Empirical Bayes Methods: Theory and Applications

Jiaying Gu, University of Toronto
Christopher R. Walters, UC Berkeley and NBER

July 28, 2022
NBER Summer Institute 2022 Methods Lectures

Credits Due

Roger Koenker

Overview

Topics:

- What is a compound decision problem and its application in economics.
- Empirical Bayes estimators and how they perform.
  - Normal mean problem: parametric vs nonparametric shrinkage
  - Computation methods.
- Other compound decisions: testing; ranking/selection
- Empirical Bayes Inference
- (If time permits) Beyond Normal model : Poisson and mixture models in general

Motivating Example: Teacher Value Added

- Let $\tilde{Y}_{ij}$ be the test outcomes of student j taught by teacher i:

$$\tilde{Y}_{ij} = \alpha_i + X_{ij}'\beta + u_{ij}, \quad i = 1, \ldots, n; \quad j = 1, \ldots J_i.$$

- X_{ij} captures family background, lagged test outcomes etc.
- α_i is the value added of teacher i (also known as teacher fixed effects).
- Estimating α_i for all i is of interest:
  - Evaluation of teachers.
  - Understand heterogeneity of teacher quality.
  - Select top/bottom quality teachers.
- Fixed effect estimator for α_i: $Y_i = J_i^{-1} \sum_j (\tilde{Y}_{ij} - X_{ij}'\hat{\beta})$.
- When J_i is reasonably large: $Y_i \mid \alpha_i, J_i \approx \mathcal{N}(\alpha_i, \sigma_u^2/J_i)$.

Compound Decision Problem

- We face n independent statistical decision problems (i.e. observe Y_i, estimate α_i).
- They have similar structure: $Y_i \mid \alpha_i \sim P_{\alpha_i}$.
- n is usually large in modern applications.
- We care about collective performance: estimator for the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ that has good performance as a whole.

- This is called **compound decision problem** (Robbins (1951)).
- In contrast to simple decision problem: ignore the ensemble of n problems and construct estimator for α_i individually.

Assumptions on $\boldsymbol{\alpha}$

Two types of assumptions on $\boldsymbol{\alpha}$:

- Fixed Effect: $\alpha_1, \ldots, \alpha_n$ treated as unknown parameters.
- Random Effect: $\alpha_1, \ldots, \alpha_n$ a vector of random variables with distribution G (i.e. $\alpha_i \sim_{iid} G$).

- Sometimes the literature reserves the name "compound decision" to the fixed effect model while using "empirical Bayes" to refer to the random effect model.

Compound Decision Problem: Formulation

- Consider a population of individuals indexed by i, and Y_i measures some outcomes with the model $Y_i \mid \alpha_i \sim P_{\alpha_i}$.
- We would like to estimate α_i for all i.
- Vector of estimators for $\boldsymbol{\alpha}$ as: $\boldsymbol{\delta}(\boldsymbol{Y}) = \Big(\delta_1(\boldsymbol{Y}), \ldots, \delta_n(\boldsymbol{Y})\Big)$ with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$.
- Let the loss function be $L(\boldsymbol{\alpha}, \boldsymbol{\delta}(\boldsymbol{Y})) = \frac{1}{n}\sum_{i=1}^{n} L(\alpha_i, \delta_i(\boldsymbol{Y}))$
  - Squared error loss: $L(\alpha_i, \delta_i) = (\alpha_i - \delta_i)^2$.
  - Absolute error loss: $L(\alpha_i, \delta_i) = |\alpha_i - \delta_i|$.
- Compound Risk (i.e. expected Loss):

$$
\begin{aligned}
R_n(\boldsymbol{\alpha}, \boldsymbol{\delta}(\boldsymbol{Y})) &= \frac{1}{n}\mathbb{E}\sum_{i=1}^{n} L(\alpha_i, \delta_i(\boldsymbol{Y})) \\
&= \frac{1}{n}\sum_{i=1}^{n} \int \ldots \int L(\alpha_i, \delta_i(y_1, \ldots, y_n)) dP_{\alpha_1}(y_1) \ldots, dP_{\alpha_n}(y_n)
\end{aligned}
$$

  - Estimation risk of each individual matters, but these risks are aggregated.

Example: Normal Mean Problem

With $P_{\alpha_i} = \mathcal{N}(\alpha_i, 1)$ and $L(\alpha_i, \delta_i) = (\alpha_i - \delta_i)^2$, we have the **normal mean problem**:

$$Y_i = \alpha_i + u_i, u_i \sim \mathcal{N}(0,1), i = 1, \ldots, n$$

- Goal: find $\boldsymbol{\delta}(\boldsymbol{Y})$ to make the compound risk small.

Two well-known estimators:

- Fixed Effect Estimator (i.e. MLE): $\delta_i^{MLE}(\boldsymbol{Y}) = Y_i$, only use information from individual i.
- Linear Shrinkage estimator (James and Stein (1961)): $\delta_i^{JS}(\boldsymbol{Y}) = (1 - \frac{n-2}{S})Y_i$ with $S = \sum_i Y_i^2$, use all $Y_1, \ldots, Y_n$ for data dependent shrinkage of each i.

- Stein (1956), James and Stein (1961): $R_n(\boldsymbol{\alpha}, \boldsymbol{\delta}^{JS}) < 1 = R_n(\boldsymbol{\alpha}, \boldsymbol{\delta}^{MLE})$ as soon as $n \geq 3$ for any vector $\boldsymbol{\alpha}$. (i.e. MLE is inadmissible).
- This is a finite sample frequentist result.

Why data dependent shrinkage?

- Consider the class of linear shrinkage estimators of the form

$$\boldsymbol{\delta}(\boldsymbol{Y}) = \{(1-b)Y_1, \ldots, (1-b)Y_n\}$$

  for some $b \geq 0$.
- Then compound risk $R_n(\boldsymbol{\delta}(\boldsymbol{Y}), \boldsymbol{\alpha}) = (1-b)^2 + b^2 \sum_i \alpha_i^2/n$.
- Minimize: $b^* = \text{argmin}_b R_n(\boldsymbol{\delta}_b(Y), \boldsymbol{\alpha}) = n/\sum_i(1+\alpha_i^2)$.
- Optimal b^* depends on $\boldsymbol{\alpha}$, but only through $\sum_i \mathbb{E}(Y_i^2)$.
  - Recall $Y_i = \alpha_i + u_i$, $u_i \sim \mathcal{N}(0,1)$, then $\mathbb{E}(Y_i^2) = \alpha_i^2 + 1$.

- James-Stein suggested $\hat{b}^* = (n-2)/\sum_i Y_i^2$ and showed

$$R_n(\delta^{JS}, \boldsymbol{\alpha}) \leq \frac{2}{n} + \frac{1}{n}\frac{(n-2)\sum_i \alpha_i^2}{(n-2)+\sum_i \alpha_i^2}$$

- When $\alpha_i = 0$ for all i, biggest improvement: Risk bounded by $2/n$ (MLE risk $= 1$).
- Always improvement provided $n \geq 3$.
- Shrinkage: introduce bias to individual estimators to improve overall performance.

Bayesian Interpretation of Frequentist's Shrinkage (Efron and Morris)

- Consider the random effect assumption: $\alpha_i \sim_{iid} G = \mathcal{N}(0, A)$.
- We have a Bayesian model: G is the subjective prior on α_i.
- A natural optimal principle: minimize Bayes risk: $\mathbb{E}_G\Big(\mathbb{E}_\alpha(L(\alpha, \delta(Y)))\Big)$ leads to optimal Bayes estimator.
- With squared error loss: optimal Bayes estimator is $\mathbb{E}[\alpha|Y_i]$.
- For nomal mean problem, $\delta_i^{Bayes}(\boldsymbol{Y}) = (1 - \frac{1}{A+1})Y_i$.
- δ_i^{Bayes} shrinks Y_i towards zero (the prior mean).

- Since $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, (A+1)I_n)$, then $S = \sum_i Y_i^2 \sim (A+1)\chi_n^2$ and $\mathbb{E}[\frac{n-2}{S}] = \frac{1}{A+1}$.
- Hence $\boldsymbol{\delta}^{JS}$ estimator replaces $\frac{1}{A+1}$ in $\boldsymbol{\delta}^{Bayes}$ by an unbiased estimator, giving rise to the name empirical Bayes.

- With $G = \mathcal{N}(0, A)$, $\boldsymbol{\delta}^{JS}$ mimics the optimal Bayes estimator.
- But recall Stein's result holds without any Bayesian assumption on $\boldsymbol{\alpha}$, there is always an improvement in MSE.

Extensions

Positive part of James-Stein estimator:

- Since $1 - \frac{1}{A+1} > 0$, $\delta_i^{JS+} = (1 - \frac{n-2}{S})_+ Y_i$ provides further improvement especially with small n.

Non-zero prior mean: Consider instead $\alpha_i \sim_{iid} G = \mathcal{N}(\alpha_0, A)$.

- The Bayes estimator shrinks towards α_0

$$\delta_i^{Bayes}(\boldsymbol{Y}) = \alpha_0 + (1 - \frac{1}{A+1})(Y_i - \alpha_0)$$

- The corresponding James-Stein estimator:

$$\boldsymbol{\delta}_i^{JS}(\boldsymbol{Y}) = \bar{Y} + (1 - \frac{n-3}{\sum_i (Y_i - \bar{Y})^2})(Y_i - \bar{Y})$$

  with $\bar{Y} = n^{-1} \sum_i Y_i$
- Dominates MLE when $n \geq 4$.

Other variants in Chris's section.

Baseball Batting Averages

- Efron and Morris (1975): 18 MLB baseball players. We observe the hits (H_i) for the first 45 at bats ($N_i = 45\ \forall i$), and we want to predict for the remainder of the season the hitting performance $H_i \sim Binom(N_i, p_i)$.
- Variance-stabilizating transformation:

$$Y_i = \sqrt{N_i} arcsin(2H_i/N_i - 1) \approx N(\alpha_i, 1), \quad \alpha_i = \sqrt{N_i} arcsin(2p_i - 1)$$

- Mean Squared Errors (using the remaining season's realized p_i to calculate α_i).

	MLE	James-Stein
$n^{-1}\sum_i(\alpha_i - \delta_i)^2$	1.11	0.348

Further Improvements?

Under fixed effect assumption:

- James-Stein estimator mimics optimal estimator in the class $\delta(\boldsymbol{Y}) = \{(1-b)Y_1, \ldots, (1-b)Y_n\}$ with $b \geq 0$.
- Can we do better if we enlarge the class of estimators?

Under random effect assumption:

- James-Stein estimator mimics the optimal Bayes rule when prior $G = \mathcal{N}$.
- Can we do better if prior $G \neq \mathcal{N}$?

