k+1,i}$, $1 \le k \le K$ with $\boldsymbol{Z}_i \sim N(0, I_{K+1})$.
- Generate $Y_i \sim \mathcal{N}(\theta_i, 1)$ and $S_{k,i} \sim \mathcal{N}(\nu_{k,i}, 1)$, $1 \le k \le K$.



Integrated Tweedie

Red: Oracle

Blue: Estimated

(n=1000)

Proposed Estimator

Based on:

Estimates $\hat{v}_i$ of the score function $\nabla_y \log f(y, \boldsymbol{s})$ at $(y_i, \boldsymbol{s}_i) : 1 \le i \le n$.

Inputs:

- $\boldsymbol{x}_i = (y_i, \boldsymbol{s}_i), 1 \le i \le n.$
- $K_\lambda(\boldsymbol{x}, \tilde{\boldsymbol{x}})$: kernel function with bandwidth λ.

Convex Optimization Criterion

$$\hat{\boldsymbol{v}}(\lambda) = \arg\min_{\boldsymbol{v} \in \boldsymbol{V}_n} \quad \boldsymbol{v}^T \boldsymbol{K}_\lambda \boldsymbol{v} + 2\boldsymbol{v}^T \nabla_1 \boldsymbol{K}_\lambda \mathbf{1}$$

- Kernel matrix of dimension $n \times n$: $(\boldsymbol{K}_\lambda)_{ij} = n^{-2} K_\lambda(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- Gradient of the Kernel: $(\nabla_1 \boldsymbol{K}_\lambda)_{ij} = \nabla_{x_{1j}} K_\lambda(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- $\boldsymbol{V}_n \subset \mathbb{R}^n$; addition convex constraints can be added;

Proposed Estimator: $\quad \hat{\boldsymbol{\delta}}(\lambda) = \{y_i + \sigma_i^2 \, \hat{v}_i(\lambda) : 1 \le i \le n\}.$

Proposed Estimator: Discussions

- Compared to other deconvolution methods, the convex program is more robust and scalable.
- Brown & Greenshtein' 09 estimated 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.
- Additional Convex Constraints such as monotonicity ($\nabla_y \log f(y, \boldsymbol{s})$ is an increasing function of y) when incorporated in the optimization reduce the volume of V_n. The estimator is more robust.
- Discrete Variables: If some of the auxiliary variables are discrete, use generalized Mahalanobis distance.
- Bandwidth selection. Use Modified Cross validation (MCV) akin to `Brown, Greenshtein, Ritov, JASA 2013.`

Rationale behind our Proposed Estimator

For any fixed λ: based on data $(\boldsymbol{y}, \boldsymbol{S})$, $\boldsymbol{K}_\lambda$, $\nabla_1 \boldsymbol{K}_\lambda$ and $\nabla_1^2 \boldsymbol{K}_\lambda$ is evaluated.

Sample Criterion: We minimize:

$$\hat{\mathbb{S}}_{1,\lambda}(\boldsymbol{v}_1) = \boldsymbol{v}_1^T \boldsymbol{K}_\lambda \boldsymbol{v}_1 + 2\boldsymbol{v}_1^T \nabla_1 \boldsymbol{K}_\lambda \mathbf{1}$$

Note that, the gradient in $\nabla_1 \boldsymbol{K}_\lambda$ is over the first coordinate.

Extend this criterion by adding K other criteria:

$$\hat{\mathbb{S}}_{k,\lambda}(\boldsymbol{v}_k) = \boldsymbol{v}_k^T \boldsymbol{K}_\lambda \boldsymbol{v}_k + 2\boldsymbol{v}_k^T \nabla_k \boldsymbol{K}_\lambda \mathbf{1}, \quad k = 2, \ldots, K+1.$$

where gradients for each of the k dimensions are considered. Study:

$$\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{v}) = \sum_{k=1}^{K+1} \hat{\mathbb{S}}_{k,\lambda}(\boldsymbol{v}_k) \text{ and } \quad \hat{\boldsymbol{v}}_{\lambda,n} = \min_{\boldsymbol{v}_k : k=1,\ldots,K+1} \hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{v})$$

