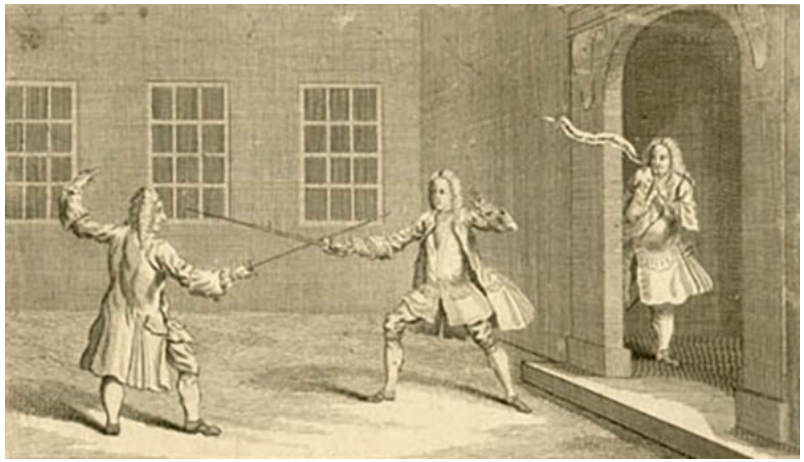
Unlikely Likelihoods

Roger Koenker

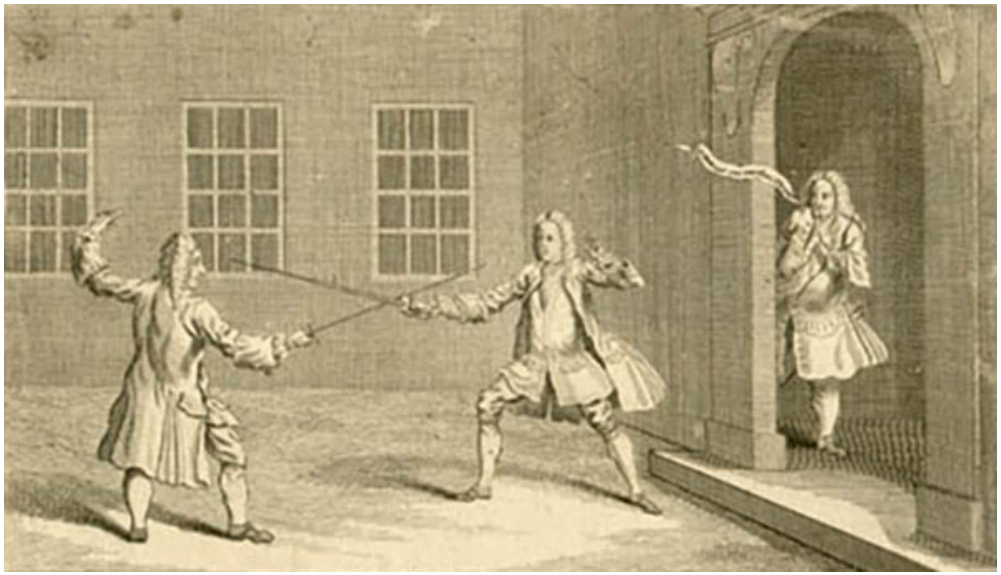
University College London

Philadelphia*: 6 August 2020

IMS Medallion Lecture



The Fundamental Duelity (sic) of Statistics



Parameters versus Distributions

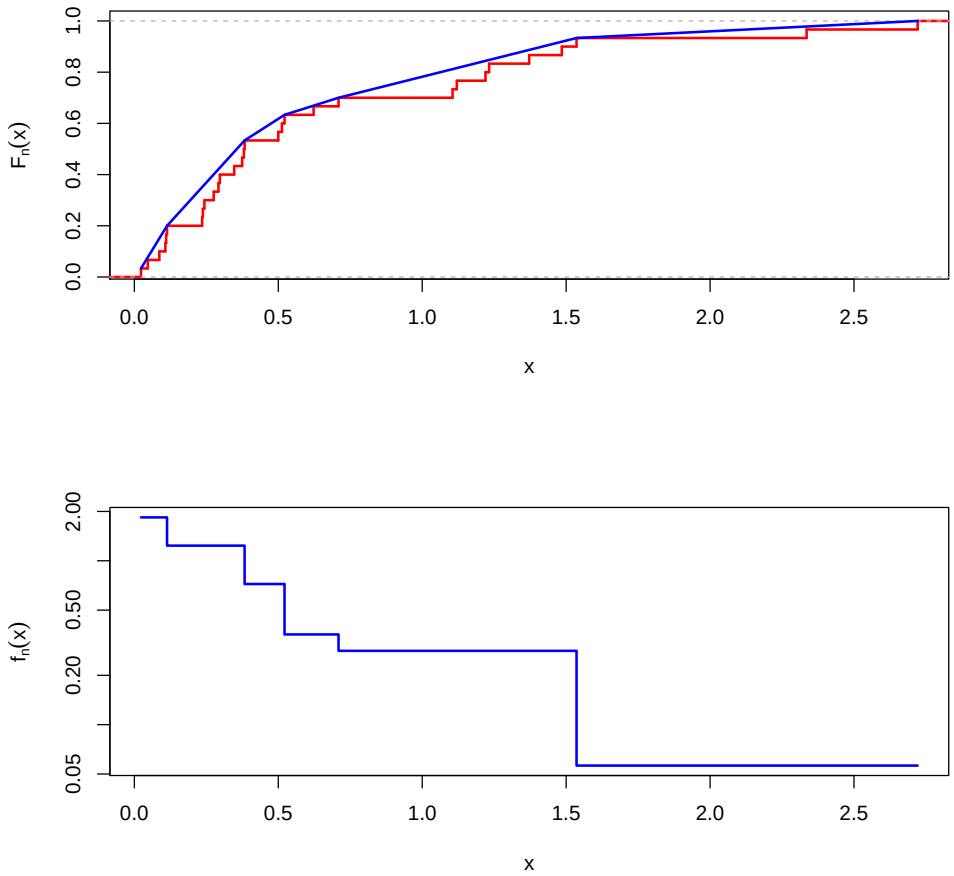
Laplace's parameters versus Quetelet's distributions [Stigler (1975]

The Fundamental Duelity (sic) of Statistics II

Parameters versus Distributions

Pearson's distributions versus Fisher's parameters [Stigler (1975]

The Prototype: Grenander's Monotone Density Estimator



Taut String: A Fully Automatic Histogram Estimator

Two Maximum Likelihood Formulations of Grenander

If you can't find a string, you can always power up your laptop and solve the maximum likelihood problem,

$$\max_{f}\{\int \log f(x)\, d\mathbb{F}_n(x) \mid f \text{ decreasing}, \int f(x)\, dx = 1\}.$$