Revisit the Compound Decision Problem Robbins (1951, 1956)

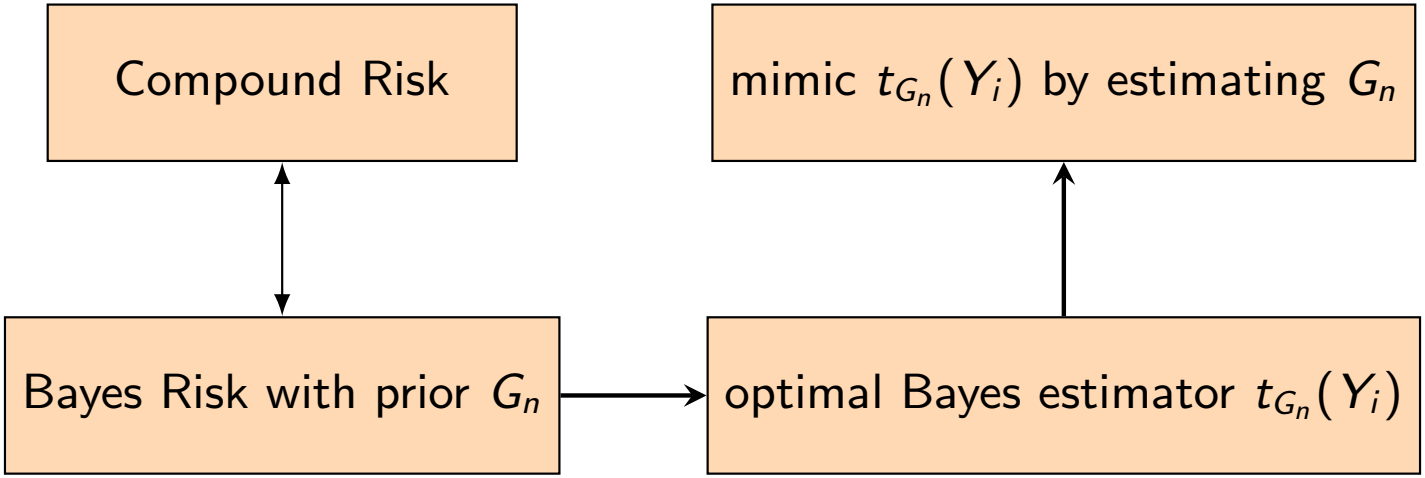
- Fixed Effect Model: $\boldsymbol{\alpha}$ is a vector of unknown parameters.
- Given α_i, suppose Y_i has density $p(\cdot \mid \alpha_i)$.
- Consider the class of *separable* estimator $\boldsymbol{\delta}(\boldsymbol{Y}) = \{t(Y_1), \ldots, t(Y_n)\}$ for some fixed function $t : \mathbb{R} \to \mathbb{R}$.
- Compound Risk:

$$
\begin{aligned}
R_n(\boldsymbol{\alpha}, \boldsymbol{\delta}(\boldsymbol{Y})) &= \frac{1}{n}\mathbb{E}\sum_{i=1}^{n} L(\alpha_i, \delta_i(\boldsymbol{Y})) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\ L(\alpha_i, t(Y_i)) && \text{(separable estimator)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \int L(\alpha_i, t(y)) p(y \mid \alpha_i) dy && \text{(expectation w.r.t. } Y_i|\alpha_i) \\
&= \int \int L(\alpha, t(y)) p(y \mid \alpha) dy dG_n(\alpha) && \text{(Bayes risk with prior } G_n)
\end{aligned}
$$

where G_n is the empirical CDF of $(\alpha_1, \ldots, \alpha_n)$ (i.e. $G_n(u) = n^{-1}\sum_i 1\{\alpha_i \le u\}$).

Fundamental Theorem of Compound Decision: Compound risk is equivalent to the Bayes risk of a single copy of the compound problem with prior G_n.

Nonparametric Empirical Bayes



- Holds for any $p(\cdot \mid \alpha_i)$ and any L.
- Not knowing the prior G_n, optimal Bayes estimator is not feasible.
- $Y_1, \ldots, Y_n$ contains information about $\alpha_1, \ldots, \alpha_n$, use them to estimate G_n.
- James-Stein estimates the second moment of G_n (recall $b^* = n/\sum_{i=1}^n (\alpha_i^2 + 1)$).
- Robbins: estimate the whole distribution G_n, hence nonparametric empirical Bayes.
- It is more ambitious than linear shrinkage as it targets the minimum of compound risk over a larger class of estimator.

Nonparametric EB for normal mean problem

With $L(\alpha, t(y)) = (\alpha - t(y))^2$, for any G_n, the optimal Bayes estimator is

$$t_{G_n}(Y_i) = \mathbb{E}[\alpha | Y_i] = \frac{\int \alpha p(Y_i \mid \alpha) dG_n(\alpha)}{\int p(Y_i \mid \alpha) dG_n(\alpha)}$$

- **g-modeling**: G_n unknown, estimate from data and plug in.
- applicable for any form of $p(y|\alpha)$.
- Estimating G_n is a deconvolution problem with $Y_i = \alpha_i + u_i$.

If $Y_i = \alpha_i + u_i$, $u_i \sim \mathcal{N}(0, 1)$, the Bayes estimator simplifies to the Tweedie formula:

$$t_{G_n}(Y_i) = Y_i + \frac{f'(Y_i)}{f(Y_i)}$$

with $f(Y) = \int \phi(Y - \alpha) dG_n(\alpha)$.

- **f-modeling**: $t_{G_n}(Y_i)$ depends on G_n only through $f(Y)$, directly estimate $f(Y)$.
- specific to the Tweedie formula and thus normal model.

Can g-modeling backfire since we need to estimate G_n nonparametrically?

Theoretical Guarantee of nonparametric empirical Bayes estimator

- Consider a plug-in rule $t_{\hat{G}_n}(Y)$ with nonparametric MLE $\hat{G}_n$.
- Minimum Risk target:

$$R^*(G_n) = \min \int \int (\alpha - t(y))^2 \phi(y - \alpha) dy dG_n(\alpha).$$

- Jiang and Zhang (2009): $t_{\hat{G}_n}(Y)$ is asymptotically optimal among all separable estimators: as $n \to \infty$,

$$r_n(t_{\hat{G}_n}) := R_n(\boldsymbol{\alpha}, t_{\hat{G}_n}) - R^*(G_n) = o(1) R^*(G_n)$$

uniformly for a wide collection of vectors $\boldsymbol{\alpha}$.

What about variance heterogeneity?

Recall Example (Teacher Value Added)

- Teacher-specific mean (i.e. Fixed Effect Estimator):

$$Y_i := J_i^{-1}(\tilde{Y}_{ij} - X_{ij}'\hat{\beta}) \approx \mathcal{N}(\alpha_i, \sigma_u^2/J_i)$$

- If we assume $J_i \perp \alpha_i$, then we can proceed similarly and deconvolve to estimate G_n of $\boldsymbol{\alpha}$ nonparametrically (deconvolution under heterogeneous variances).

- We can also consider a richer model:

$$\tilde{Y}_{ij} = \alpha_i + X_{ij}'\beta + u_{ij}, u_{ij} \sim \mathcal{N}(0, \theta_i), (\alpha_i, \theta_i) \sim G_n$$

and estimate (β, G_n) from the panel data (Gu and Koenker (2017), Soloff et al. (2021)).

Parametric or Nonparametric?

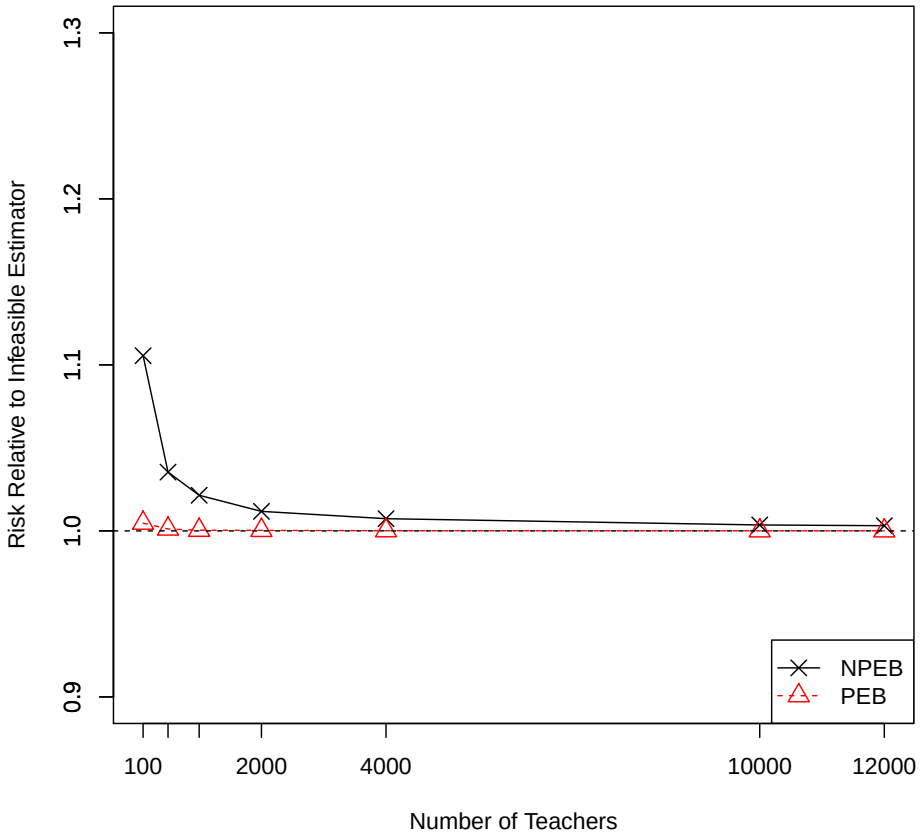
In finite samples, the performance difference between parametric and nonparametric EB depends on the underlying G and the sample size n.

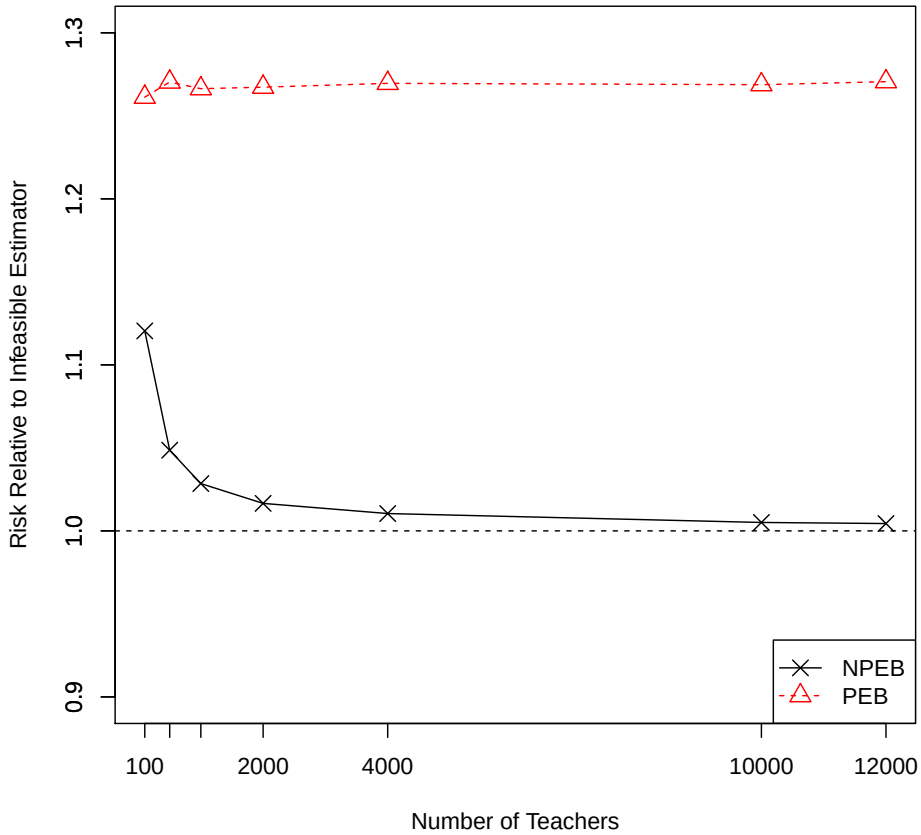
NPEB offers an adaptive approach: as long as n is not too small,

- not worse off when $G \approx \mathcal{N}$
- but can be much better for other G.

