On a Problem of Robbins:
Or How I Learned to Stop Worrying and Love (Empirical) Bayes

Roger Koenker

University of Illinois, Urbana-Champaign

Hong Kong: 23 May 2014

On a Problem of Robbins:
Or How I Learned to Stop Worrying and Love (Empirical) Bayes

Roger Koenker

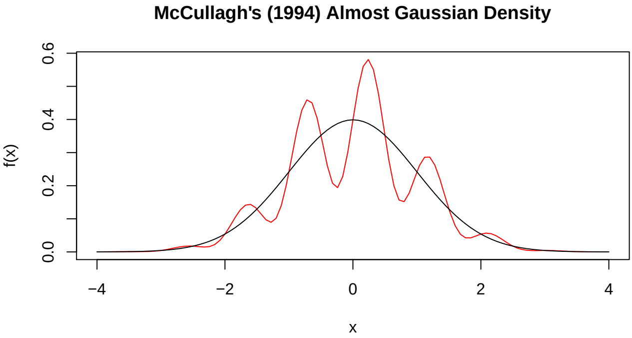
University of Illinois, Urbana-Champaign

Hong Kong: 23 May 2014

Outline

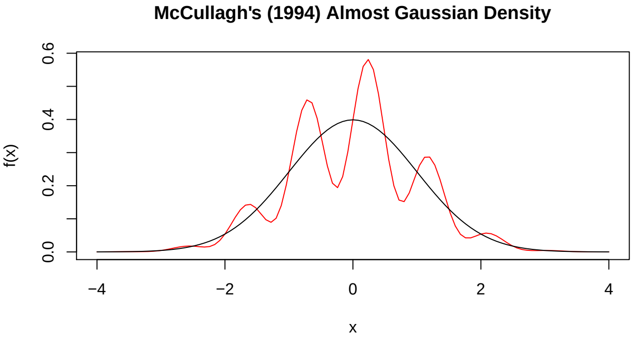
- Prologue or Provocation?
  - Partial Identification and Gaussian Moment Matching
  - Moment Equalities and Inequalities
  - Discrete Distributions and their Aliases
- Robbins's (1951) Compound Decision Problem
  - Minimax Rules and their Discontents
  - Mixture Models and the Kiefer-Wolfowitz GMLE
  - Applications to Classification and Multiple Testing

Where are we when we are “in the moment?”



$$f(x) = \varphi(x)(1 + \frac{1}{2}\sin(2\pi x))$$

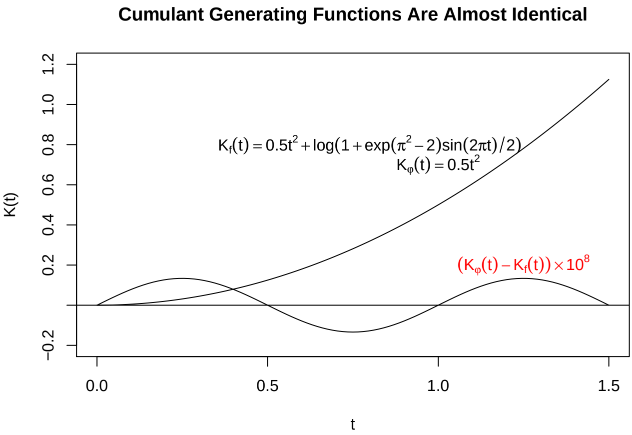
Where are we when we are “in the moment?”



$$f(x) = \varphi(x)(1 + \tfrac{1}{2}\sin(2\pi x))$$

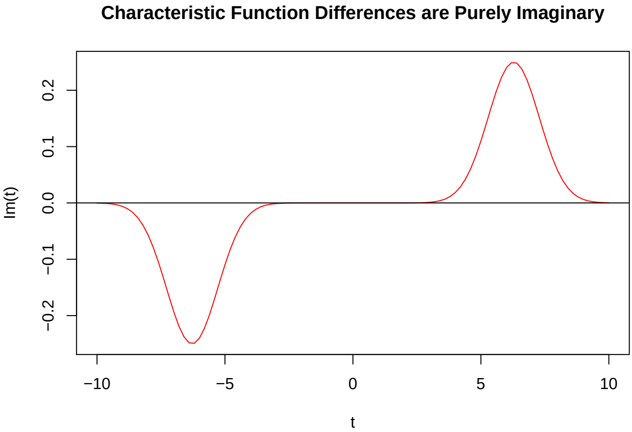
Densities f and φ have identical even moments, odd moments up to 9 are nearly zero.