Jumps in $\hat{f}$ occur at order statistics of the sample and at the origin.

Two Maximum Likelihood Formulations of Grenander

If you can't find a string, you can always power up your laptop and solve the maximum likelihood problem,

$$\max_f \{ \int \log f(x) d\mathbb{F}_n(x) \mid f \text{ decreasing}, \int f(x) dx = 1 \}.$$

Jumps in $\hat{f}$ occur at order statistics of the sample and at the origin. An alternative formulation also grounded in maximum likelihood involves writing our target density, f, as a scale mixture of uniforms,

$$\max_{G \in \mathcal{G}} \{ \int \log f(x) d\mathbb{F}_n(x) \mid f(x) = \int t^{-1} I(0 \leqslant x \leqslant t) dG(t) \},$$

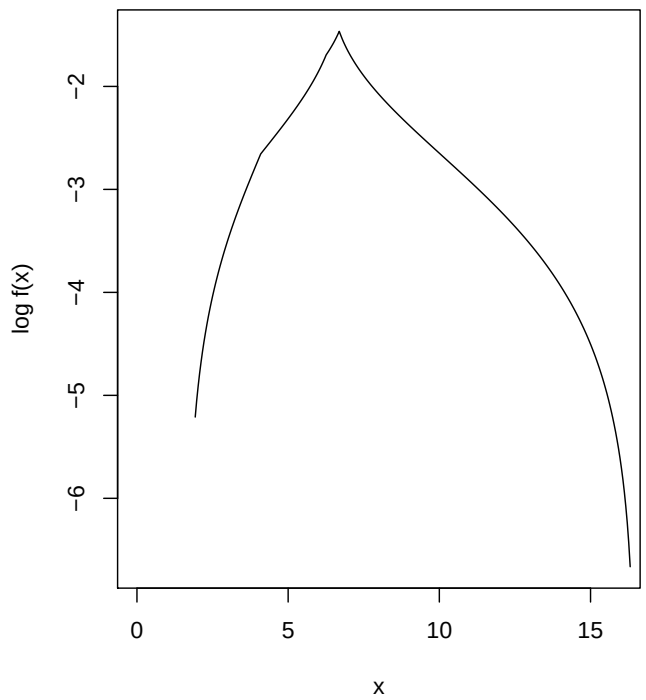
where $\mathcal{G}$ constitutes the set of proper distribution functions. This second formulation anticipates the nonparametric maximum likelihood estimator of Robbins (1950) and Kiefer and Wolfowitz (1956) that will be a main theme of the talk.