A Simulation Example (Gilraine, Gu, McMillan (2020))

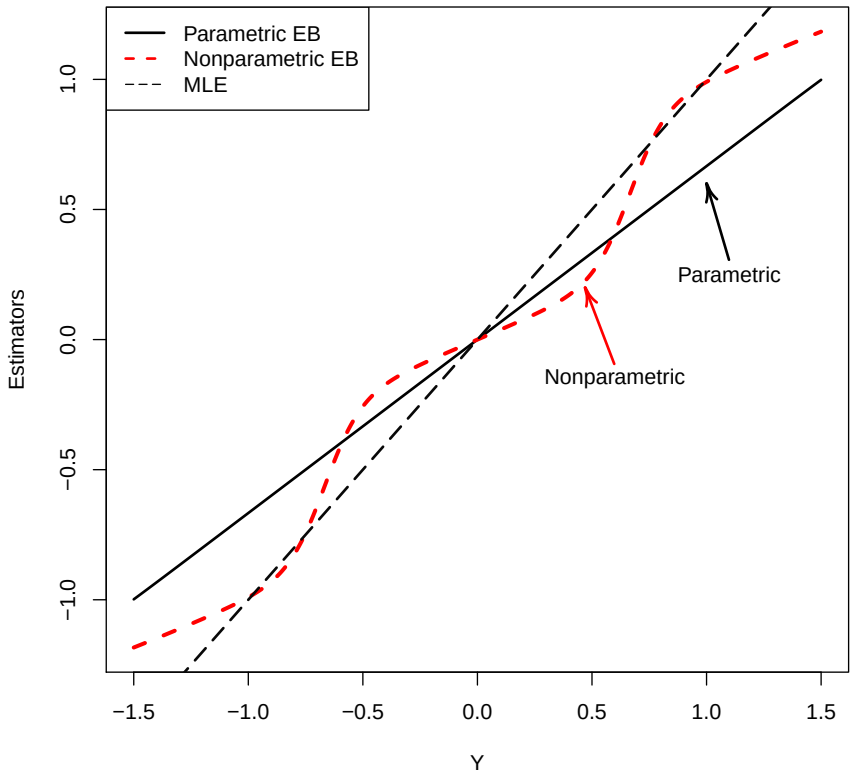
- $Y_i = \alpha_i + u_i$, $i = 1, \ldots, n = \{100, 500, 1000, 2000, 10000, 12000\}$, $u_i \sim \mathcal{N}(0, 0.25/J)$, $J \in \{8, 16\}$ with equal prob.
- DGP 1: Normal $G = \mathcal{N}(0, 0.08)$
- DGP 2: mixed-Normal $G = 0.95\mathcal{N}(0, 0.03) + 0.025\mathcal{N}(-1, 0.03) + 0.025\mathcal{N}(1, 0.03)$ [Same mean and variance as G in DGP 1].
- $\hat{G}_n$ estimated using NPMLE in `REBayes` R package.
- Plots Risk ratio (Estimator vs Infeasible with known G)



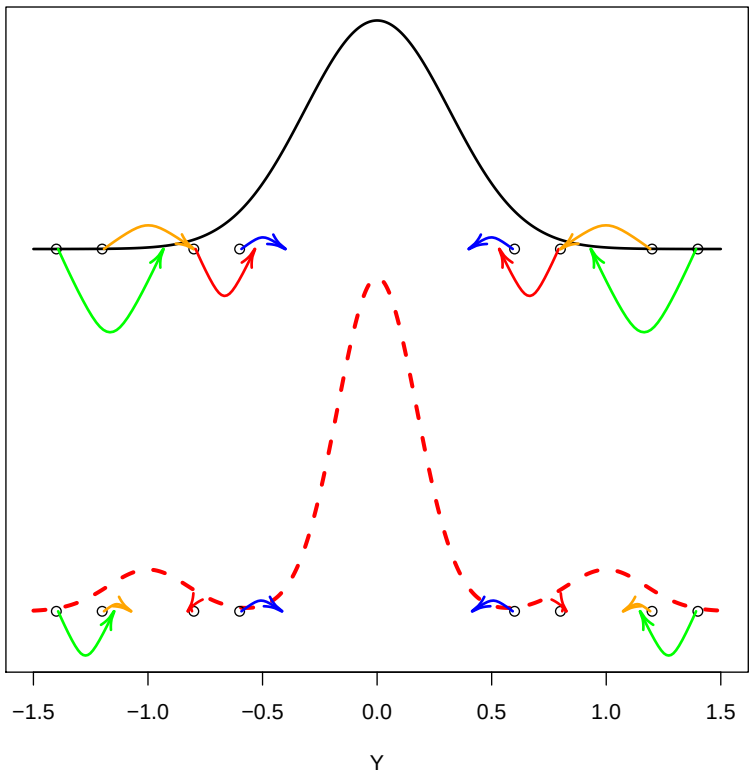


Visualize Shrinkage

Consider G in DGP 2.
Left: solid (–)PEB, dashed (- -) NPEB, long-dash (- -)MLE 45% line.
Right: upper shows amount of shrinkage under PEB, below shows that of NPEB.



(c) PEB vs NPEB vs MLE



(d) Amount of Shrinkage

Teacher Value Added (Gilraine, Gu, McMillan (2020))

- Los Angeles School District Primary School (Grade 3 - 5) student-teacher matched datasets: 11,000 teachers.
- Control for a rich set of covariates, teacher specific mean leads to

$$Y_i = \alpha_i + u_i, u_i \sim \mathcal{N}(0, \sigma_u^2/J_i), J_i \perp \alpha_i$$

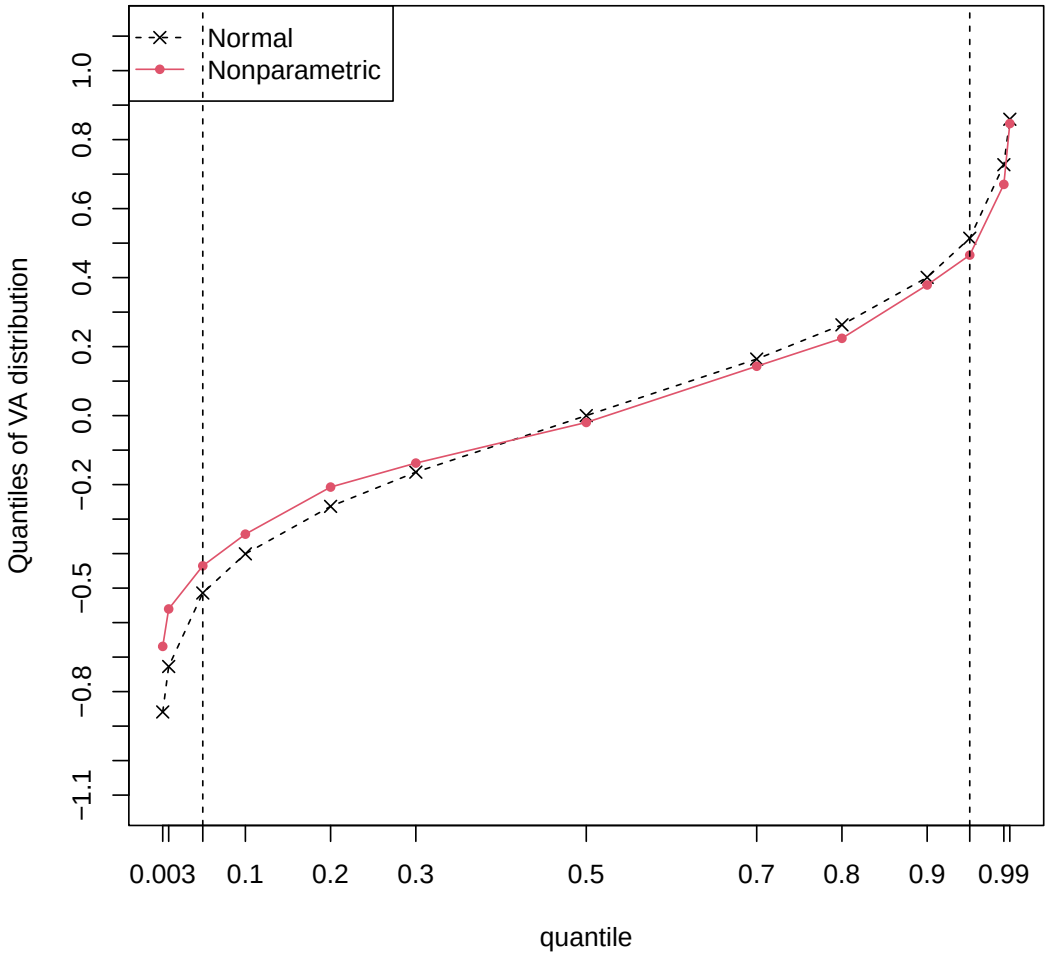
with Y_i is average test outcome (after controling for covariates) and J_i being the total number of students taught by teacher i.

We compare the differences of

- Parametric EB: assume $\alpha_i \sim \mathcal{N}(0, A)$. Kane and Kraiger (2008), Chetty et al. (2014).
- Nonparametric EB: estimate G nonparametrically. Gilraine, Gu, McMillan (2020)

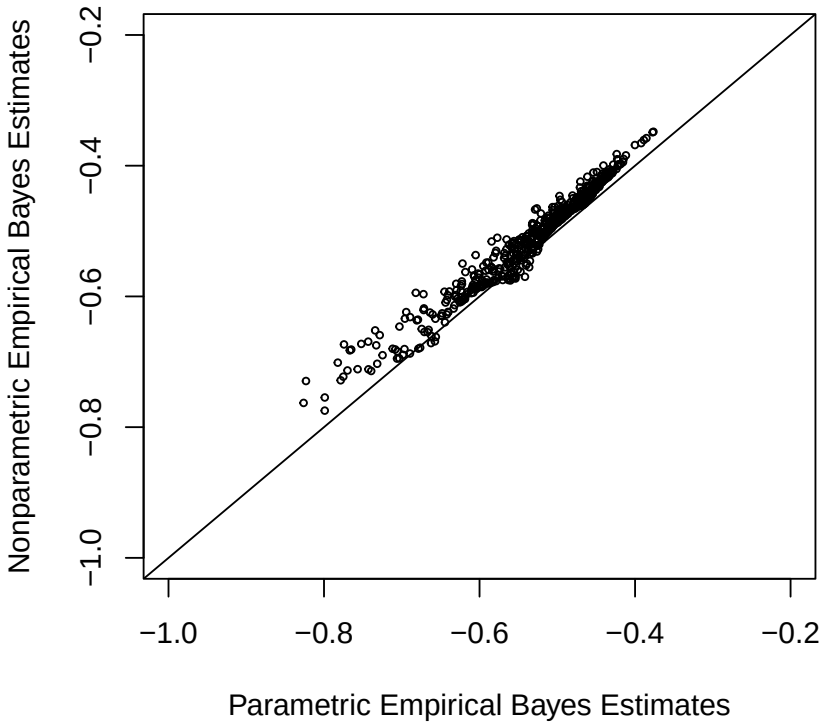
Estimated VA distribution for LA data

- Dash (-×-): Normality imposed; Solid (-•-) : Nonparametric MLE of G.
- Vertical dotted line: top/bottom 5%.
- Nonparametric $\hat{G}_n$ estimates an asymmetric distribution with smaller left tail [i.e. not as many low quality teachers].

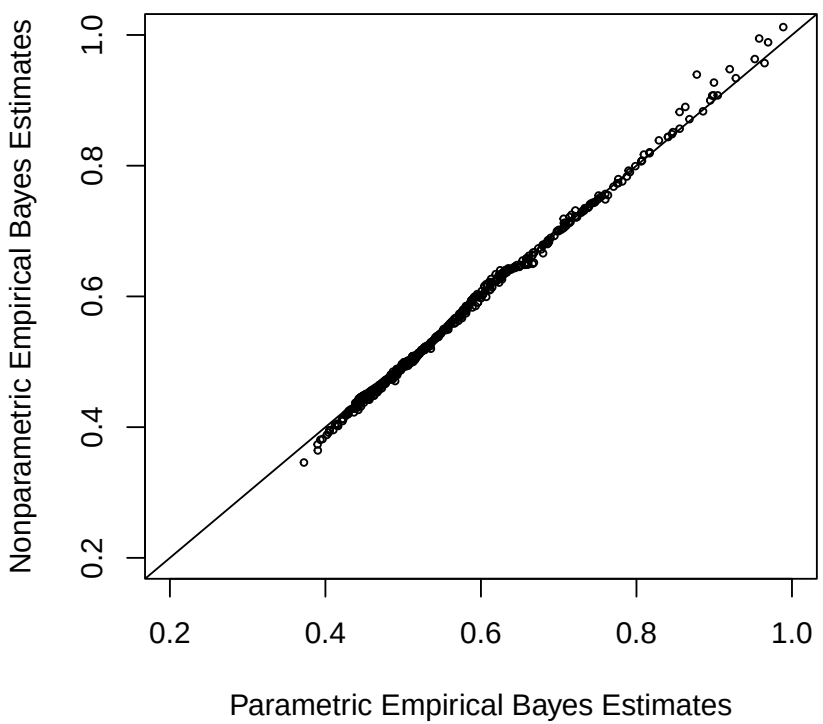


Parametric vs Nonparametric EB estimates

- Each dot represents a teacher within bottom/top 5% in Y_i.
- NPEB is more optimistic about teachers at the bottom.