Cumulants Too



$$|K_f(t) - K_\varphi(t)| < \epsilon = 10^{-8}$$

But the Characteristic Function Reveals All



Real parts are identical, only the imaginary part is informative.

Momentary Bounds for Distribution Functions

The McCullagh example raises the question: If F and G have the same first $2p$ moments how big can $|F(x) - G(x)|$ be? Lindsay and Basak (2000), building on prior work of Akhiezer, offer the answer for continuous G,

$$\frac{1}{2} w_p(x) \leqslant \sup_{F \in \mathcal{F}_p} |F(x) - G(x)| \leqslant w_p(x),$$

where $w_p(x) = (v_p(x)^\top H_p^{-1} v_p(x))^{-1}$, $v_p(x) = (1, x, x^2, \cdots, x^p)$ and H_p is the Hankel matrix,

$$H_p = \begin{bmatrix} 1 & m_1 & \cdots & m_p \\ m_1 & m_2 & \cdots & m_{p+1} \\ \vdots & & & \vdots \\ m_p & m_{p+1} & \cdots & m_{2p} \end{bmatrix}$$

with $m_k = \int x^k dG(x)$, but Lindsay comments that finding such F's is "numerically challenging."

How Challenging Is It? Two Approaches

- 20th Century Brute Force (Method of Moment Spaces)

$$\min\{c^\top w \mid Aw = m,\ w \in \mathcal{S}\}$$

where $A = (x_i^j)$, $i = 1, \cdots, n,\ j = 1, \cdots, 2p$ and $\{x_i\}$ constitute a fairly fine equally spaced grid on, say $[-8, 8]$.

How Challenging Is It? Two Approaches

- 20th Century Brute Force (Method of Moment Spaces)

$$\min\{c^\top w \mid Aw = m,\ w \in \mathcal{S}\}$$

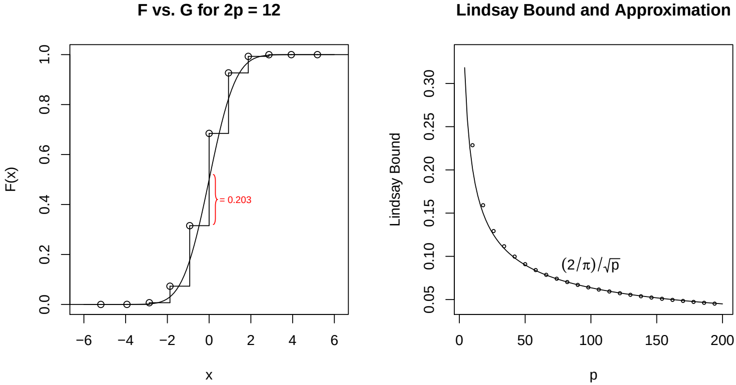
where $A = (x_i^j), i = 1, \cdots, n,\ j = 1, \cdots, 2p$ and $\{x_i\}$ constitute a fairly fine equally spaced grid on, say $[-8, 8]$.

- 19th Century Finesse (Gaussian Quadrature)

$$F(x) = \sum_{i=1} w_i \delta_{x_i}(x)$$

where x_i are the roots of a Hermite polynomial of order, $2p + 1$, and the w_i are given by the standard formulae for Gaussian quadrature. If not “known to Gauss” probably “obvious to Jacobi.”

The Akhiezer-Lindsay Bound is Sharp



Theorem: The Akhiezer-Lindsay bound is attained by the discrete "Gaussian quadrature" density.

The Moral Take-away

- Downside
  - Moments are informative about the tails of distributions, but not much else,
  - Higher moments relevant for large deviation results,
  - For distributions with unbounded support, moments aren't estimable, i.e. are not identified, Bahadur and Savage (1956).

The Moral Take-away

- Downside
  - Moments are informative about the tails of distributions, but not much else,
  - Higher moments relevant for large deviation results,
  - For distributions with unbounded support, moments aren't estimable, i.e. are not identified, Bahadur and Savage (1956).
- Upside
  - Discrete distributions effectively encode seemingly more complex continuous distributions, cf. Sims's rational inattention.

The Robbins (1951) Compound Decision Problem

Suppose we observe, $y = (y_1, \cdots, y_n)$ from,

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

and we would like to estimate $\theta \in \{-1, 1\}^n$ under loss,

$$L(\hat{\theta}_i, \theta_i) = n^{-1} \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|.$$

Robbins notes that for $n = 1$ the minimax procedure is,

$$\delta_{1/2}(y) = \text{sgn}(y),$$

and he shows that this rule remains minimax for $n > 1$.