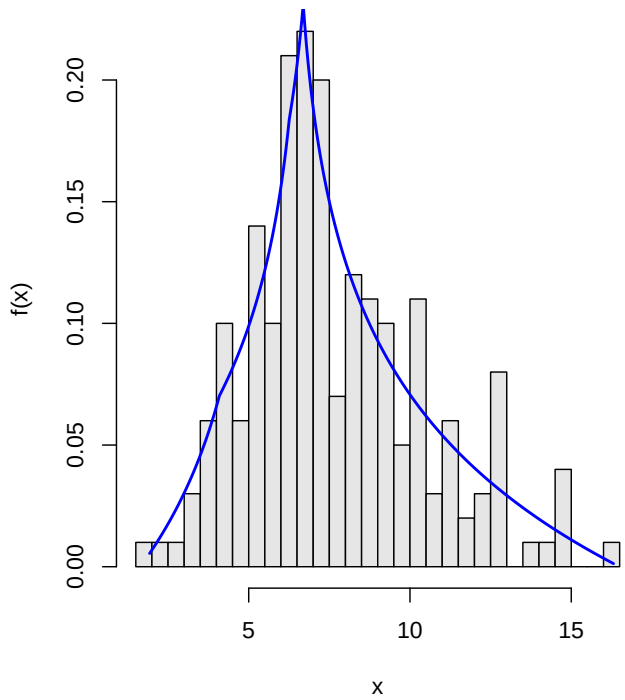
Log-Concave Density Estimation

What if we would like an MLE for unimodal densities?

$$\max_{f}\{\int \log f(x)\, d\mathbb{F}_n(x) \mid f \text{ log-concave}, \int f(x)\, dx = 1\}.$$

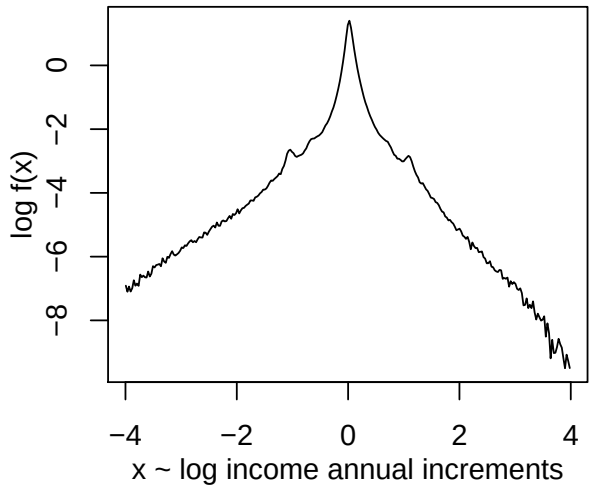
This can reformulated as just another convex optimization problem so computation is again quite easy and solutions are piecewise exponential as in this estimate of a gamma density.

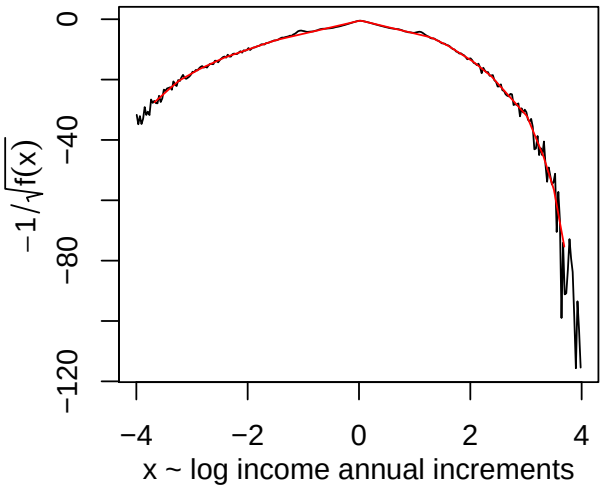




Log-Concaves Can't have Algebraic Tails

Weaker notions of concavity are needed to accommodate heavy tailed behavior. For example to model annual log income increments for households in the U.S. it is preferable to impose concavity on $-1/\sqrt{f(x)}$.