(e) Bottom 5%



(f) Top 5%

Computation Methods

Computation Methods: Overview

Goal: nonparametrically estimate the Bayes rule:

$$t_G(Y) = \frac{\int \alpha p(y \mid \alpha) dG(\alpha)}{\int p(y \mid \alpha) dG(\alpha)}$$

when G is unknown.

Available tools:

- f-modeling in normal model: Brown and Greenshtein (2009).
- g-modeling: nonparametrically estimate G.
  - Nonparametric MLE: Kiefer and Wolfowitz (1956).
  - EM algorithm (Laird (1978)).
  - Interior Point algorithm: Koenker and Mizera (2014) `REBayes` R package
  - Efron's logspline g: `deconvolveR` package

Demo on two examples

What if G is not point identified?

f-modeling

- Recall in the normal mean problem, the Tweedie formula takes the form

$$t_G(Y) = \mathbb{E}[\alpha|Y] = \frac{\int \alpha\phi(Y-\alpha)dG(\alpha)}{\int \phi(Y-\alpha)dG(\alpha)} = Y + \frac{f'(Y)}{f(Y)}$$

  where $f(Y) = \int \phi(Y-\alpha)dG(\alpha)$ is the marginal density of Y.
- Brown and Greenshtein (2009): Estimate $f(Y)$ and $f'(Y)$ using kernel methods.
- Disadvantage:
  - Kernel method doesn't respect the hierarchical structure of the model (i.e. the fact that f is a mixture density).
  - As a consequence, estimator does not respect the fact that $t_G(Y)$ is monotone in Y ($t'_G(Y) = Var(\alpha|Y) > 0$).
- Improvement:
  - Koenker and Mizera (2014) imposes shape constraint to enforce monotonicity for the normal mean problem.

g-modeling: nonparametric MLE for G

$$\hat{G}_n := \underset{G \in \mathcal{G}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \log \left(\int \phi(y_i - \alpha) dG(\alpha) \right) \right\}$$

- For the normal mean problem G is identified: Gaussian deconvolution $Y = \alpha + u$.
- Kiefer and Wolfowitz (1956): As $n \to \infty$, $\hat{G}_n$ is a consistent estimator of G.
- Lindsay (1985): Solution $\hat{G}_n$ exists and is a discrete probability measure, with no more than n mass points in the interval $[\min_i y_i, \max_i y_i]$.
- Polyanskiy and Wu (2020): If the true G is sub-gaussian, then number of mass points is $\mathcal{O}(\log n)$ with high probability.

- We can generalize to

$$\max_{G \in \mathcal{G}} \left\{ \sum_{i=1}^{n} \log \left(\int p(y_i \mid \alpha) dG(\alpha) \right) \right\}$$

- As long as $p(\cdot \mid \alpha)$ belongs to the exponential family, Lindsay's result holds.

EM algorithm: Laird (1978)

- EM algorithm has been widely used for finite mixture model: e.g. Heckman and Singer (1984) for Weibull mixture.
- We pick a K (number of mass points) and the algorithm iteratively optimizes for locations and weights.
- We slowly increase K until likelihood no longer improves.

- This is a tough problem: non-convex.
  - the set of discrete distribution with K mass points is not a convex set.

- Relaxing finite mixture problem to infinite mixture problem restores convexity.

Interior Point algorithm

- Dual problem:

$$\max\left\{\sum_{i=1}^{n}\log v_i \mid \sum_{i=1}^{n} v_i\phi(y_i-\alpha)\le n \text{ for all } \alpha\right\}$$

- n variables, infinite number of dual constraints.
- Koenker and Mizera (2014) suggested taking a fixed and fine grid and enforce dual constraints on the grid:

$$\hat{G}_L = \min_{G\in\mathcal{G}_L}\left\{-\sum_{i=1}^{n}\log\left(\int\phi(y_i-\alpha)dG(\alpha)\right)\right\},$$

  where $\mathcal{G}_L$ is the class of probability measures supported on the fixed grid.

- This is still a convex optimization problem: modern interior point algorithm scales well with n and is very efficient.
- Implementation in `REBayes` R package using mosek.
- Generalizes to a variety of $p(\cdot \mid \alpha)$: Gaussian, Poisson, Binomial, Weibull, Gamma...

Efron's log-spline

- Again take a fixed grid $\{u_1, \ldots, u_L\}$.
- Assume G has a density g, belonging to exponential family with parameter $\boldsymbol{\mu} \in \mathbb{R}^p$:

$$g(\boldsymbol{\mu}) = \exp(Q\boldsymbol{\mu} - \psi(\boldsymbol{\mu}))$$

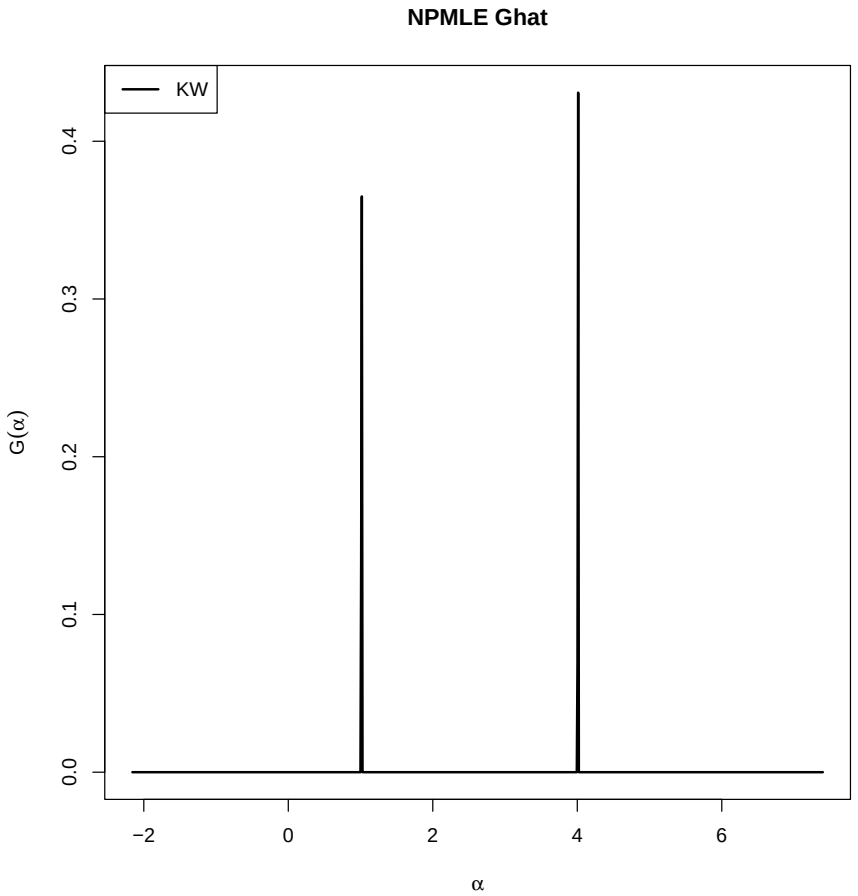
  where Q is a matrix of dimension $L \times p$. The ℓ-th entry is g evaluated at u_ℓ.
- Efron (2014) suggested using natural splines with p parameters.
- p is a user choice (larger p means more flexible).
- We then optimize $\boldsymbol{\mu}$ with a penalized likelihood:

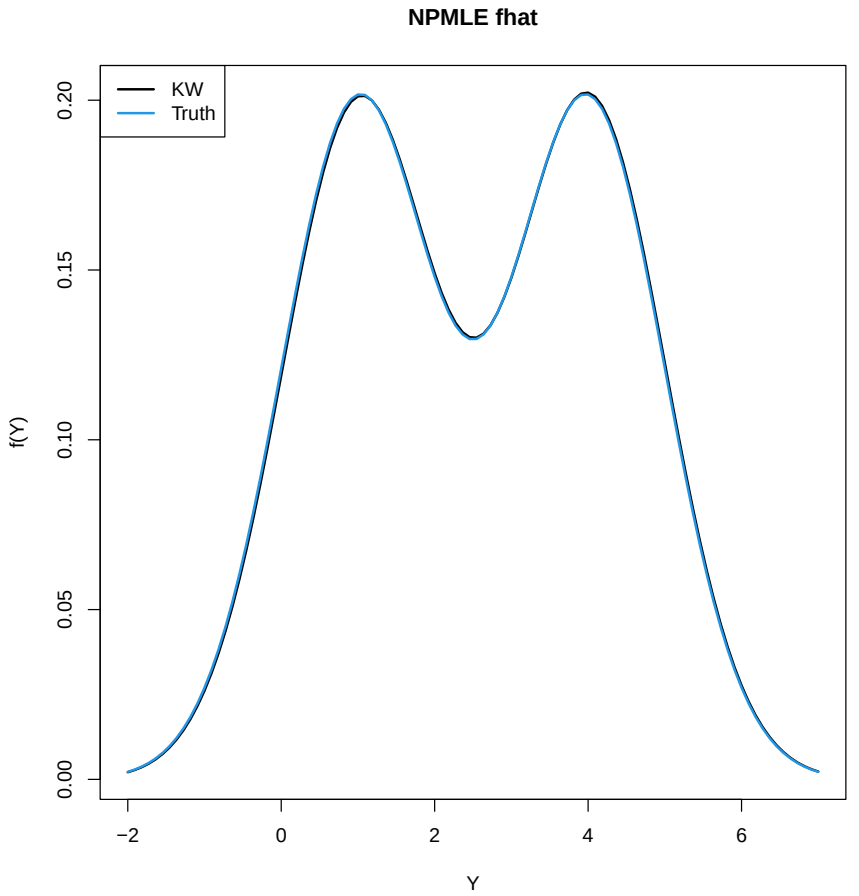
$$\max_{\boldsymbol{\mu}} \left\{ \sum_{i=1}^{n} \log \sum_{\ell=1}^{L} \phi(y_i - u_\ell) g_\ell(\boldsymbol{\mu}) - \lambda \|\boldsymbol{\mu}\| \right\}$$

- λ is another tuning parameter, shrinking $\boldsymbol{\mu}$ towards the origin, and penalize $\hat{G}$ towards uniform.

