Nonparametric empirical Bayes estimation and ranking
A new method for evaluating teachers

Jiaying Gu

University of Toronto

Banff, April 9, 2019

Joint work with Michael Gilraine (NYU), Roger Koenker (UCL), Rob McMillan (UoT)

Preliminary and Incomplete

Teacher Value Added (TVA)

- Leading research questions in education economics:
  - how to use student test results to evaluate teachers performance?
  - what is the short and long term impacts of teachers?
- Typical data environment:
  - Detailed administrative data with longitudinal structure
    - We have obtained data from North Carolina which covers all public school students from fourth and fifth grade from 1996 - 2010 with many detailed demographic data. ($\approx$ 2.7 million student-year observations and 35,000 teachers)
    - Data of similar quality from Los Angeles (11,000 teachers) is also available.
  - Focus on primary school where it is easy to match student with teacher.

Motivation

- Current statistical approach of the TVA literature: James-Stein shrinkage estimator assuming Gaussian teacher effect (Kane and Staiger (2008), citation 804; Chetty et al. (2014), citation 729)
  Question on effect estimation: To what extent are parametric shrinkage methods different from Robbins' nonparametric shrinkage estimator for TVA in real data?

- TVA is used in high-stakes decision making:
  - As of 2017, thirty nine states require TVA to be incorporated into teacher evaluation scores and incentive pay schemes.
  - TVA is used to evaluate education policies (releasing teachers for test score gains).

  Question on ranking: how do we implement such policy - select the best and worst.

Statistical Model

- Index student by i, teacher by j and year by t:

$$A^*_{ijt} = X_{ijt}^\top \beta + \alpha_j + \epsilon_{ijt}, \quad i = 1, 2, \ldots, n_{jt}$$

- A^*_{ijt} are students' test scores centered and scaled for each grade-year.
- X includes polynomials of lagged test scores, students' demographic background, teacher's experience etc.
- Test score residuals $A_{ijt} = A^*_{ijt} - X_{ijt}^\top \hat{\beta} \approx \alpha_j + \epsilon_{ijt}$.
- We work with $y_{jt} = \frac{1}{n_{jt}} \sum_{i=1}^{n_{jt}} A_{ijt} \approx \mathcal{N}(\alpha_j, \sigma_\epsilon^2 / n_{jt})$ to estimate TVA α_j.
- Classroom size n_{jt} in the range of [8, 39] for both NC and LA data.

Effect Estimation: The Compound Decision Problem

- Longitudinal Data: $y_{jt} \sim \mathcal{N}(\alpha_j, \sigma_\epsilon^2 / n_{jt})$, $t = 1, 2, \ldots, T_j$.
- MLE for α_j: $y_j := \sum_t n_{jt} y_{jt} / \sum_t n_{jt} \sim \mathcal{N}(\alpha_j, \sigma_j^2)$, $\quad \sigma_j^2 = \sigma_\epsilon^2 / \sum_t n_{jt}$
- For teachers with small total class size $\sum_t n_{jt}$, MLE is going to be a poor estimator for α_j.
- If $\alpha_j \overset{iid}{\sim} G$, then we can borrow strength from each other.
- Compound decision problem with heterogeneous variances (Jiang and Zhang (2010), Xie, Kou, Brown (2012, 2016), Weinstein, Ma, Brown, Zhang(2018)): a shrinkage estimator for α_j performs better than MLE under $\mathcal{L}_2$ loss $N^{-1} \sum_j (\hat{\alpha}_j - \alpha_j)^2$.
- The loss function considers all teachers and treats every teachers equally.

Linear shrinkage estimator

- If $\alpha_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$, then we get a linear shrinkage rule

$$\hat{\alpha}_j = y_j \frac{\sigma_\alpha^2}{\sigma_j^2 + \sigma_\alpha^2}$$

- Larger total class size $\sum_t n_{jt}$, less shrinkage towards the common mean (zero).
- Practical implementation in the TVA literature: plug-in estimator with MLE, MoM, SURE for $(\sigma_\alpha^2, \sigma_\epsilon^2)$.
- Deviation from Gaussian α_j: Xie, Kou, Brown (2016) suggests a linear rule $(1 - b_j)y_j$ with optimal b_j minimizing $\mathcal{L}_2$ loss subject to $b_j \leq b_k$ if $\sigma_j^2 \leq \sigma_k^2$.