Guvenen et al (2015) show that U.S. earnings histories exhibit essentially Cauchy tail behavior.

Primal and Dual Log Concave Problems

The primal convex optimization problem for log concave densities is:

$$\min\left\{ n^{-1} \sum_{i=1}^{n} g(X_i) + \int e^{-g(x)} dx \;\middle|\; g \in \mathcal{K}(X) \right\}, \tag{P_1}$$

where $g = -\log f$ and $\mathcal{K}(X)$ denotes set of closed convex functions on the empirical support of the observations with associated dual problem,

$$\max\left\{ \int -f \log f dx \;\middle|\; f = \frac{d(\mathbb{Q}(X) - G)}{dx},\ G \in \mathcal{K}(X)^o \right\}, \tag{D_1}$$

where $\mathbb{Q}(X) = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical probability measure, $\mathcal{K}(X)^o$ is the polar cone associated with $\mathcal{K}(X)$. So the problem becomes one of maximizing Shannon entropy or minimizing the Kullback-Leibler divergence to the uniform subject to the concavity constraint.

Rényi "Likelihoods" and Quasi Concave Densities

Replacing Shannon entropy with a version of Rényi entropy we obtain the new dual and primal pairings, Koenker and Mizera (2010) consider,

$$\max\Big\{\frac{1}{\alpha}\int f^{\alpha}(y)\,dy \;\Big|\; f = \frac{d(\mathbb{Q}(X) - G)}{dy}, \quad G \in \mathcal{K}(X)^{o}\Big\}, \tag{D_α}$$

and

$$\min\Big\{\sum_{i=1}^{n} g(X_i) + \frac{|1-\alpha|}{\alpha}\int g^{\beta}\,dx \;\Big|\; g \in \mathcal{K}(X)\Big\}. \tag{P_α}$$

where f is now g^{β}, g is convex, f is α-concave and α and β are conjugate in the usual sense that $1/\alpha + 1/\beta = 1$. As α decreases we allow larger and larger classes of densities, culminating with $\alpha = -\infty$ by inclusion of all quasi-concave densities.

Beyond Shape Constraints – on to Mixtures

As we saw with the Grenander estimator maximum likelihood can be a vital tool for nonparametric estimation of mixture models. Robbins (1950) anticipated this and Kiefer and Wolfowitz (1956) filled in many details. Consider mixtures of the form,

$$f(x) = \int \varphi(x, \theta) dG(\theta),$$

where φ is a known parametric distribution and G is an unknown mixing distribution, we have the primal problem

$$\min_{G \in \mathcal{G}} \{ - \sum_{i=1}^{n} \log f(x_i) \mid f(x_i) = \int \varphi(x_i, \theta) dG(\theta), \ i = 1, ..., n \},$$

The associated dual problem is, Lindsay (1981),

$$\max \{ \sum_{i=1}^{n} \log v_i \mid \sum_{i=1}^{n} v_i \varphi(x_i, \theta) \leqslant n \text{ for all } \theta \}$$

Laird (1978) proposed an EM computational method; modern interior point and gradient descent methods offer efficient alternative methods.

Compound Decisions and the Gaussian Sequence Model

Following Robbins (1956) and the empirical Bayes approach to compound decisions, suppose we observe $\{y_1, y_2, \cdots, y_n\}$ with $Y_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \cdots, n$, and face the quadratic loss function, for nonlinear shrinkage,