Demonstration: discrete G

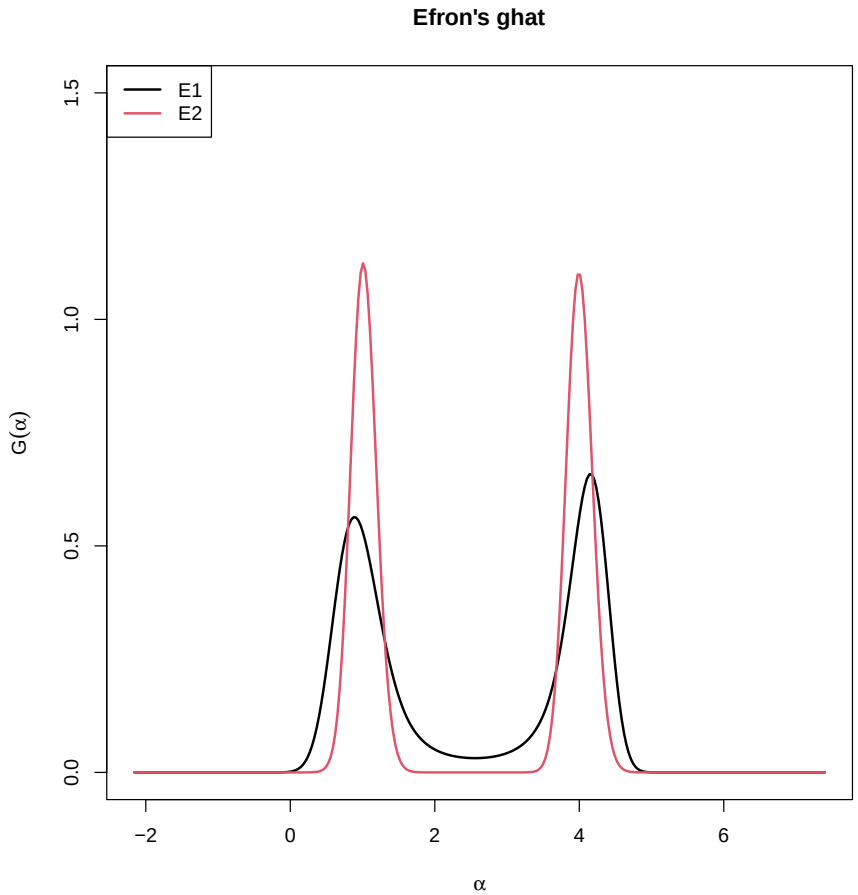
- $Y_i = \alpha_i + u_i, u_i \sim \mathcal{N}(0,1), \alpha_i \sim G = 0.5\delta_1 + 0.5\delta_4$.
- The NPMLE estimator $\hat{G}_n$ and the implied $\hat{f}_n$ with $n = 4000$. Truth (blue).

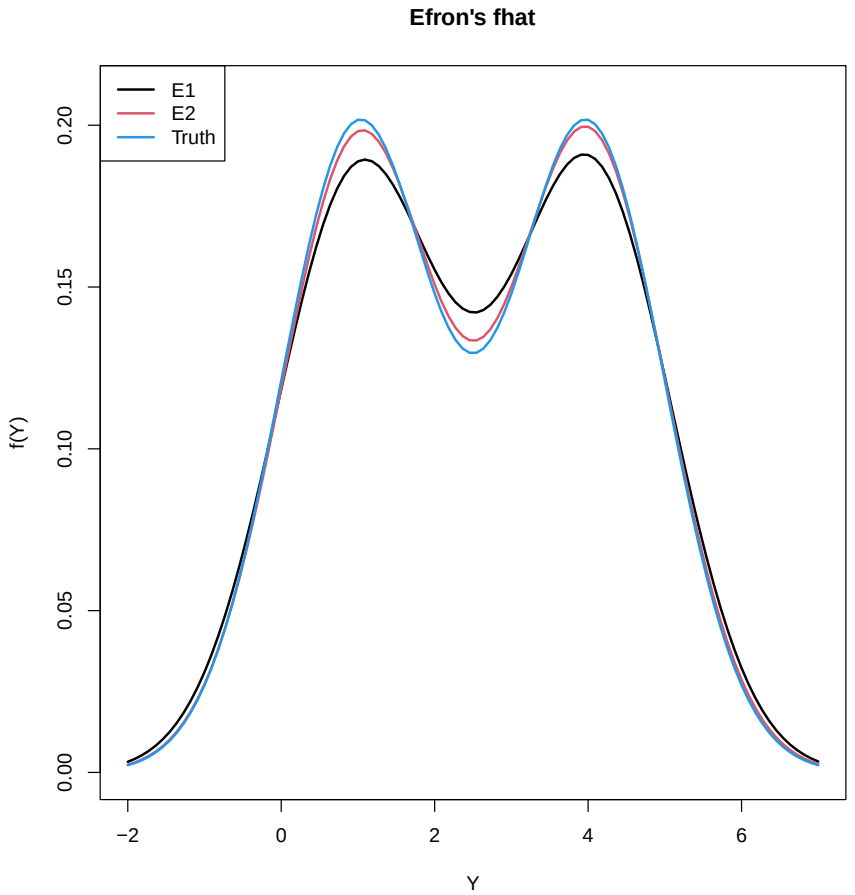




Demonstration: discrete G

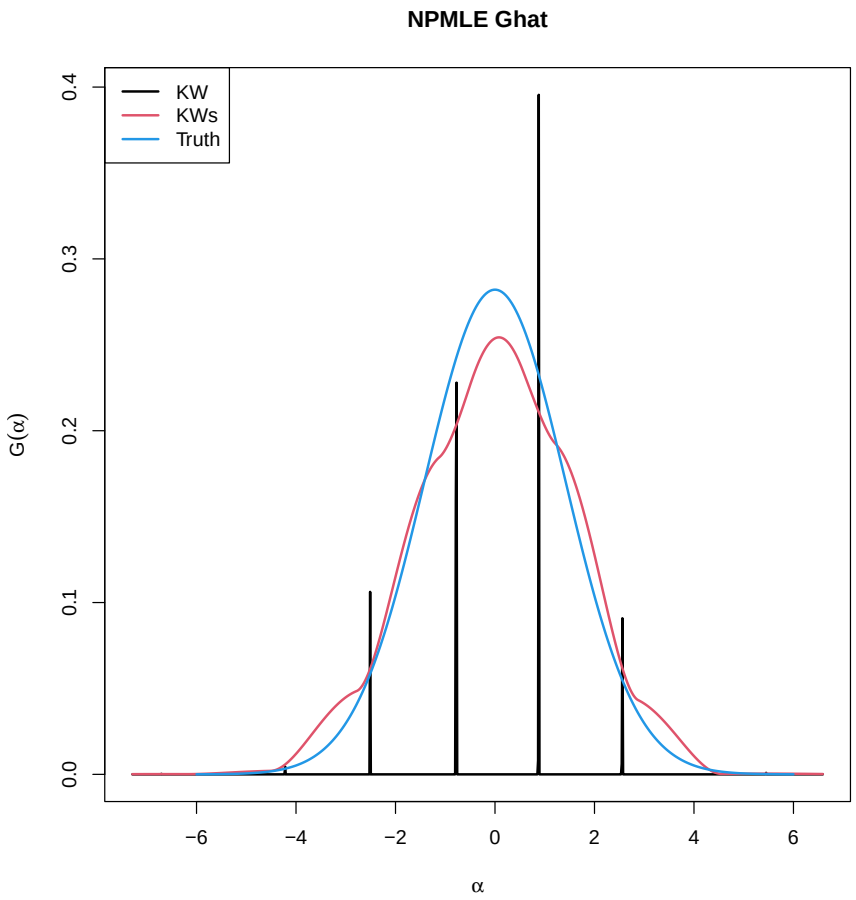
- $Y_i = \alpha_i + u_i, u_i \sim \mathcal{N}(0,1), \alpha_i \sim G = 0.5\delta_1 + 0.5\delta_4$.
- Efron's estimator $\hat{G}_n$ and the implied $\hat{f}_n$ with $n = 4000$.
- E1 (black): $p = 5, \lambda = 0.1$. E2 (red): $p = 10, \lambda = 0.1$. Truth (blue).

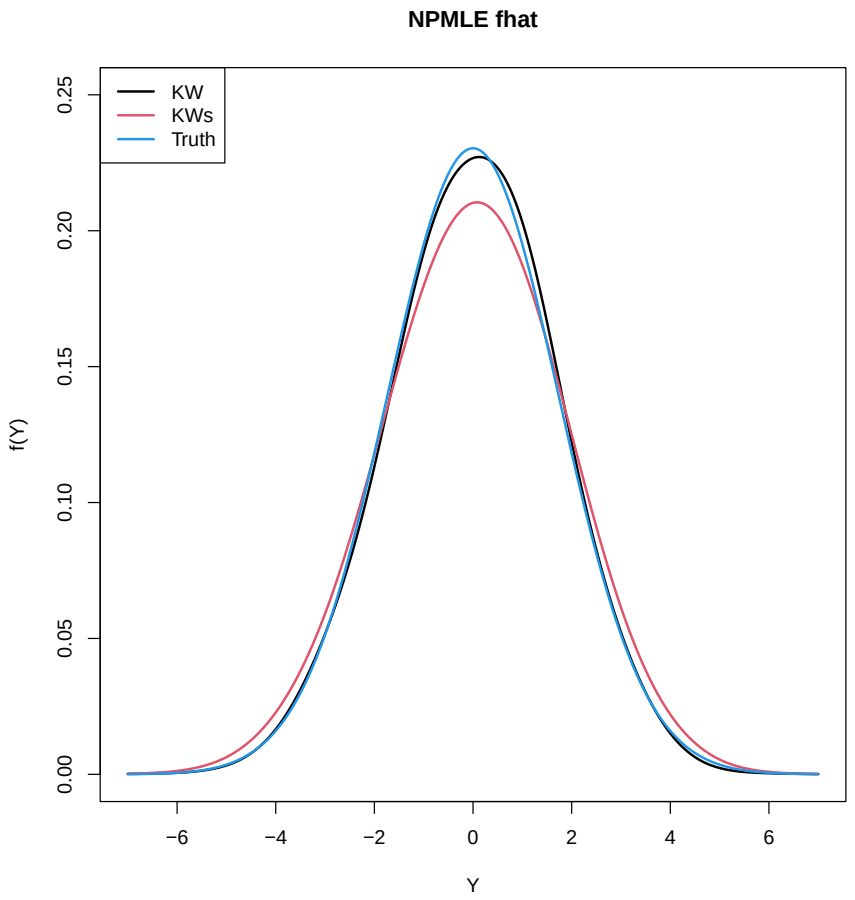




Demonstration: continuous G

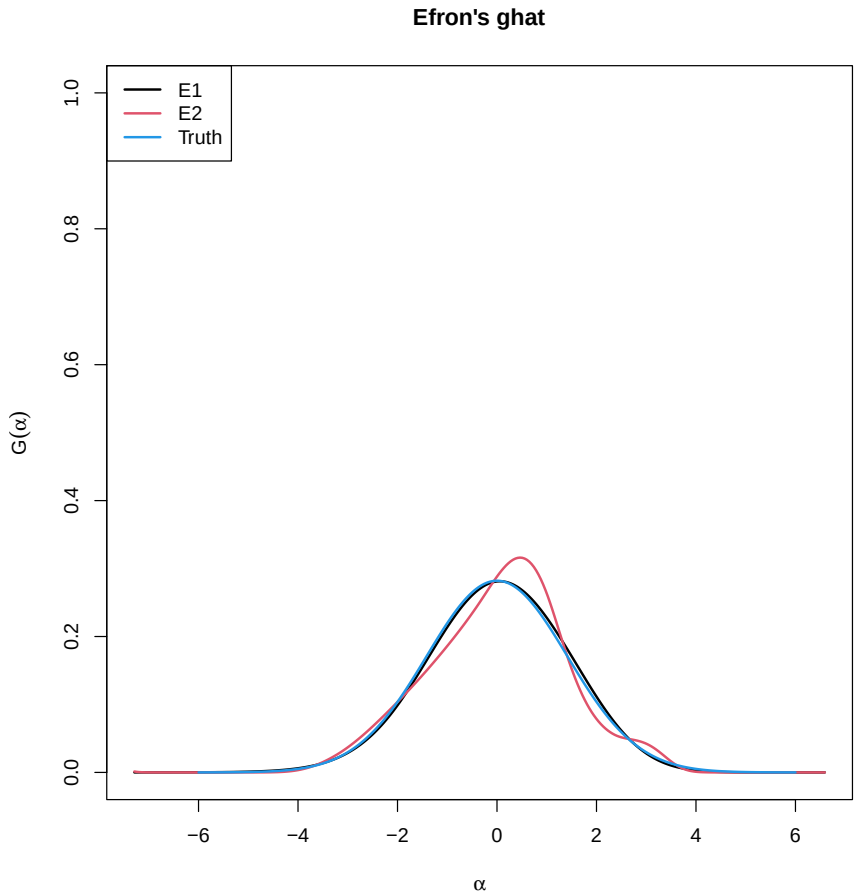
- $Y_i = \alpha_i + u_i, u_i \sim \mathcal{N}(0,1), \alpha_i \sim G = \mathcal{N}(0,2)$.
- The NPMLE estimator (black) $\hat{G}_n$ and its kernel smoothed version (red) and their implied $\hat{f}_n$ with $n = 4000$.