Let's be Bayesian

Lacking further information we may be willing to assume that the Y_i are exchangeable, and thus that the θ_i are iid Bernoulli (p). The minimax principle presumes that malevolent nature has chosen $p = 1/2$.

Let's be Bayesian

Lacking further information we may be willing to assume that the Y_i are exchangeable, and thus that the θ_i are iid Bernoulli (p). The minimax principle presumes that malevolent nature has chosen $p = 1/2$. Robbins observes that if we knew p,

$$P(\theta = 1|y, p) = \frac{p\varphi(y-1)}{p\varphi(y-1) + (1-p)\varphi(y+1)}$$

we should guess $\hat{\theta}_i = 1$ if this probability exceeds 1/2, or equivalently,

$$\delta_p(y) = \text{sgn}(y - \tfrac{1}{2}\log((1-p)/p))$$

Let's be Bayesian

Lacking further information we may be willing to assume that the Y_i are exchangeable, and thus that the θ_i are iid Bernoulli (p). The minimax principle presumes that malevolent nature has chosen $p = 1/2$. Robbins observes that if we knew p,

$$P(\theta = 1 | y, p) = \frac{p\varphi(y-1)}{p\varphi(y-1) + (1-p)\varphi(y+1)}$$

we should guess $\hat{\theta}_i = 1$ if this probability exceeds 1/2, or equivalently,

$$\delta_p(y) = \text{sgn}(y - \tfrac{1}{2}\log((1-p)/p))$$

But we don't know p.

Hierarchical Bayes Methods

We have the log likelihood,

$$\ell_n(p|y) = \sum_{i=1}^{n} \log(p\varphi(y_i - 1) + (1-p)\varphi(y_i + 1))$$

a symmetric beta prior is convenient,

$$\log \pi(p) = a \log(p) + a \log(1-p) - \log B(a, a).$$

Hierarchical Bayes Methods

We have the log likelihood,

$$\ell_n(p|y) = \sum_{i=1}^{n} \log(p\varphi(y_i - 1) + (1-p)\varphi(y_i + 1))$$

a symmetric beta prior is convenient,

$$\log \pi(p) = a \log(p) + a \log(1-p) - \log B(a, a).$$

The posterior for θ_i is,

$$p(\theta_i = 1 \mid y_1, \ldots, y_n) = \frac{\varphi(y_i - 1)\bar{p}_i}{\varphi(y_i - 1)\bar{p}_i + \varphi(y_i + 1)(1 - \bar{p}_i)},$$

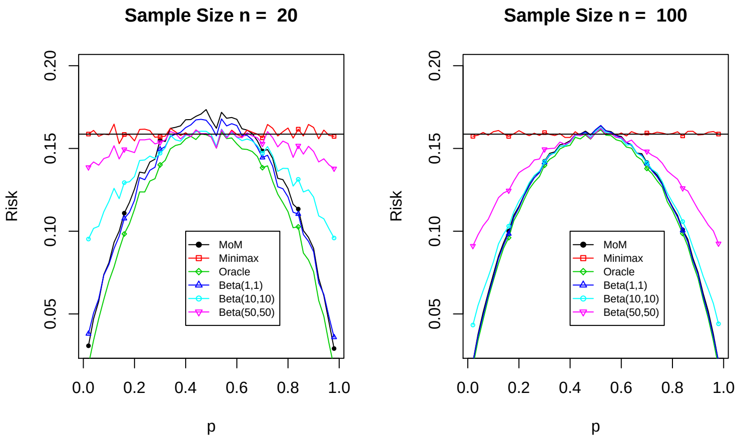
where $\bar{p}$ is the posterior mean of p given the data y.

$$\bar{p}_i = \frac{\int p \prod_{j \neq i}(p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1))\pi(p)dp}{\int \prod_{j \neq i}(p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1))\pi(p)dp}.$$

and we have a plug-in cutoff Bayes rule,

$$\delta_{\bar{p}_i}(y_i) = \operatorname{sgn}(y_i - \tfrac{1}{2}\log((1 - \bar{p}_i)/\bar{p}_i)).$$

Empirical Risk for Several Decision Rules



Mean absolute loss over 1000 replications.

A Grouped Robbins Problem

Suppose we now have a panel structure, n groups each with J members

$$Y_{ij} = \theta_{ij} + u_{ij}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, J,$$

with $\theta_{ij} \in \{-1, 1\}$ and $u_{ij} \sim \mathcal{N}(0, 1)$. Each group is allowed its own p_i, but – preserving exchangeability – drawn from a distribution G, so marginally,

$$Y_i \sim f(y|p) = \int_0^1 \prod_{j=1}^{J} (p\varphi(y_j - 1) + (1-p)\varphi(y_j + 1))dG(p).$$

Robbins (1951), anticipating Kiefer and Wolfowitz (1956), proposed that G could be estimated (nonparametrically) by maximum likelihood.