$$L(\hat{\theta}, \theta) = n^{-1} \sum_{i=1}^{n} (\hat{\theta} - \theta)^2.$$

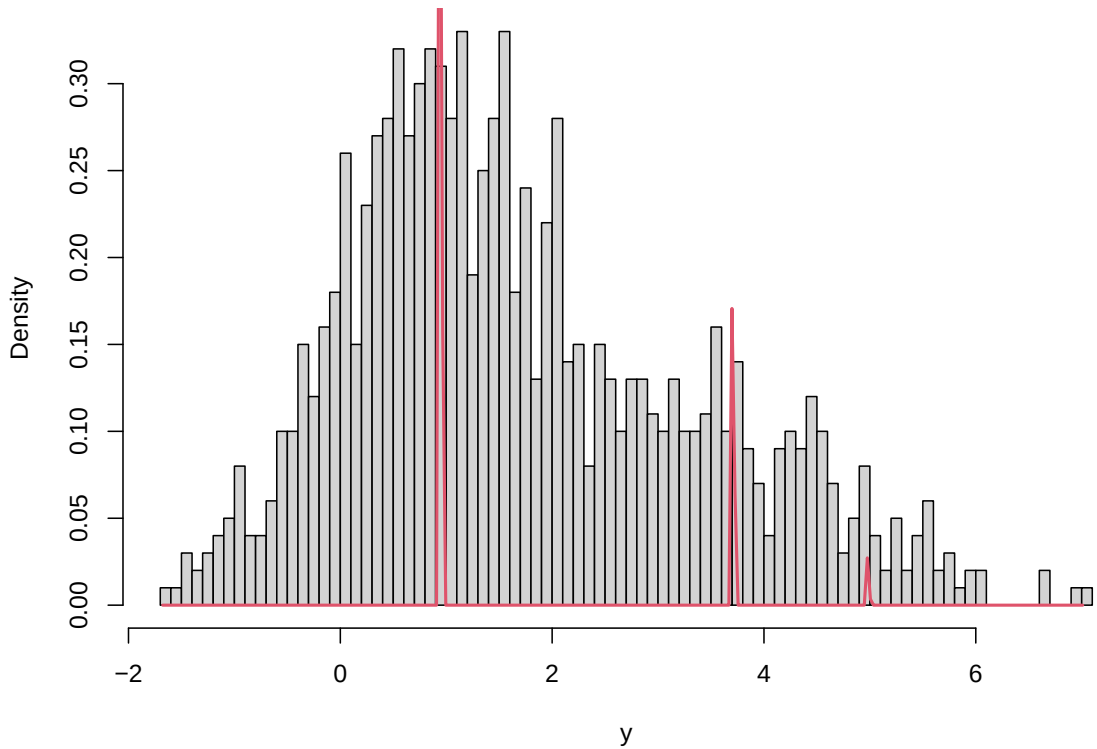
Adopting the presumption that the θ_i are drawn iidly from G, the Bayes rule is given by Tweedie's formula,

$$\hat{\theta}_i = y_i + f'(y)/f(y).$$

Rather than estimating each of the incidental parameters θ_i, independently, thereby entailing a loss of 1, we estimate their distribution, G, and then "borrow strength from the ensemble" to estimate the Bayes rule. In the terminology of Efron (2016, 2019) this is called g-modeling.

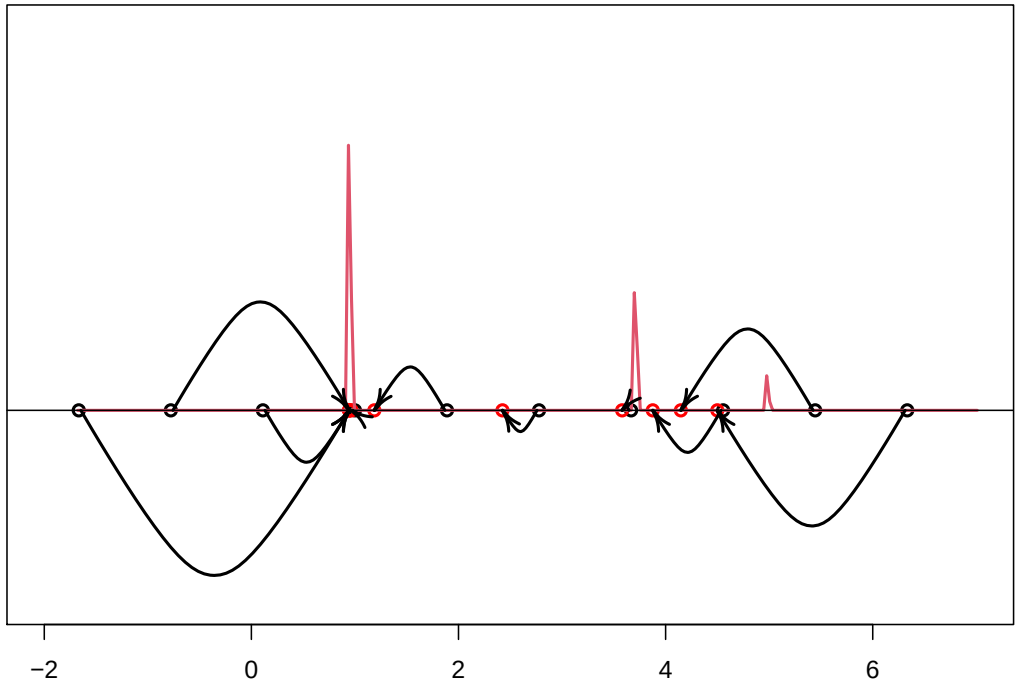
A Simple Discrete Mixture Example

Consider the simple model, $Y_k \ \mathcal{N}(\theta, 1)$, with $\theta \in \{1, 3\}$ with probabilities $(0.75, 0.25)$ respectively. We draw a sample of $n = 1000$, Y's, plot their histogram, and then overplot the Kiefer-Wolfowitz NPMLE in red.