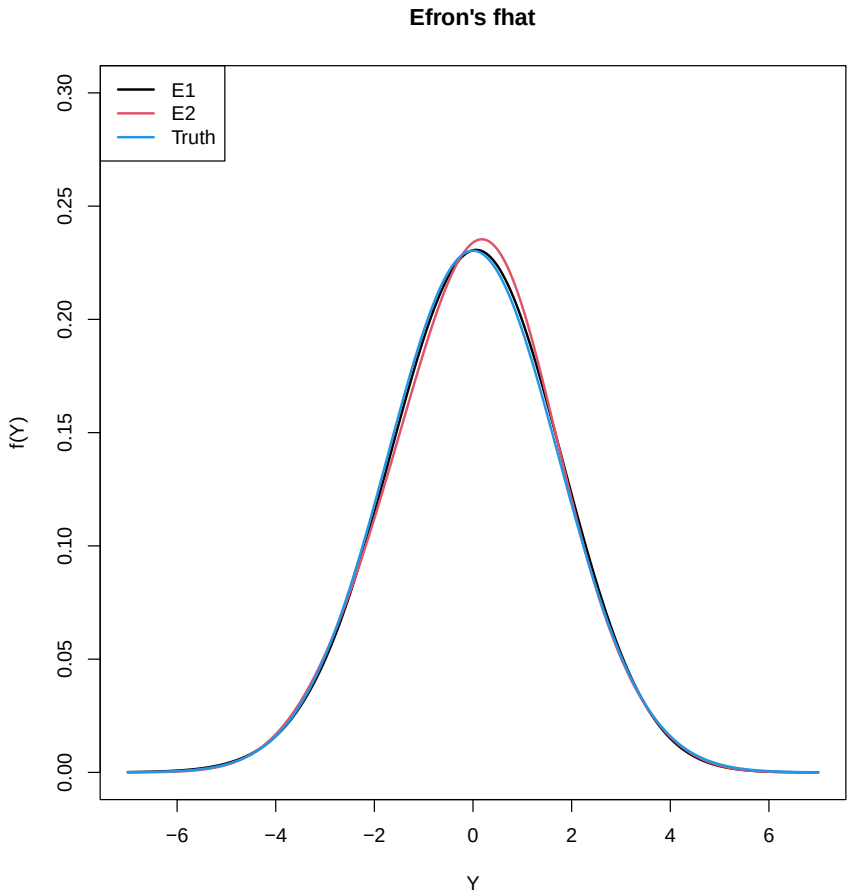




Demonstration: continuous G

- $Y_i = \alpha_i + u_i, u_i \sim \mathcal{N}(0,1), \alpha_i \sim G = \mathcal{N}(0,2)$.
- Efron's estimator $\hat{G}_n$ and the implied $\hat{f}_n$ with $n = 4000$.
- E1 (black): $p = 5, \lambda = 0.1$. E2 (red): $p = 10, \lambda = 0.1$. Truth (blue).





When G is not point identified

- When outcome variables are discrete, it is more likely for G to be not identified.
- Binomial mixture: $Y_i \sim Binomial(k, p_i)$, e.g. Kline and Walters (2021) where Y_i are job recalls.
- We can only learn k moments of G from the frequency of $\boldsymbol{Y}$.

- We have a set of distributions G that are observationally equivalent even when we know population frequency.
- If the parameter of interest is a function of G, then it is only partially identified and we can construct a bound.

Summary

- Use Empirical Bayes estimators when collective performance matters, not individual performance.
- We have tools to estimate G_n nonparametrically, and it can be beneficial when n is moderately large.

Isn't deconvolution hard?

- Deconvolution for G_n is a hard problem: rate for $\hat{G}_n$ is logarithmic

$$f_{G_1} \approx f_{G_2} \not\Rightarrow G_1 \approx G_2$$

- But: if parameter of interest results from smoothing G_n, performance can be good.

- Marginal density $f_{\hat{G}_n}(y) = \int \phi(y-\alpha) d\hat{G}_n(\alpha)$.
  - Hellinger risk bound $\log^2 n/n$ (Zhang (2009), Polyanskiy and Wu (2020)).
- Linear functionals $\int g(\alpha) d\hat{G}_n(\alpha)$ are discussed in van der Geer (2000, Chapter 11).

Beyond normal mean problem

Learning G opens door to other inquiries about α_i's:

- Heterogeneity: $Var(\alpha) = \int \alpha^2 dG - (\int \alpha dG)^2$.
- Tail probability:

$$\mathbb{P}(\alpha \geq u|Y) = \int 1\{\alpha \geq u\} p(y \mid \alpha) dG / \int p(y \mid \alpha) dG$$

- Tail mean:

$$\mathbb{E}[\alpha \mid Y \geq u] = \int \int \alpha 1\{y \geq u\} p(y \mid \alpha) dy dG / \int \int 1\{y \geq u\} p(y \mid \alpha) dy dG$$

Learning G can be useful in other compound decision problems.

Other compound decision problems: testing, ranking/selection

Another Simple Compound Decision Problem (Robbins 1951)

$$Y_i = \alpha_i + u_i, \alpha_i \in \{-1, 1\}, u_i \sim \mathcal{N}(0, 1)$$

- Loss function: $L(\alpha_i, \delta_i) = \frac{1}{2}|\alpha_i - \delta_i|$ for $\delta_i \in \{-1, 1\}$: incur loss one if δ_i makes a wrong guess, otherwise zero.
- G_n of $\boldsymbol{\alpha}$ is characterized by one number: $p_n = n^{-1} \sum_i 1\{\alpha_i = 1\}$.
- Compound risk for $\delta_i(\boldsymbol{Y}) = t(Y_i)$ for some function t:

$$\begin{aligned} R_n(\boldsymbol{\alpha}, \boldsymbol{\delta}) &= (2n)^{-1}\mathbb{E}\Big(\sum_i |\alpha_i - t(Y_i)|\Big) \\ &= \frac{1}{2}\int\int |\alpha - t(y)|\phi(y - \alpha)dydG_n \\ &= p_n \int |1 - t(y)|\phi(y - 1)dy + (1 - p_n)\int |-1 - t(y)|\phi(y + 1)dy \end{aligned}$$

- Fundamental theorem of compound decision applies: compound risk = Bayes risk with prior p_n.

Bayes rule with prior p_n

- For any $p_n \in (0,1)$, the optimal Bayes estimator (with prior p_n) takes the form

$$t_{p_n}(Y) = sgn(Y + \frac{1}{2}\log\frac{p_n}{1-p_n})$$

- If $p_n > 1/2$, then $\frac{1}{2}\log\frac{p}{1-p} > 0$, "shrunk Y" upwards; otherwise downwards.

Robbins's rule

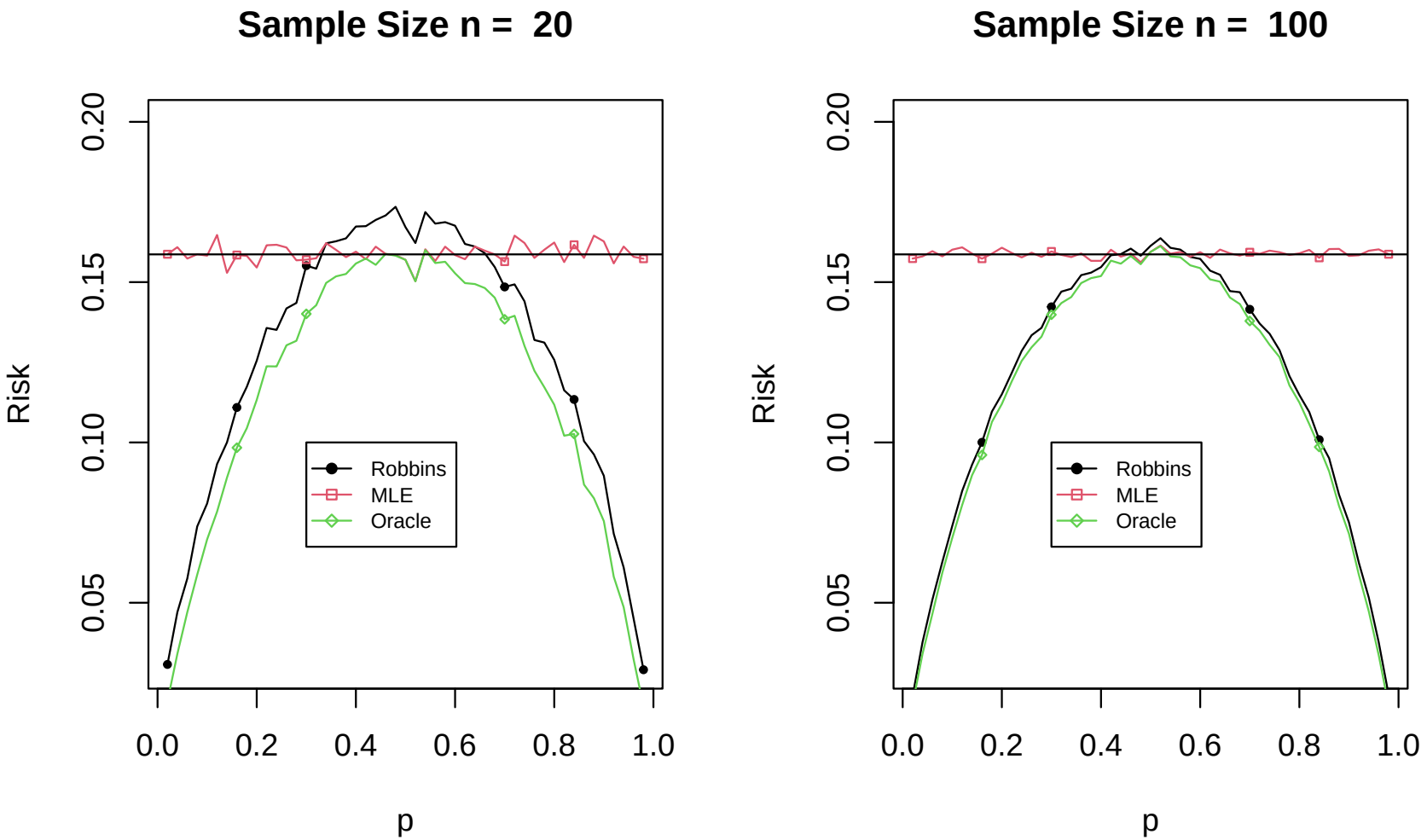
- Not knowing p_n, Robbins suggested estimating it using the whole data $Y_1, \ldots, Y_n$.
- A method of moments estimator: $\hat{p}_n = (n^{-1} \sum_i Y_i + 1)/2$.
- Hannan and Robbins (1955): As long as $\hat{p}_n$ is consistent, as $n \to \infty$,

$$R_n(\boldsymbol{\alpha}, t_{\hat{p}_n}(\boldsymbol{Y})) = \min R_n(\boldsymbol{\alpha}, t(\boldsymbol{Y})) + o_p(1)$$

- When $p_n = 1/2$, $t_{1/2}(Y) = sgn(Y)$ corresponds to the MLE of α_i, using $t_{\hat{p}_n}(Y)$ we are at most $\epsilon = o_p(1)$ worse off.
- When $p_n \neq 1/2$, we can do much better using $t_{\hat{p}_n}(Y)$.

Risk Comparison

Figure: From Gu and Koenker (2016). Mean absolute loss over 1000 replications. Black (–) : Robbins; Green (–): Optimal Bayes; Red (–) MLE



Connection to Multiple Testing

Consider again the normal mean problem

$$Y_i = \alpha_i + u_i, u_i \sim \mathcal{N}(0,1), \alpha_i \sim G$$

- We are now interested in which $\alpha_i \notin \mathcal{A}$.
- Example: with $\mathcal{A} = \{0\}$ we are testing for significance for each individual α_i.
- Transformed parameters: $H_i = 1\{\alpha_i \notin \mathcal{A}\}$ (i.e. not α_i).
- Compound decision with loss

$$L(\delta_i, H_i) = \begin{cases} 1-\tau & \text{if } \delta_i = 1, H_i = 0 \quad \text{(type I error)} \\ \tau & \text{if } \delta_i = 0, H_i = 1 \quad \text{(type II error)} \end{cases}$$

- Distribution of H_i: $p_0 = \mathbb{P}(H_i = 1) = \mathbb{P}(\alpha_i \notin \mathcal{A}) = G(\mathcal{A}^c)$.