Generalized MLE's for Mixture Models

When the number of groups, n, is small we can proceed as before with group specific MLE's. But for larger n it is preferable to "borrow strength" across groups and estimate the mixing distribution, G, from all the data. There are two options:

- Parametric Random Effects: Assume G takes some parametric form and estimate its "hyperparameters." This is the traditional hierarchical Bayes option.
- Nonparametric Random Effects: Try to estimate G nonparametrically. This is the Robbins (1951) and Kiefer and Wolfowitz (1956) empirical Bayes option.

Kiefer and Wolfowitz Generalized MLE's for Mixture Models

- Generic Problem

$$Y_i|\theta \sim f(y|\theta), \quad \theta \sim G, \quad Y_i \sim h(y) = \int f(y|\theta)dG(\theta)$$

$$\max_{G \in \mathcal{G}} \{\sum_{i=1}^{n} \log h(y_i) \mid h(y) = \int f(y|\theta)dG(\theta)\}$$

- Generic Solutions
  - Objective is strictly convex and constraints are polyhedral, so solutions are unique.
  - Constraints are implemented on a fairly fine grid, so solutions are discrete with only a few mass points.
  - Rather than impose a prior for G, we estimate it, *quelle horreur*.

The Grouped Robbins Problem

In the grouped Robbins problem with a mixture over the p_i's we solve,

$$\max\{\sum_{i=1}^{n} \log(h_i) \mid Ap = h, p \in \mathcal{S}\}$$

where $h_i = h(y_{i1}, \cdots, h_{iJ})$, A denotes the n by m matrix with typical element

$$A_{ik} = \prod_{j=1}^{J} (p_k \varphi(y_{ij} - 1) + (1 - p_k)\varphi(y_{ij} + 1))$$

and p is an m-vector, constituting a grid on $[0, 1]$, and living on the $m - 1$ dimensional simplex, $\mathcal{S}$.

Some Simulation Evidence

As a simple example suppose that we have $n = 200$ groups with $J \in \{5, 10, 100\}$ observations per group, and the group p_i are iid with $\mathcal{P}(\theta_{ij} = 1) \equiv p_i \sim \frac{1}{4}\delta_{0.1} + \frac{3}{4}\delta_{0.3}$. We compare risk performance for estimating the θ_{ij} relative to an oracle rule for:

- (Wald) minimax rule,
- Robbins method of moments rule applied separately to each group,
- Empirical characteristic function, ECF, rule of Jin and Cai (2007),
- GMLE empirical Bayes rule based on Robbins, Kiefer and Wolfowitz.

n	J	Minimax	MoM	ECF	GMLE
200	5	1.668	1.599	1.472	1.357
200	10	1.300	1.290	1.224	1.043
200	100	1.305	1.036	1.048	1.011

Free the θ's: The Gaussian Sequence Model

Restricting the θ_{ij}'s to live in $\{-1, 1\}$ seems a bit cruel, why not let them roam free? Suppose that,

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

so marginally $Y_i \sim f(y) = \int \varphi(y - \theta) dG(\theta)$. Under squared error loss Robbins (1956) shows that the optimal Bayes rule estimator of the θ's is given by,

$$\delta(y) = y + f'(y)/f(y).$$

Efron (2011) calls this Tweedie's formula; it provides a general shrinkage strategy for Gaussian noise models, encompassing various parametric Stein rule procedures. When G is known we're good to go, otherwise we need to estimate our prior, G.

Needless [sic] and Haystacks

It is commonly assumed that G contains a large mass point concentrated at zero, the haystack, and a smaller mass well separated from zero, i.e. the needles. Castillo and van der Vaart (2012) compare several Bayes and empirical Bayes procedures in this setting.

	s = 25			**s = 50**			**s = 100**		
	3	4	5	3	4	5	3	4	5
PM1	111	96	94	176	165	154	267	302	307
PM2	106	92	82	169	165	152	269	280	274
EBM	103	96	93	166	177	174	271	312	319
PMed1	129	83	73	205	149	130	255	279	283
PMed2	125	86	68	187	148	129	273	254	245
EBMed	110	81	72	162	148	142	255	294	300
HT	175	142	70	339	284	135	676	564	252
HTO	136	92	84	206	159	139	306	261	245
GMLE	80	57	30	122	81	40	174	112	53

Mean squared error of several estimators considered by Castillo and van der Vaart and the GMLE procedure of Robbins. Sample size $n = 500$ throughout, with s non-null observations concentrated at $\theta \in \{3, 4, 5\}$. Based on 100 replications for the first eight Castillo and van der Vaart procedures, and 1000 replications for the GMLE.