But why linear?

Nonlinear shrinkage estimator I

- For $\alpha_j \overset{iid}{\sim} G$, the Bayes rule is (Tweedie formula)

$$\delta_j = \mathbb{E}(\alpha | y_j, \sigma_j) = y_j + \sigma_j^2 f_j'(y_j)/f_j(y_j) \quad \text{with } f_j(y_j) = \int \frac{1}{\sigma_j}\phi((y_j - \alpha_j)/\sigma_j) dG(\alpha_j)$$

- Marginal density $f_j(y)$ is difficult to estimate due to heterogeneous variances.
- Nonparametric empirical Bayes estimator through NPMLE of G: Robbins (1956), Jiang and Zhang (2010), Gu and Koenker (2017a)

$$\hat{\delta}_j = \frac{\int \alpha\phi((y_j - \alpha)/\sigma_j) d\hat{G}(\alpha)}{\int \phi((y_j - \alpha)/\sigma_j) d\hat{G}(\alpha)}$$

- Convex optimization for $\hat{G}$ (Koenker and Mizera (2014))

$$\hat{G} = \underset{G \in \mathcal{G}}{\operatorname{argmax}} \Big\{ \sum_{j=1}^{N} \log f_j(y_j) \mid f_j(y) = \int \phi((y - \alpha)/\sigma_j)/\sigma_j dG(\alpha) \Big\}$$

- Restriction: iidness of α_j imposes independence between α_j and σ_j.

Nonlinear shrinkage estimator II

- Exploit the longitudinal structure:

$$y_{jt} = \alpha_j + u_{jt}, \quad u_{jt} \sim \mathcal{N}(0, \theta_j / n_{jt})$$

- Sufficient statistics for (α_j, θ_j)

$$y_j = \sum_t n_{jt} y_{jt} / \sum_t n_{jt} \sim \mathcal{N}(\alpha_j, \theta_j / \sum_t n_{jt})$$

$$S_j = \frac{1}{T_j} \sum_t (y_{jt} - y_j)^2 n_{jt} \sim \gamma(r_j, \theta_j / r_j) \text{ with } r_j = (T_j - 1)/2$$

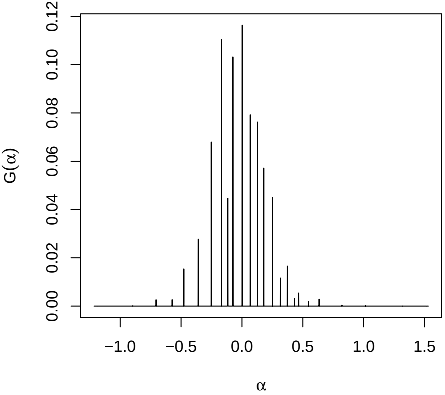
- We can identify and nonparametrically estimate the joint distribution of $(\alpha_j, \theta_j) \overset{iid}{\sim} G$ where arbitrary dependence is allowed. (Gu and Koenker (2017a, b))
- Bayes rule is a nonlinear function of (y_j, S_j): $\delta_j = \mathbb{E}(\alpha | y_j, S_j)$.

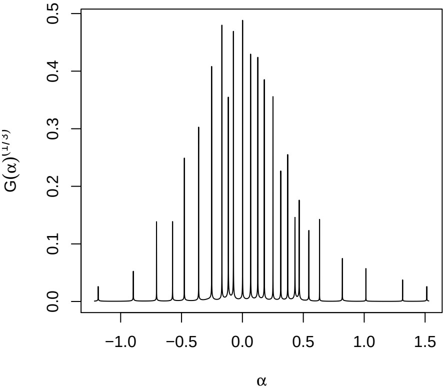
Unbalanced Panel

# of occurrence	Absolute	%	% cumulative
1	10180	29.00	29.00
2	6486	18.50	47.50
3	4706	13.40	61.00
4	3217	9.20	70.10
5	2446	7.00	77.10
6	1910	5.40	82.60
7	1281	3.70	86.20
8	1120	3.20	89.40
9	975	2.80	92.20
10	735	2.10	94.30
11	676	1.90	96.20
12	588	1.70	97.90
13	302	0.90	98.80
14	249	0.70	99.50
15	182	0.50	100.00
Total	35053		100%