Connection to Multiple Testing

Optimal Bayes estimator:

$$\delta_i^{Bayes} = 1\{\mathbb{P}(\alpha_i \in \mathcal{A}|Y_i) \leq \tau\}$$

with

$$\mathbb{P}(\alpha_i \in \mathcal{A}|Y_i) = \int_{\mathcal{A}} \phi(Y_i - \alpha) dG(\alpha) / \int_{\mathbb{R}} \phi(Y_i - \alpha) dG(\alpha)$$

- The quantity $\mathbb{P}(\alpha_i \in \mathcal{A}|Y_i)$ is also known as local false discovery rate (Lfdr).
- Once we estimate G, we can estimate Lfdr.
- Empirical Bayes Multiple Testing: rank individuals by Lfdr and then threshold.
- Choice of τ leads to different false discovery rate (proportion of false rejections out of all rejections).

Connection to ranking and selection

Gu and Koenker (2022) considers EB methods for ranking and selection problem.

- Among all teachers, we'd like to estimate the top q % performers.
- Parameter of interest: $H_i = 1\{\alpha_i \geq G^{-1}(1-q)\}$.
- Rank teachers with Lfdr $\mathbb{P}(\alpha_i < G^{-1}(1-q)|Y_i)$ and then threshold by τ.

Ranking devices:

- **Homogenous variances of** u_i: ranking with $Y_i \Leftrightarrow$ ranking with $g(Y_i)$ with monotone g
  - i.e. parametric/nonparametric EB, local false discovery rate all give same ranking.
- **Heterogenous variances of** u_i: EB methods can produce different rankings.

Role of τ:

- τ controls false discovery rate.
- τ controls proportion of selection to make.

Design loss function to achieve two goals:

- Capacity constraint: we want to select at most q %.
- FDR constraint: avoid making too many false discoveries among selection.

Empirical Bayes Inference

Empirical Bayes Inference: normal mean with normal prior

- Assume $\alpha_i \sim_{iid} G = \mathcal{N}(0, A)$, then Bayes estimator $\delta_i = (1 - \frac{1}{1+A})Y_i$.
- Bayesian credible set for α_i with 95% coverage if we knew A:

$$[q_{0.025}(y_i, A), q_{0.975}(y_i, A)] = \delta_i \pm 1.96\sqrt{A/(1+A)}$$

with $q_\tau(y_i, A)$ being the τ quantile of posterior distribution

$$\alpha_i | y_i, A \sim \mathcal{N}\Big(\frac{A}{1+A}y_i, \frac{A}{1+A}\Big)$$

- Naive EB confidence interval $[q_{\tau/2}(y_i, \hat{A}), q_{1-\tau/2}(y_i, \hat{A})]$ does not account for the fact that A is estimated, and this gives poor coverage when n is small.
- Morris (1983) provides a finite sample correction: add a correction term to $\sqrt{\hat{A}/(1+\hat{A})}$.

More Robust Empirical Bayes Inference

Armstrong et al. (2022) shows if $G \neq \mathcal{N}$, coverage of the naive EBCI can be poor.

- They provide a robust EBCI:

$$CI_{i,\tau} = \delta_i \pm \hat{c}_\tau \sqrt{\hat{A}/(1+\hat{A})}$$

such that $\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(\alpha_i \in CI_{i,\tau}) \geq 1-\tau$.

EBCI as a compound decision (Jiang (2021)) :

$$\min_{a,b} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[b(Y_i) - a(Y_i)]$$

$$\text{subject to } \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(a(Y_i) \leq \alpha_i \leq b(Y_i)) = 1-\tau$$

Conclusion

- Empirical Bayes methods can be useful for economics application as we become more interested in unobserved heterogeneity.
- Herbert Robbins is ahead of his time in proposing the nonparametric EB method in 1950s.
- Large scale individualized datasets become increasingly available.
- Modern computation methods make it feasible to apply his method.

Selected References I

Armstrong, T. B., Kolesár, M. & Plagborg-Møller, M. (2020), 'Robust empirical bayes confidence intervals', *arXiv preprint arXiv:2004.03448* .

Brown, L. (2008), 'In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies', *The Annals of Applied Statistics* **2**, 113–152.

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

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

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge U. Press: Cambridge.

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

Efron, B. (2014), 'Two modeling strategies for empirical bayes estimation', *Statistical science* **29**(2), 285.

Efron, B. & Morris, C. (1973), 'Stein's estimation rule and its competitiors - an empirical bayes approach', *Journal of the American Statistical Association* **68**, 117–130.

Efron, B. & Morris, C. (1975), 'Data analysis using stein's estimator and its generalizations', *Journal of the American Statistical Association* **70**(350), 311–319.

Gilraine, M., Gu, J. & McMillan, R. (2020), A new method for estimating teacher value-added. NBER Working Paper Series Number 27094.

Gu, J. & Koenker, R. (2016), 'On a problem of Robbins', *International Statistical Review* **84**, 224–244.

Gu, J. & Koenker, R. (2017*a*), 'Empirical bayesball remixed: Empirical bayes methods for longitudinal data', *Journal of Applied Econometrics* **32**(3), 575–599.

Selected References II

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

Gu, J. & Koenker, R. (2020), 'Invidious comparisons: Ranking and selection as compound decisions', *arXiv preprint arXiv:2012.12550* .

Hannan, J. F. & Robbins, H. (1955), 'Asymptotic solutions of the compound decision problem for two completely specified distributions', *The Annals of Mathematical Statistics* pp. 37–51.

Heckman, J. & Singer, B. (1984), 'A method for minimizing the impact of distributional assumptions in econometric models for duration data', *Econometrica* **52**, 63–132.

Ignatiadis, N. & Wager, S. (2022), 'Confidence intervals for nonparametric empirical bayes analysis', *Journal of the American Statistical Association* pp. 1–18.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proceedings of the fourth Berkeley symposium on mathematical statistics and probability', Univ of California Press, p. 361.

Jiang, W. (2019), 'Comment: Empirical bayes interval estimation', *Statistical science* **34**(2), 219–223.

Jiang, W. & Zhang, C.-H. (2009), 'General Maximum Likelihood Empirical Bayes Estimation of Normal Means', *The Annals of Statistics* **37**, 1647–1684.

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

Koenker, R. & Gu, J. (2015), Rebayes: An r package for empirical Bayes methods. Available from `https://cran.r-project.org/package=REBayes`.

Koenker, R. & Gu, J. (2017), 'Rebayes: an r package for empirical bayes mixture methods', *Journal of Statistical Software* **82**, 1–26.

Koenker, R. & Gu, J. (2019), 'Comment: Minimalist g-modeling', *Statistical science* **34**(2), 209–213.

Selected References III

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

Laird, N. (1978), 'Nonparametric maximum likelihood estimation of a mixing distribution', *Journal of the American Statistical Association* **73**, 805–811.

Lindsay, B. (1983), 'The geometry of mixture likelihoods: A general theory', *The Annals of Statistics* **11**, 86–94.

Lindsay, B. (1995), Mixture models: Theory, geometry and applications, *in* 'NSF-CBMS regional conference series in probability and statistics'.

Matias, C. & Taupin, M.-L. (2004), 'Minimax estimation of linear functionals in the convolution model', *Mathematical Methods of Statistics* **13**(3), 282–328.

Morris, C. (1983), 'Parametric empirical Bayes inference: Theory and applications', *Journal of the American Statistical Association* **78**, 47–55.

Neyman, J. & Scott, E. L. (1948), 'Consistent estimates based on partially consistent observations', *Econometrica: Journal of the Econometric Society* pp. 1–32.

Robbins, H. (1951), Asymptotically subminimax solutions of compound statistical decision problems, *in* 'Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press: Berkeley, pp. 131–149.

Robbins, H. (1956), An empirical Bayes approach to statistics, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press: Berkeley, pp. 157–163.

Stein, C. (1956), Inadmissibility of the usual estimator of the mean of a multivariate normal distribution, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press: Berkeley, pp. 197–206.

Zhang, C.-H. (2009), 'Generalized maximum likelihood estimation of normal mixture densities', *Statistica Sinica* pp. 1297–1318.

Optional slides: Compound decision with Poisson model

Motivating Example 2: Evaluation of dialysis centers

- We have information on 3230 dialysis centers in the U.S. over a few time periods.
- Outcome Y_{it}: number of deaths.
- Covariates X_{it}: measures of observed patient heterogeneity.
- Poisson regression model:

$$\mathbb{E}[Y_{it}|X_{it}] = \exp(log(\alpha_i) + X_{it}'\beta)$$

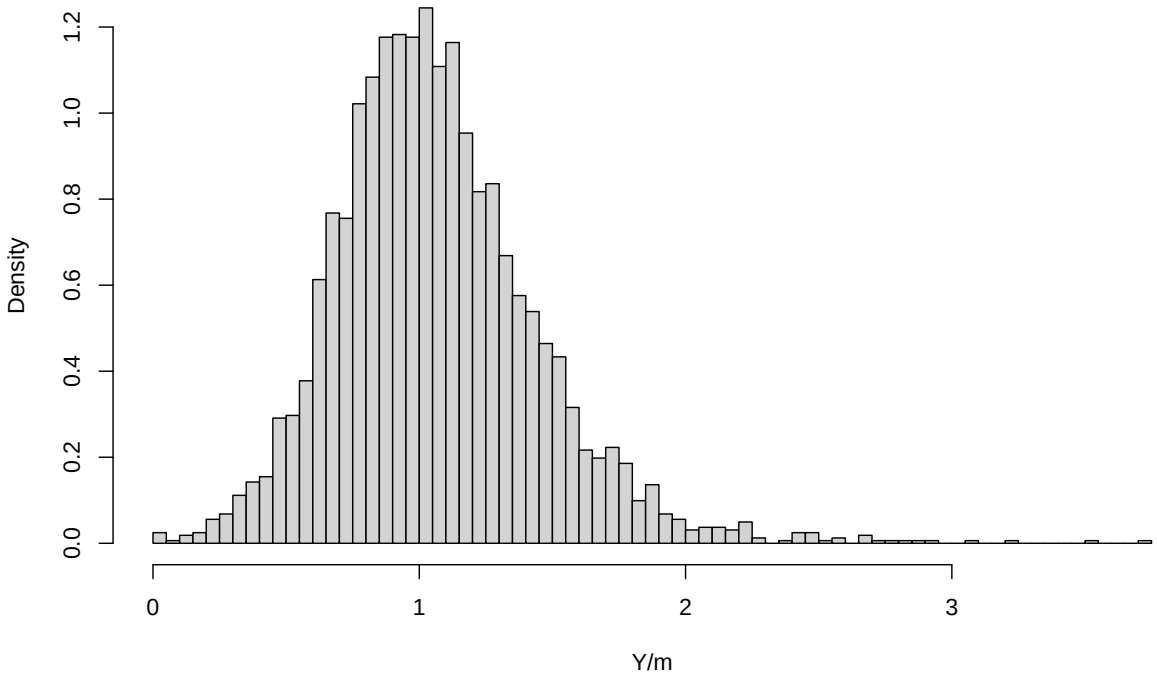
- α_i measures unobserved center quality.
- $Y_{it} \sim Poisson(\alpha_i m_{it})$ with $m_{it} = \exp(X_{it}'\hat{\beta})$.
- Policy maker may be interested in:
  - Diversity of quality.
  - An estimate of each center's quality α_i.
  - Shut down/expand a subset of centers having low/high quality.