Population Version of this extended criterion is
Kernelized Stein Discrepancy (KSD) measure of score functions

Rationale behind our Proposed Estimator

Population Version: Kernelized Stein Discrepancy (KSD)

- $\boldsymbol{f}$ true $K+1$ dimension joint density; true score function $\boldsymbol{h_f}$
- q any other $K+1$ density with score function $\boldsymbol{h} : \mathbb{R}^{K+1} \to \mathbb{R}^{K+1}$

Stein's Discripancy Distance: $\boldsymbol{u}, \tilde{\boldsymbol{u}} \in \mathbb{R}^{K+1}$.

$$\kappa_\lambda[\boldsymbol{h}](\boldsymbol{u}, \tilde{\boldsymbol{u}}) = (\boldsymbol{h}(\boldsymbol{u}) - \boldsymbol{h}(\tilde{\boldsymbol{u}}))^T \boldsymbol{K}_\lambda(\boldsymbol{u}, \tilde{\boldsymbol{u}})(\boldsymbol{h}(\boldsymbol{u}) - \boldsymbol{h}(\tilde{\boldsymbol{u}}))$$

Population Criterion: $\mathbb{M}_\lambda[\boldsymbol{h}] = \mathbb{E}_{\boldsymbol{u}, \tilde{\boldsymbol{u}} \overset{i.i.d.}{\sim} \boldsymbol{f}} \{\kappa_\lambda[\boldsymbol{h}](\boldsymbol{u}, \tilde{\boldsymbol{u}})\}$

Substituting f by $\hat{f}_n$ and $\boldsymbol{h}$ by $\boldsymbol{v}_{\lambda,n}$ above, we get $\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{v})$ for large n .

Properties:

- $\mathbb{M}_\lambda[\boldsymbol{h}] = 0$ iff $\boldsymbol{h} = \boldsymbol{h_f}$.
- $\hat{\mathbb{M}}_{\lambda,n}(\boldsymbol{v})$ is a V-statistic.
- Asymptotic closeness in RKHS norm:

The minimizer $\hat{\boldsymbol{v}}_{\lambda,n}$ is close to h_f in RKHS norm (let $\mathbf{x}_j = (y_j, \boldsymbol{s}_j)$)

$n^{-2} \sum_{i,j} (\hat{\boldsymbol{v}}_{\lambda,n}[i] - \boldsymbol{h_f}(\mathbf{x}_i))^T K_\lambda(\mathbf{x}_i, \mathbf{x}_j)(\hat{\boldsymbol{v}}_{\lambda,n}[j] - \boldsymbol{h_f}(\mathbf{x}_j)) = O(n^{-1})$

Theoretical Challenges: A snippet

The solution $\hat{\boldsymbol{v}}_{\lambda,n}$ is close to $\boldsymbol{h_f}$ in RKHS norm. The rate is parametric $O(n^{-1})$ and does not depend on K for any fixed K.

$$n^{-2}\sum_{i,j}(\hat{\boldsymbol{v}}_{\lambda,n}[i]-\boldsymbol{h_f}(\mathbf{x}_i))^T K_\lambda(\mathbf{x}_i,\mathbf{x}_j)(\hat{\boldsymbol{v}}_{\lambda,n}[j]-\boldsymbol{h_f}(\mathbf{x}_j)) = O(n^{-1})$$

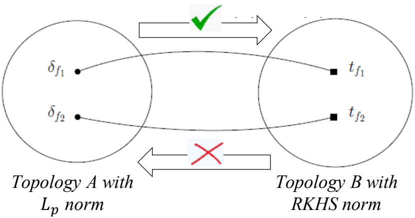
Challenges: We want the score functions to be close in L_2 norm. But, closer in RKHS norm does not trivially imply convergence in L_2 norm ☹

f_1, f_2: two densities
$\delta_{f_1}, \delta_{f_2}$: their tweedie estimator
t_{f_1}, t_{f_2}: their score function