Tweedie Shrinkage for Posterior Means

Given our $\hat{\mathrm{G}}$ we can compute a posterior mean estimate for any value of y. What does this look like?



Tweedie shrinkage is quite smart about adapting shrinkage to the form of the posterior. No longer are we simply shrinking toward one fixed value.

Minimalist G-Modeling and Alternatives

When φ is Gaussian we have a classical deconvolution problem, but Fourier methods perform poorly, while maximum likelihood in several forms performs quite brilliantly.

- Efron's logspline approach expresses $g = G'$ as a natural spline:

$$\log g(\theta) = \sum_{j=1}^{p} \alpha_j \psi_j(\theta),$$

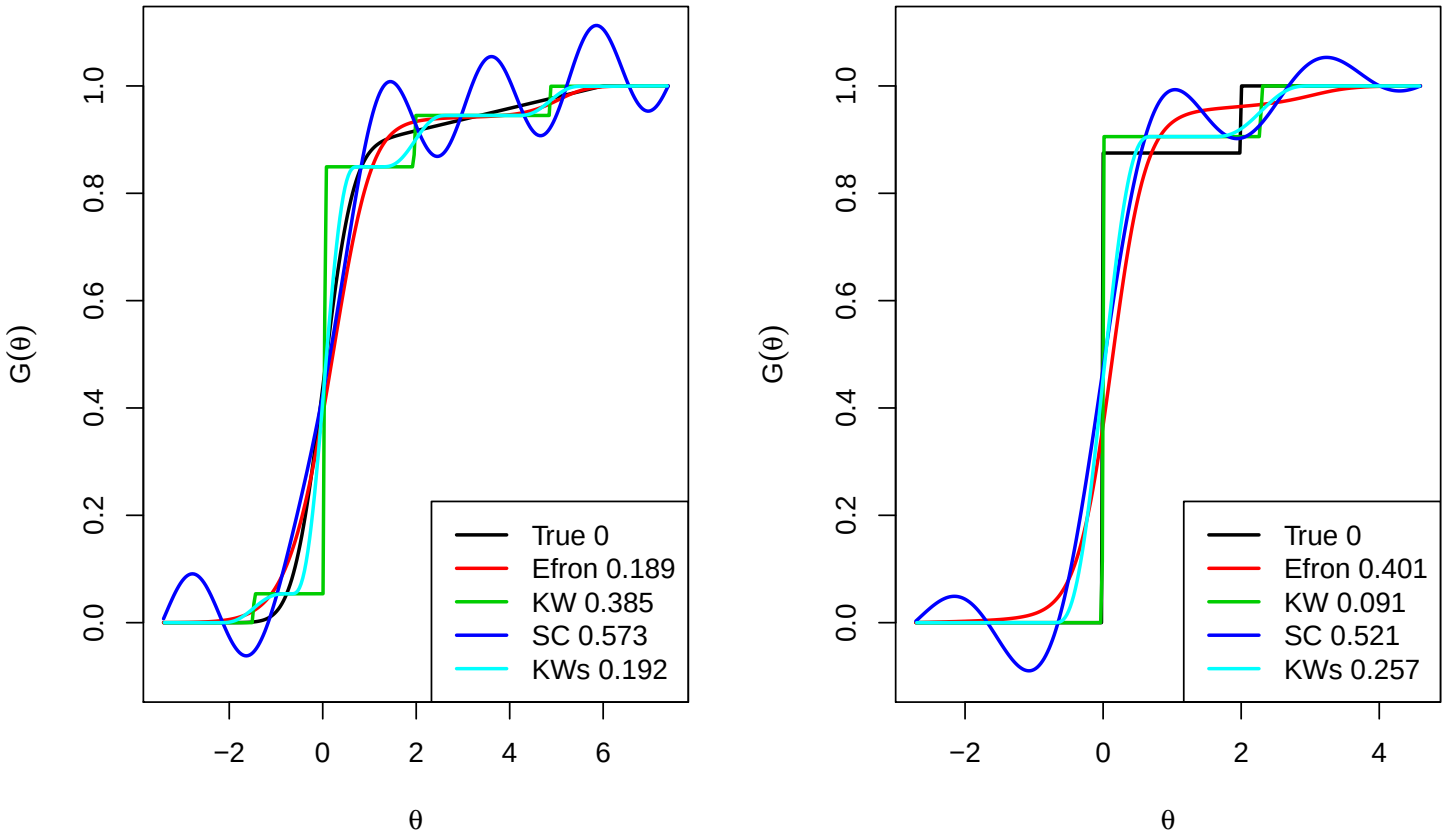
  and estimates the parameters $\alpha \in \mathbb{R}^p$ by penalized maximum likelihood.
- The Kiefer and Wolfowitz NPMLE yields a discrete G typically with only a few atoms, and has the advantage that it is tuning parameter free.