Beyond normal models: Poisson for counts

- For simplicity, using 1 year of data:

$$Y_i \sim Poisson(\alpha_i m_i), \alpha_i \sim G$$

  where Y_i is number of deaths and m_i is called exposure, measuring expected number of deaths at each center.
- We are interested in α_i: imagine all centers want to purchase insurance and insurer tries to decide individual specific premium depending on α_i.



Parametric empirical Bayes

- Consider prior $\alpha_i \sim \mathit{Gamma}(\alpha, \beta)$.
- Then Bayes estimator

$$\delta_i^{Bayes} = \frac{Y_i + \alpha}{m_i + \beta}$$

- Parametric empirical Bayes: $\delta_i^{PEB} = \frac{Y_i + \hat{\alpha}}{m_i + \hat{\beta}}$ with $\hat{\alpha}, \hat{\beta}$ be MLE estimates.

Poisson: nonparametric empirical Bayes

- The "Tweedie formula" for Poission for ($\alpha_i \sim G$) leads to

$$t_G(y, m) = \mathbb{E}[\alpha | Y, m] = \frac{\int \alpha P(Y = y \mid \alpha, m) dG(\alpha)}{\int P(Y = y \mid \alpha, m) dG(\alpha)}$$

- If $m_i = m$ were all the same (absolute into $\alpha_i = \alpha_i m$), then

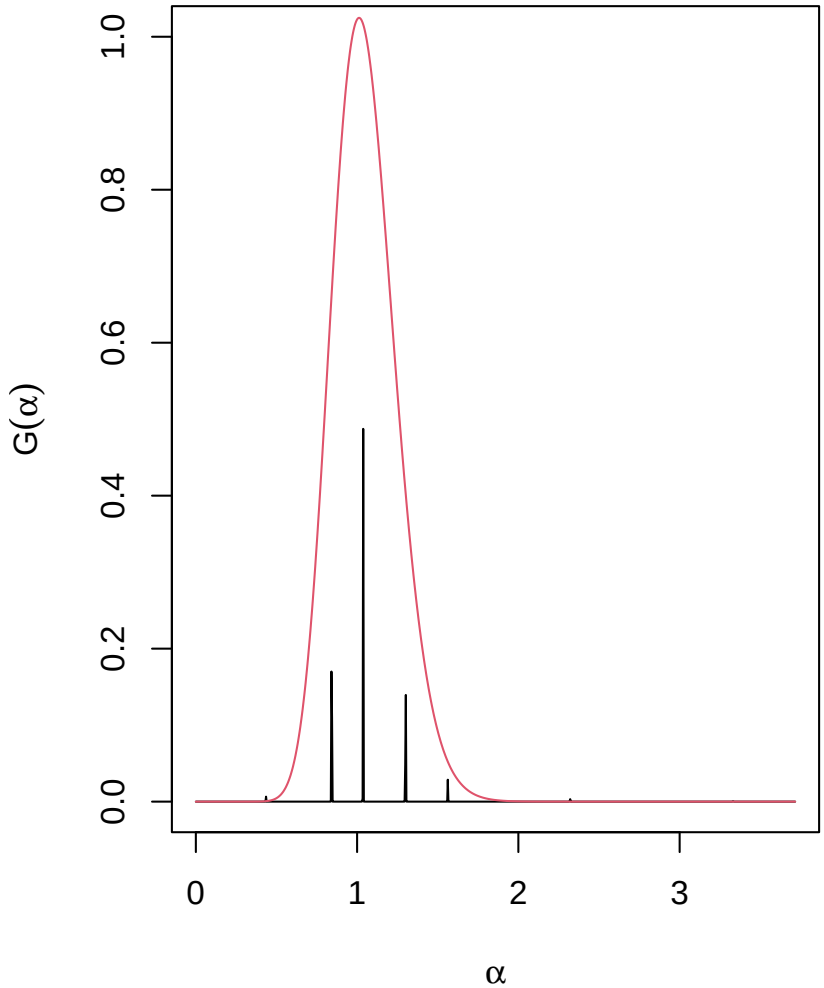
$$t_G(y) = \frac{(y+1) P_Y(y+1)}{P_Y(y)}$$

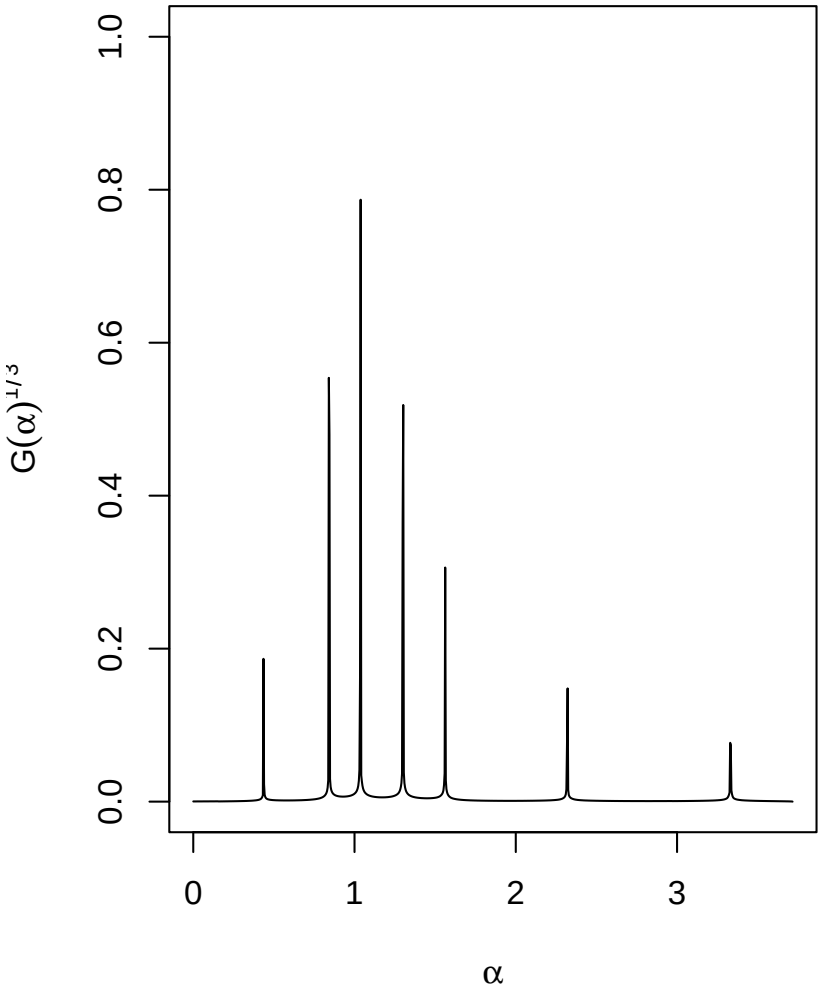
  with $P_Y(y) = P(Y_i = y)$.
- f-modeling: directly estimate P_Y; g-modeling: estimate G and back out P_Y.
- Robbins (1956) proposed to use f-modeling, but empirical frequency of $\boldsymbol{Y}$ doesn't necessarily respect monotonicity of $t_G(y)$.
- Further improvement provided by Brown et al. (2013).

- g-modeling: automatically respects monotonicity. Also handles heterogeneous m_i very well.

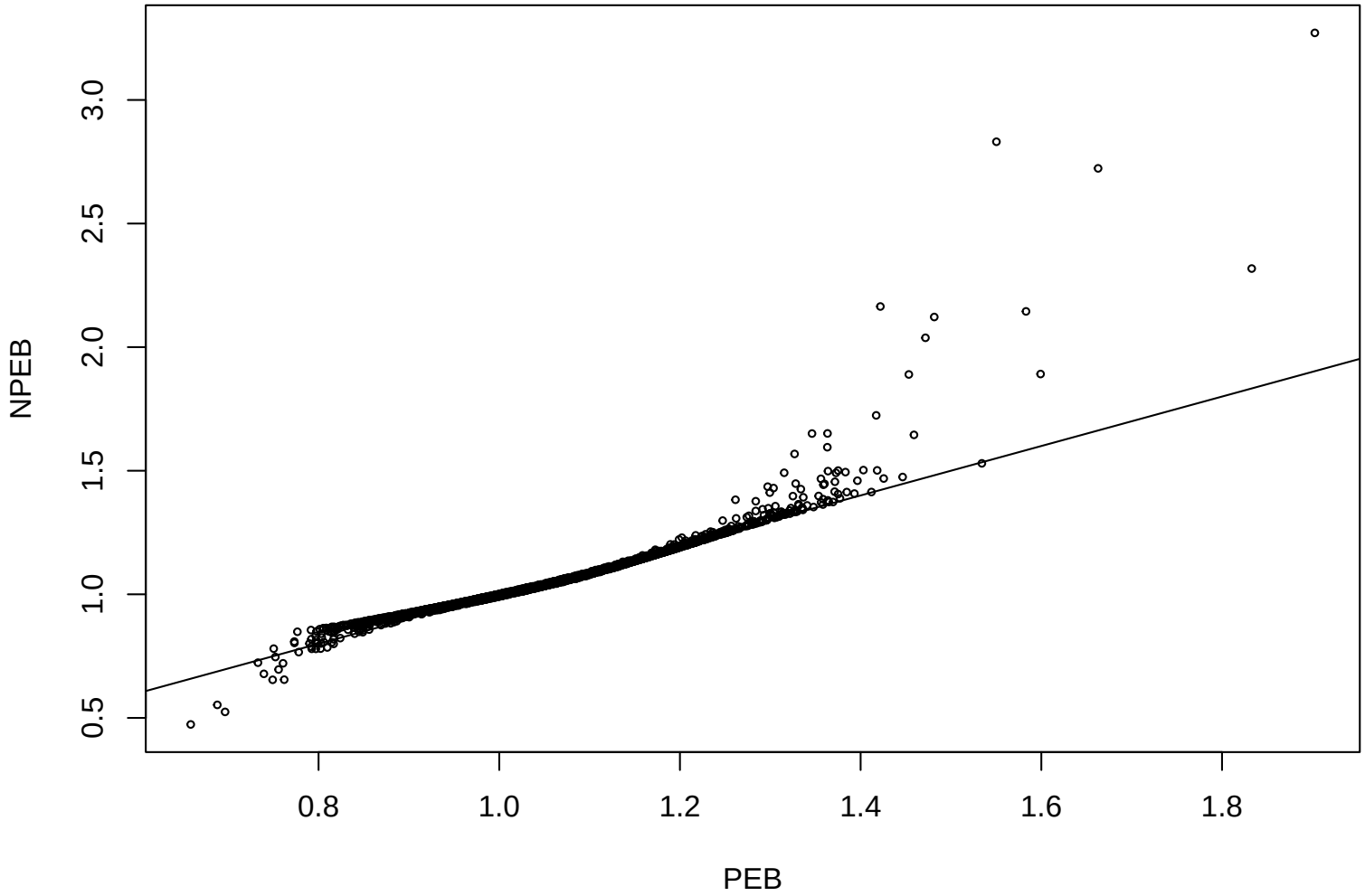
Parametric vs nonparametric MLE G

Left: Gamma (in red) vs NPMLE (in black). Right: Plots in the scale of $G^{1/3}$ to review mass points.





Parametric vs Nonparametric



Poisson for counts

- If we are interested in $\mathbb{E}[\alpha|Y < t, m]$, then we can show

$$\mathbb{E}[\alpha|Y < t, m] = \frac{\sum_{y=0}^{t-1} P_Y(y; m) t_G(y, m)}{\sum_{y=0}^{t-1} P_Y(y; m)} = \mathbb{E}[t_G(Y, m)|Y < t, m]$$

  with $P_y(y; m) = \int P(Y = y \mid \alpha, m) dG(\alpha)$.
- In contrast, selection bias using Y:

$$\mathbb{E}[\alpha|Y < t, m] - \mathbb{E}[Y|Y < t, m] = \frac{t P_Y(t; m)}{\sum_{y=0}^{t-1} P_Y(y; m)} > 0$$

- Posterior mean can be shown to cure selection bias (Dawid (1994)), only when G is reasonable. Poor prior can produce poor posterior.