$d_B(t_{f_1}, t_{f_2}) \le d_A(\delta_{f_1}, \delta_{f_2})$

but not always $d_A(\delta_{f_1}, \delta_{f_2}) \le C_0 d_B(t_{f_1}, t_{f_2})$

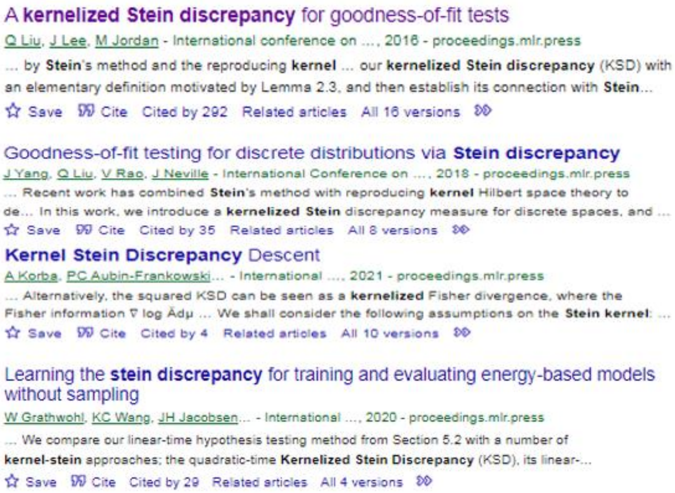
Inverse Problem



Heuristic Explanation:

Choose $K_\lambda(\boldsymbol{u},\tilde{\boldsymbol{u}}) \to 1$ whenever $\boldsymbol{u} \neq \tilde{\boldsymbol{u}}$. This means we need $\lambda \to 0$.

The Role of Statistical Decision Theory



A **kernelized Stein discrepancy** for goodness-of-fit tests

Q Liu, J Lee, M Jordan - International conference on ..., 2016 - proceedings.mlr.press

... by **Stein**'s method and the reproducing **kernel** ... our **kernelized Stein discrepancy** (KSD) with an elementary definition motivated by Lemma 2.3, and then establish its connection with **Stein**...

☆ Save ❞ Cite Cited by 292 Related articles All 16 versions »

Goodness-of-fit testing for discrete distributions via **Stein discrepancy**

J Yang, Q Liu, V Rao, J Neville - International Conference on ..., 2018 - proceedings.mlr.press

... Recent work has combined **Stein**'s method with reproducing **kernel** Hilbert space theory to de... In this work, we introduce a **kernelized Stein** discrepancy measure for discrete spaces, and ...

☆ Save ❞ Cite Cited by 35 Related articles All 8 versions »

Kernel Stein Discrepancy Descent

A Korba, PC Aubin-Frankowski... - International ..., 2021 - proceedings.mlr.press

... Alternatively, the squared KSD can be seen as a **kernelized** Fisher divergence, where the Fisher information ∇ log Ădμ ... We shall consider the following assumptions on the **Stein kernel**: ...

☆ Save ❞ Cite Cited by 4 Related articles All 10 versions »

Learning the **stein discrepancy** for training and evaluating energy-based models without sampling

W Grathwohl, KC Wang, JH Jacobsen... - International ..., 2020 - proceedings.mlr.press

... We compare our linear-time hypothesis testing method from Section 5.2 with a number of **kernel-stein** approaches; the quadratic-time **Kernelized Stein Discrepancy** (KSD), its linear-...

☆ Save ❞ Cite Cited by 29 Related articles All 4 versions »

Though Kernelized Stein Discrepancy (KSD) has been a very vibrant research area in machine learning, all recent applications assume without exact proofs that estimators whose projections are close in the RKHS are also themselves close in Lp or related metrics

Asymptotic Properties

- As $n \to \infty$, set $\lambda_n = O\big(1/(n^{\frac{1}{K+2}} \log(n))\big)$.

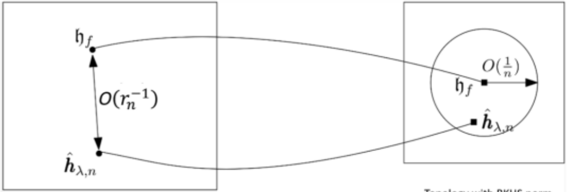
- Let $\boldsymbol{h_f}$ be the true score function of the $(K+1)$ dimensional unknown joint density $\boldsymbol{f}$ of $\boldsymbol{w} = (y, \boldsymbol{S})$. Assume sub-gaussianity.