Both approaches share the advantage that they are applicable to the general class of mixture problems, not only to Gaussian deconvolution.

Two variants of a simulation setting from Efron (2016)

$$Y_i = \theta_i + u_i, \quad u_i \sim \mathcal{N}(0,1), \quad \theta_i \sim G(\theta) = \tfrac{1}{8}\Phi(\theta/6) + \tfrac{7}{8}\Phi(2\theta).$$

$$Y_i = \theta_i + u_i, \quad u_i \sim \mathcal{N}(0,1), \quad \theta_i \sim G(\theta) = \tfrac{7}{8}I(\theta > 0) + \tfrac{1}{8}I(\theta > 2)$$



Wasserstein (L_1) distances between $\hat{G}_n$ and true G in legend

Two variants of a simulation setting from Efron (2016)

The take-away

- Efron is better for smooth G, KW is better for discrete G,
- Kernel deconvolution a la Stefanski and Carroll (1990) is awful,
- Kernel smoothing of the KW NPMLE is good for smooth G, but reintroduces a bandwidth choice.

Should we believe any of this based on one realization? A simulation with $n = 1000$ and 1000 replications may be more convincing?

	Efron	Kernel	NPMLE	NPMLEs
Smooth	0.185	0.591	0.342	0.180
Discrete	0.409	0.718	0.156	0.280

Mean Wasserstein (L_1) Error

Binary Response Modeling with Random Coefficients

There are many other applications of the Kiefer-Wolfowitz NPMLE, to survival models, longitudinal data, multiple testing, etc. But I would like to briefly touch upon some recent work on binary response with Jiaying Gu because it represents an extreme variant of my main theme: the unlikely likelihood, with distribution as parameter.

Binary Response Modeling with Random Coefficients

There are many other applications of the Kiefer-Wolfowitz NPMLE, to survival models, longitudinal data, multiple testing, etc. But I would like to briefly touch upon some recent work on binary response with Jiaying Gu because it represents an extreme variant of my main theme: the unlikely likelihood, with distribution as parameter.

The model: we observe $(y_i, x_i, w_i) : i = 1, \cdots n$ where $y_i \in \{0, 1\}$, $x_i \in \mathbb{R}^{d+1}$, $w_i \in \mathbb{R}^p$ and suppose,