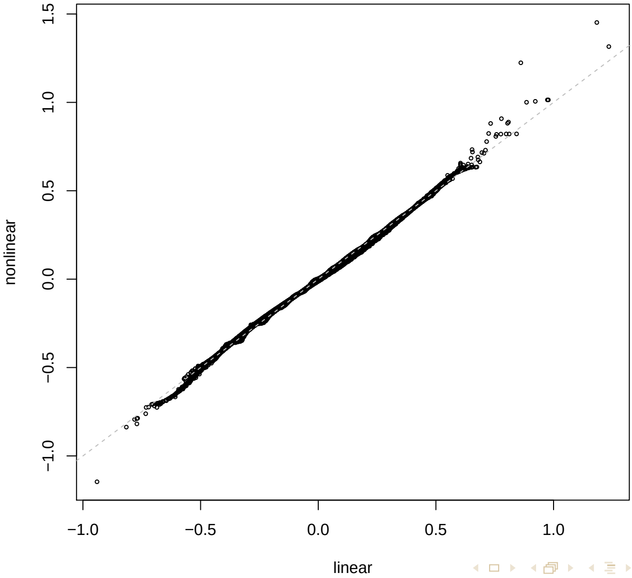
Estimated Distribution using North Carolina Data

- Linear shrinkage under Gaussian assumptions
  - $\alpha_j \sim \mathcal{N}(0, 0.047)$.
  - $\hat{\sigma}_\epsilon^2 = 0.249$.
- NPMLE $\hat{G}$

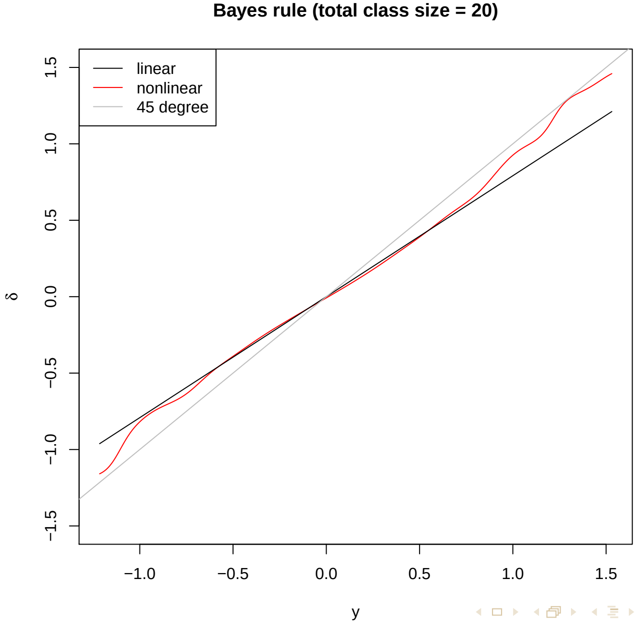




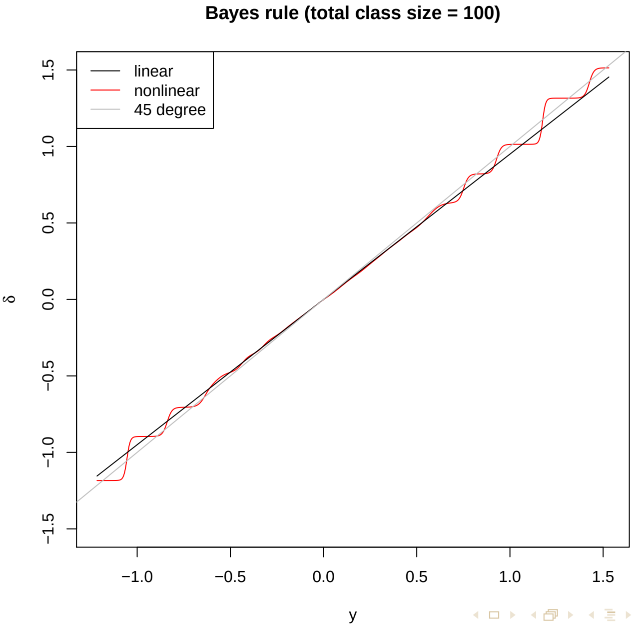
Effect estimation: linear vs nonlinear



Bayes Rule: linear vs nonlinear



Bayes Rule: linear vs nonlinear



Policy Evaluation

- All the action seems to be in the tail. But this is exactly what is relevant for educational policy (Chetty et al. 2014).
- Left tail policy: evaluate the magnitude of test score gains by replacing bottom $q\%$ of the teachers by a mean quality teacher (zero effect).