- Let $\hat{\boldsymbol{v}}_i = (\hat{v}_{1,i}, \hat{v}_{2,i}, \ldots, \hat{v}_{K+1,i})$ be the solution from our proposed optimization with λ_n for the i th observation.

- Rate of convergence: $r_n = \dfrac{\log^{2K+5}(n)}{n^{1/(K+2)}}$ near parametric rate for K=0

Convergence of the Score function

$$\frac{1}{n}\sum_{i=1}^{n} \left\| \hat{\boldsymbol{v}}_i - \boldsymbol{h_f}(y_i, \boldsymbol{S}_i) \right\|_1 = O_p(r_n) \text{ as } n \to \infty$$


$\mathfrak{h}_f$
$O(r_n^{-1})$
$\hat{\boldsymbol{h}}_{\lambda,n}$
$O(\frac{1}{n})$
$\mathfrak{h}_f$
$\hat{\boldsymbol{h}}_{\lambda,n}$

Topology with L1 norm

Topology with RKHS norm

Asymptotic Properties

- As $n \to \infty$, set $\lambda_n = O\big(1/(n^{\frac{1}{K+2}} \log(n))\big)$.

- Let $\boldsymbol{h_f}$ be the true score function of the $(K+1)$ dimensional unknown joint density $\boldsymbol{f}$ of $\boldsymbol{w} = (y, \boldsymbol{S})$. Assume sub-gaussianity.

- Let $\hat{\boldsymbol{v}}_i = (\hat{v}_{1,i}, \hat{v}_{2,i}, \ldots, \hat{v}_{K+1,i})$ be the solution from our proposed optimization with λ_n for the i th observation.

- Rate of convergence: $r_n = \dfrac{\log^{2K+5}(n)}{n^{1/(K+2)}}$ `near parametric rate for K=0`

Convergence of the Score function

$$\frac{1}{n} \sum_{i=1}^{n} \left\| \hat{\boldsymbol{v}}_i - \boldsymbol{h_f}(y_i, \boldsymbol{S}_i) \right\|_1 = O_p(r_n) \text{ as } n \to \infty$$

Convergence of the Loss function of our proposed estimator $\hat{\boldsymbol{\delta}}$

$$|\mathcal{L}_n^2(\hat{\boldsymbol{\delta}}(\lambda_n), \boldsymbol{\theta}) - \mathcal{L}_n^2(\hat{\boldsymbol{\delta}}^{\mathsf{IT}}, \boldsymbol{\theta})| = O_p(r_n) \text{ as } n \to \infty$$

Asymptotic Properties: Discussions

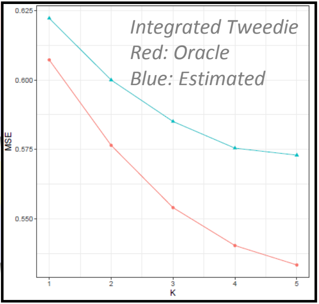
- $\lambda_n = O\left(1/(n^{\frac{1}{K+2}}\log(n))\right)$ and $r_n = \frac{\log^{2K+5}(n)}{n^{1/(K+2)}}$

Convergence of the Score function

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\boldsymbol{v}}_i - \boldsymbol{h}_{\boldsymbol{f}}(y_i,\boldsymbol{S}_i)\right\|_1 = O(r_n) \text{ as } n \to \infty$$

Convergence of the Loss of our proposed estimator $\hat{\boldsymbol{\delta}}$

$$|\mathcal{L}_n^2(\hat{\boldsymbol{\delta}},\boldsymbol{\theta}) - \mathcal{L}_n^2(\hat{\boldsymbol{\delta}}^{\mathsf{IT}},\boldsymbol{\theta})| = O_p(r_n) \text{ as } n \to \infty$$



Integrated Tweedie
Red: Oracle
Blue: Estimated

As the number of auxiliary sequences increases:

- The large sample ($n \to \infty$) risk of the proposed estimator decreases or stays same (if the auxiliary sequence is useless). It does not aggravate ☺
- The rate of convergence however decreases exponentially with increase in K ☹; *curse of dimensionality*.
- Possible remedy: construct a single new auxiliary sequence summarizing the information in all auxiliary sequences. It can be lossy reduction. Use it instead.