Multiple Testing

Suppose instead of estimating the θ_i's we only are required to classify them:

H_0: $\theta_i \in A$ so Y_i is regarded as uninteresting

H_1: $\theta_i \notin A$ so Y_i is regarded as interesting

Given $Y_1, \cdots, Y_n$ we need a decision rule, $\delta(Y_i) = 1$ if we think Y_i is interesting and $\delta(Y_i) = 0$ otherwise, subject to asymmetric loss,

$$L(\delta, H) = \begin{cases} 1 - \tau & \text{if } \delta = 1, \text{ and } H = 0, \text{ Type I error,} \\ 0 & \text{otherwise,} \\ \tau & \text{if } \delta = 0, \text{ and } H = 1, \text{ Type II error.} \end{cases}$$

Assume the H_i are Bernoulli(p) so, $Y_i | H_i \sim (1 - H_i)F_0 + H_i F_1$ where

$$dF_0 = f_0 = (1-p)^{-1} \int_A \varphi(y - \theta) dG(\theta),$$

$$dF_1 = f_1 = p^{-1} \int_{A^c} \varphi(y - \theta) dG(\theta),$$

FDR and the New Deal on Testing

The local false discovery rate, Lfdr, is given by,

$$T_i = (1-p)f_0(Y_i)/f(Y_i)$$

where $f(y) = (1-p)f_0(y) + pf_1(y)$ and it is conventional to reject $H_i = 0$ when $\delta_i = I(T_i < c_\alpha = T_{(k)}) = 1$ where,

$$k = \text{argmin}\{k | k^{-1} \sum_{i=1}^{k} T_{(i)} < \alpha\}$$

This approach has a nice interpretation in terms of Bayes factors, Efron (2010), and as shown by Genovese and Wasserman (2002)

$$\text{Mfdr} = \frac{\mathbb{E}\sum_i (1-H_i)\delta_i}{\mathbb{E}\sum_i \delta_i} = \text{FDR} + O_p(n^{-1/2})$$

Can't Find the Oracle?

- Implementation requires estimation of the quantities, p, f_0 and f. and has generally led to deconvolution methods using empirical characteristic functions, e.g. Jin and Cai (2007) and Cai and Sun (2009).

Can't Find the Oracle?

- Implementation requires estimation of the quantities, p, f_0 and f. and has generally led to deconvolution methods using empirical characteristic functions, e.g. Jin and Cai (2007) and Cai and Sun (2009).
- In mixture model settings where deconvolution is appropriate, the Kiefer-Wolfowitz GMLE is an attractive alternative,

Can't Find the Oracle?

- Implementation requires estimation of the quantities, p, f_0 and f. and has generally led to deconvolution methods using empirical characteristic functions, e.g. Jin and Cai (2007) and Cai and Sun (2009).
- In mixture model settings where deconvolution is appropriate, the Kiefer-Wolfowitz GMLE is an attractive alternative,
- GLME mixing distributions are discrete, but this may be a feature, not a bug, in some applications,

Can't Find the Oracle?

- Implementation requires estimation of the quantities, p, f_0 and f. and has generally led to deconvolution methods using empirical characteristic functions, e.g. Jin and Cai (2007) and Cai and Sun (2009).
- In mixture model settings where deconvolution is appropriate, the Kiefer-Wolfowitz GMLE is an attractive alternative,
- GLME mixing distributions are discrete, but this may be a feature, not a bug, in some applications,
- Empirical Bayes methods coupled with GMLE computational techniques provide powerful tools for addressing a wide variety of estimation and testing problems involving unobserved heterogeneity.

Can't Find the Oracle?

- Implementation requires estimation of the quantities, p, f_0 and f. and has generally led to deconvolution methods using empirical characteristic functions, e.g. Jin and Cai (2007) and Cai and Sun (2009).
- In mixture model settings where deconvolution is appropriate, the Kiefer-Wolfowitz GMLE is an attractive alternative,
- GLME mixing distributions are discrete, but this may be a feature, not a bug, in some applications,
- Empirical Bayes methods coupled with GMLE computational techniques provide powerful tools for addressing a wide variety of estimation and testing problems involving unobserved heterogeneity.
- R package **REBayes** available from CRAN.