$$\mathbb{E}[\alpha 1\{\alpha > G^{-1}(q)\}] = \int_{G^{-1}(q)}^{+\infty} \alpha dG(\alpha)$$

- Depending on the thickness of the true distribution tail, this gain can be over/under estimated if the Gaussian effect assumption is misplaced.

How do we pick these teachers?

Empirical Bayes Ranking

- One approach is to rank the teachers by posterior mean. But although $\mathcal{L}_2$ loss is natural for effect estimation,it may not be natural for selecting good/bad teachers.
- There are some available alternatives in the literature, notably posterior expected rank: Laird and Louis (1989), Xie, Singh, Zhang (2009)
- We've come up with two types of loss function that leads to
  - ranking criteria based on posterior tail probability $\mathbb{P}(\alpha \leq G^{-1}(q)|y, \sigma)$ (see also Henderson and Newton (2016))
  - ranking criteria based on posterior expected shortfall $\mathbb{E}[\alpha 1\{\alpha \leq G^{-1}(q)\}|y, \sigma]$.
- How to choose loss function? Economists/education policy maker may be able to link loss function specification to welfare consideration.

Tail probability rule

- Let $\alpha_q := G^{-1}(q)$, consider loss function for a binary action $\delta_j : (y_j, \sigma_j^2) \mapsto \{0, 1\}$

$$L(\delta_j, \alpha_j) = (1 - \delta_j)1\{\alpha_j \leq \alpha_q\}$$

- Loss function only considers the tail population instead of the whole.
- Minimizing the Bayes risk subject to a size constraint $\mathbb{P}(\delta_j = 1) = q$ leads to the Bayes rule $\delta_j = 1\{v_q(y_j, \sigma_j) \geq \lambda_q\}$ with

$$v_q(y_j, \sigma_j) = \mathbb{P}(\alpha \leq \alpha_q | y_j, \sigma_j) = \frac{\int_{-\infty}^{\alpha_q} \phi((y_j - \alpha)/\sigma_j) dG(\alpha)}{\int \phi((y_j - \alpha)/\sigma_j) dG(\alpha)}$$

- Choose λ_q to satisfy the size constraint.
- Under mild conditions, which are satisfied for the normal model, there is a nested structure of the set $\Omega_q = \{j : v_q(y_j, \sigma_j) \geq \lambda_q\}$: $\Omega_{q_2} \subseteq \Omega_{q_1}$ for $q_1 > q_2$.
- A close connection to multiple testing problem: $v_q(y_j, \sigma_j)$ is one minus the local FDR (Efron et al. 2001, Sun and McLain 2012))
  - Composite one-sided null $H_{0j} : \alpha_j \geq \alpha_q$.
  - FDR literature: thresholding value on $v_q(y_j, \sigma_j)$ is chosen to satisfy FDR size restriction.

Expected shortfall rule

- Introduce effect size weights into the previous loss function, focusing on lower tail $\alpha_q < 0$