Real Data Example: Estimating Gene Expressions

Gene expression of 3000 well expressed genes based on RNA-Seq analysis & filtering

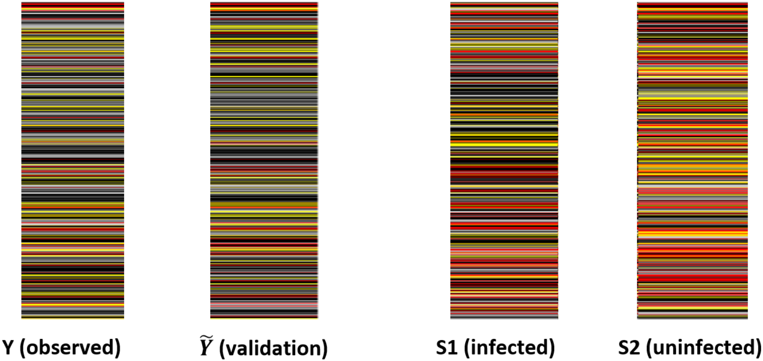
Goal: estimate expression in VZV infected cells when Interferon alpha-1 (INFA) gene is knocked out

Primary Data: Two vectors but one is used for validation and the other for estimation

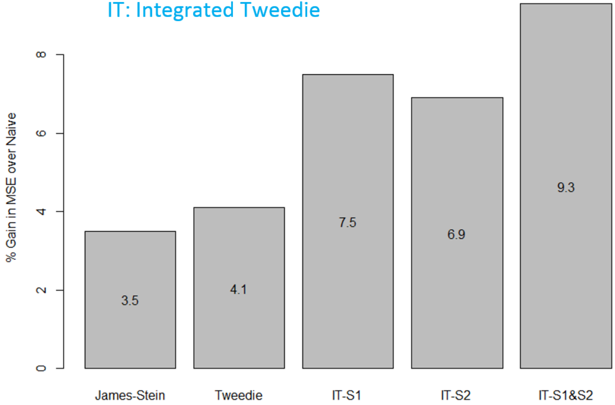
Auxiliary Data: Two sequences

(a) expression in uninfected cells

(b) Expression in infected cells without any gene knocked out.



Real Data Example: Estimating Gene Expressions



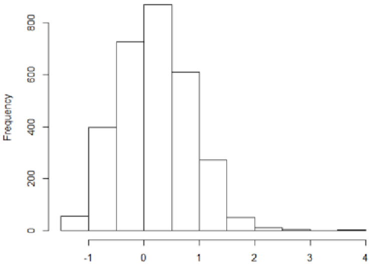
Naïve: using Y to estimate θ

IT: Integrated Tweedie

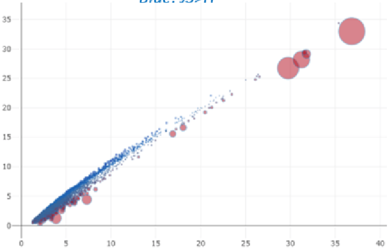
5.2% additional reduction in MSE using transfer learning

Real Data Example: Estimating Gene Expressions

Histogram of the difference between effect size estimates:
D=Integrated Tweedie – James Stein

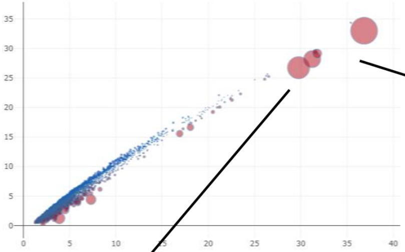


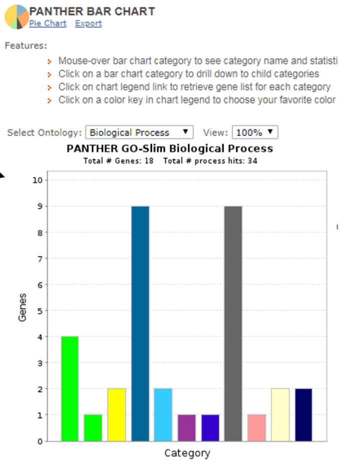
X axis: Integrated Tweedie (IT) Estimate
Y axis: James Stein (JS) Estimate
Size of bubble: exp(|D|)
Red: IT>JS
Blue: JS>IT



- There is difference among the estimates.
- The differences are more pronounced at the tails.
- The differences have more up-regulated Integrative Tweedie estimates.

Real Data Example: Estimating Gene Expressions





PANTHER BAR CHART

Pie Chart Export

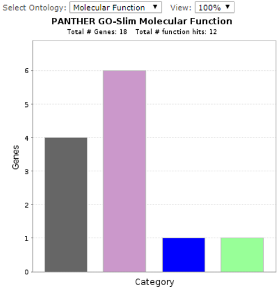
Features:

- Mouse-over bar chart category to see category name and statisti
- Click on a bar chart category to drill down to child categories
- Click on chart legend link to retrieve gene list for each category
- Click on a color key in chart legend to choose your favorite color

Select Ontology: Biological Process ▼ View: 100% ▼

PANTHER GO-Slim Biological Process

Total # Genes: 18 Total # process hits: 34



Select Ontology: Molecular Function ▼ View: 100% ▼

PANTHER GO-Slim Molecular Function

Total # Genes: 18 Total # function hits: 12

Select the genes whose difference in effect size between JS and TL is significant at 5% level.

In human cells those genes are involved in:

- 34 Biological Processes
- 12 Molecular Functions

Real Data Example: Estimating Store Sales

Target: Estimate Monthly Sale of Beers at 866 stores of a retailer across US.

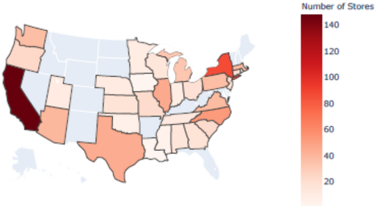
Auxiliary Data: Monthly Sales of three other products in these store

(a) Milk (b) deodorant (c) hotdog

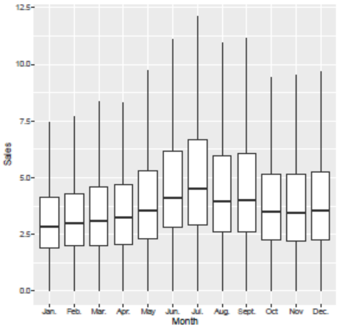
Set-up: We have 12 months of data from Jan-Dec 2006.

We use first six months for estimating $\{\sigma_i : 1 \leq i \leq 866\}$;

For the next 6 months, we use the previous month's data to predict the next.



Distribution of stores across US



Boxplot of Beer sales across stores

Real Data Example: Estimating Store Sales

Target: Estimate Monthly Sale of Beers at 866 stores of a retailer across US.

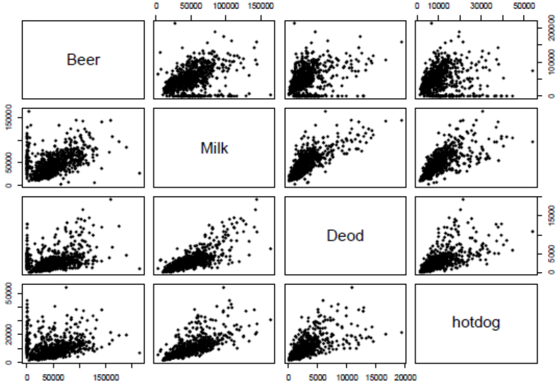
Auxiliary Data: Monthly Sales of three other products in these store

(a) Milk (b) deodorant (c) hotdog

Set-up: We have 12 months of data from Jan-Dec 2006.

We use first six months for estimating $\{\sigma_i : 1 \leq i \leq 866\}$;

For the next 6 months, we use the previous month's data to predict the next.



Sales in July

Real Data Example: Estimating Store Sales

Target: Estimate Monthly Sale of Beers at 866 stores of a retailer across US.