$$L(\delta_j, \alpha_j) = -\alpha_j(1 - \delta_j)1\{\alpha_j \leq \alpha_q\}$$

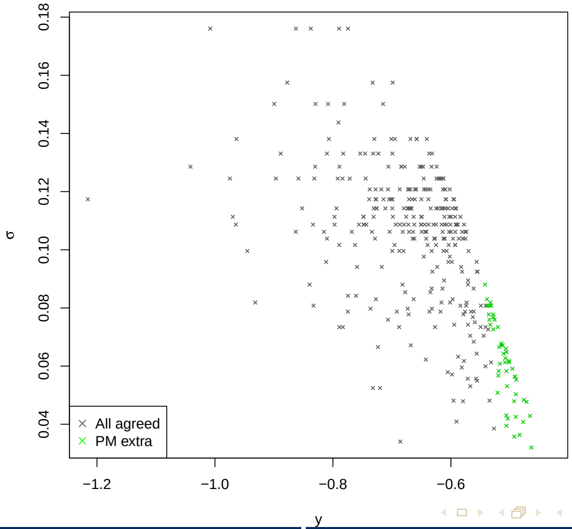
- Loss function has the interpretation as the lost gain of not replacing teacher j with a (better) mean teacher.
- Minimizing the Bayes risk subject to a size constraint $\mathbb{P}(\delta_j = 1) = q$ leads to the Bayes rule $\delta_j = 1\{s_q(y_j, \sigma_j) \geq \tau_q\}$ with

$$s_q(y_j, \sigma_j) = -\mathbb{E}(\alpha 1\{\alpha \leq \alpha_q\} | y_j, \sigma_j) = -\frac{\int_{-\infty}^{\alpha_q} \alpha\phi((y_j - \alpha)/\sigma_j)dG(\alpha)}{\int \phi((y_j - \alpha)/\sigma_j)dG(\alpha)}$$

- Choose τ_q to satisfy the size constraint.

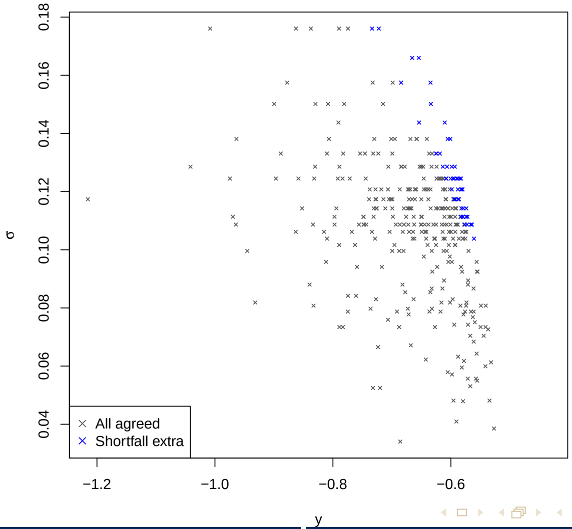
Comparison: q = 1% Posterior Mean

- grey points: agreed by both tailp, shortfall and posterior mean (201 teachers)
- green points: extra 49 teachers selected by posterior mean criteria.



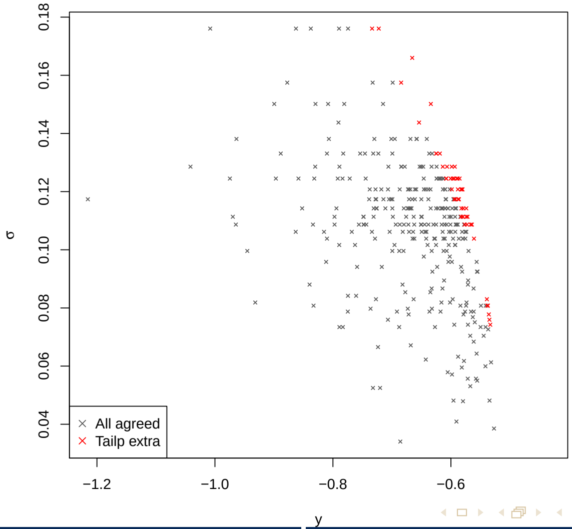
Comparison: q = 1% Shortfall

- grey points: agreed by both tailp, shortfall and posterior mean (301 teachers)
- blue points: extra 49 teachers selected by shortfall criteria.



Comparison: q = 1% Tailp

- grey points: agreed by both tailp, shortfall and posterior mean (301 teachers).
- red points: extra 49 teachers selected by tailp criteria.



Conclusions

- Teacher evaluation is involved in high-stakes decision making.
- We show the possibility of deviating from the Gaussian assumption and linear shrinkage rules and that it is empirically relevant.
- Efron's G-modeling: We take a nonparametric approach for G, which seems to open doors to many different Bayes rules depending on the type of loss function under consideration.