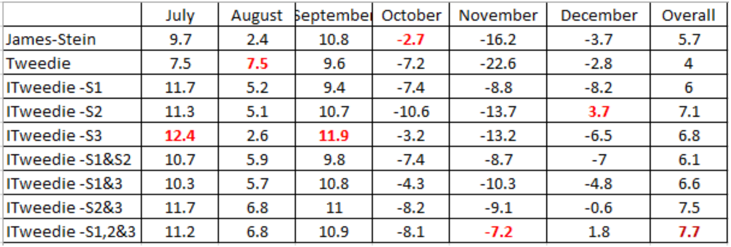
Auxiliary Data: Monthly Sales of three other products in these store

(a) Milk (b) deodorant (c) hotdog

Set-up: We have 12 months of data from Jan-Dec 2006.

We use first six months for estimating $\{\sigma_i : 1 \leq i \leq 866\}$;

For the next 6 months, we use the previous month's data to predict the next.



	July	August	eptembe	October	November	December	Overall
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
ITweedie -S1	11.7	5.2	9.4	-7.4	-8.8	-8.2	6
ITweedie -S2	11.3	5.1	10.7	-10.6	-13.7	3.7	7.1
ITweedie -S3	12.4	2.6	11.9	-3.2	-13.2	-6.5	6.8
ITweedie -S1&S2	10.7	5.9	9.8	-7.4	-8.7	-7	6.1
ITweedie -S1&3	10.3	5.7	10.8	-4.3	-10.3	-4.8	6.6
ITweedie -S2&3	11.7	6.8	11	-8.2	-9.1	-0.6	7.5
ITweedie -S1,2&3	11.2	6.8	10.9	-8.1	-7.2	1.8	7.7

% Gain in MSE over Naive

Overall, integrative analysis yields 3.9% improvement over Tweedie

Real Data Example: Estimating Store Sales

Target: Estimate Monthly Sale of Beers at 866 stores of a retailer across US.

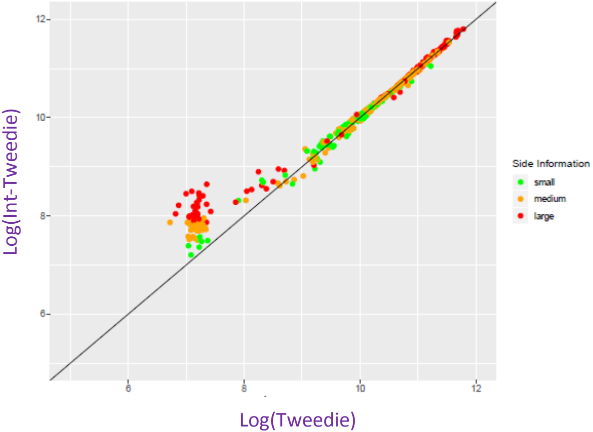
Auxiliary Data: Monthly Sales of three other products in these store

(a) Milk (b) deodorant (c) hotdog

Set-up: We have 12 months of data from Jan-Dec 2006.

We use first six months for estimating $\{\sigma_i : 1 \leq i \leq 866\}$;

For the next 6 months, we use the previous month's data to predict the next.



Again, we find that the difference is mainly
when Naïve-Tweedie underestimates
Compared to Integrated Tweedie

Closing Remarks

Summary

- In this talk:
  - discussed a framework for assimilating information from auxiliary sequences to improve EB estimation;
  - directly estimated score functions using convex optimization
  - used Kernelized Stein discripancy and RKHS theory to establish asymptotic optimality.
- Statistical decision theory provides much needed mathematical support for complex information pooling algorithms.

REFERENCES

Luo J, Mukherjee G and Sun W. Transfer learning for empirical Bayes estimation: an integrative nonparametric Tweedie approach. preprint 2022, `gmukherjee.github.io`

Banerjee T, Mukherjee G and Sun W. Adaptive Sparse Estimation with Side Information. Journal of American Statistical Association, 2019. R-Package: ASUS.

Banerjee T, Liu Q, Mukherjee G and Sun W. A General Framework for Empirical Bayes Estimation in the Discrete Linear Exponential Family. Journal of Machine Learning Research, 2021, R-package: NPEB;

Banerjee T. Nonparametric EB Prediction in Mixed Models, in review, 2022.