$$y_i = 1(x_i^\top \beta_i + w_i \theta_0 \geqslant 0).$$

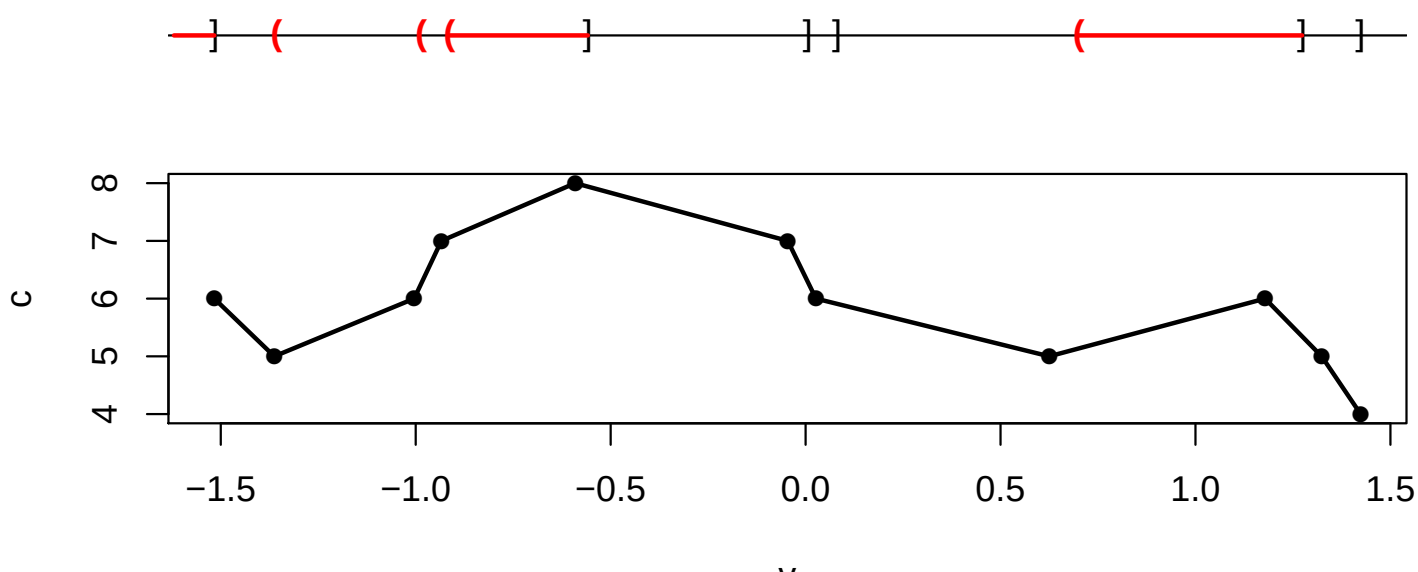
The random coefficients β_i are drawn independently of x_i and w_i and iidly from a distribution G_0. We will need to normalize β_i since it is only identified up to scale. Our objective is to estimate the pair (θ_0, G_0), I will (almost entirely) ignore the role of θ.

The Current Status (Cosslett) Model

The simplest setting has $x_i = (1, -v_i)^\top$, with no w_i, and normalized so that $\beta_i = (\eta_i, 1)^\top$,

$$\mathbb{P}(y = 1|v) = \int 1(\eta \geqslant v) dG_\eta(\eta).$$

Given a sample $\{(y_i, v_i)\ i = 1, \cdots, n\}$, Each point defines a half line: if $y_i = 0$ then the interval is $R_i = (-\infty, v_i)$, while if $y_i = 1$ the interval is $R_i = [v_i, \infty)$. The R_i's form a partition of $n + 1$ intervals of $\mathbb{R}$, that we denote by I_j, for $j = 1, \cdots, n + 1$. Counts are defined for each interval.



Only (Some) Locally Maximal Intervals Really Count

The NPMLE for the mixing distribution solves

$$\min_{g \in \mathcal{S}_m} \{ - \sum_{i=1}^{n} \log f_i \mid Ag = f \},$$

where $A = (a_{ij}) = 1(\eta_j > v_i,\ y_i = 1) + 1(\eta_j \leqslant v_i,\ y_i = 0)$, $\mathcal{S}_m$ denotes the unit simplex and g_j denotes the mass associated with interval with endpoint η_j.

The non-negativity requirement on the elements of $g \in \mathcal{S}_m$ assures that only a few of the remaining locally maximal intervals receive postive mass at a NPMLE solution. No further regularization is required. For Gaussian data the number of strictly possitive mass points is roughly of order, $\mathcal{O}(\sqrt{n})$. Intervals may be prescreened since only intervals that have locally maximal counts can receive positive mass for the NPMLE.

NPMLE for Bivariate G_η

When the random parameter η is bivariate things get more interesting. We have half spaces instead of half lines and polygons instead of intervals so computation become more complicated. Our binary response is generated as,

$$\mathbb{P}(y_i = 1 | z_i, v_i) = \mathbb{P}(\eta_{1i} + z_i \eta_{2i} \geqslant v_i).$$

Each pair, (z_i, v_i), defines a plane that divides $\mathbb{R}^2$ into two halfspaces, an "upper" one corresponding to realizations of $y_i = 1$, and a "lower" one for $y_i = 0$. Let R_i denote these halfspaces and $F_\eta\{R_i\}$ be the probability assigned to each R_i by the distribution G_η, so the log likelihood is,

$$\ell(G_\eta) = \sum_{i=1}^{n} \log F_\eta\{R_i\}.$$

Theorem The NPMLE assigns positive mass only to polygons with locally maximal counts of the number of their intersecting halfspaces.

“Facing Up to Arrangements”

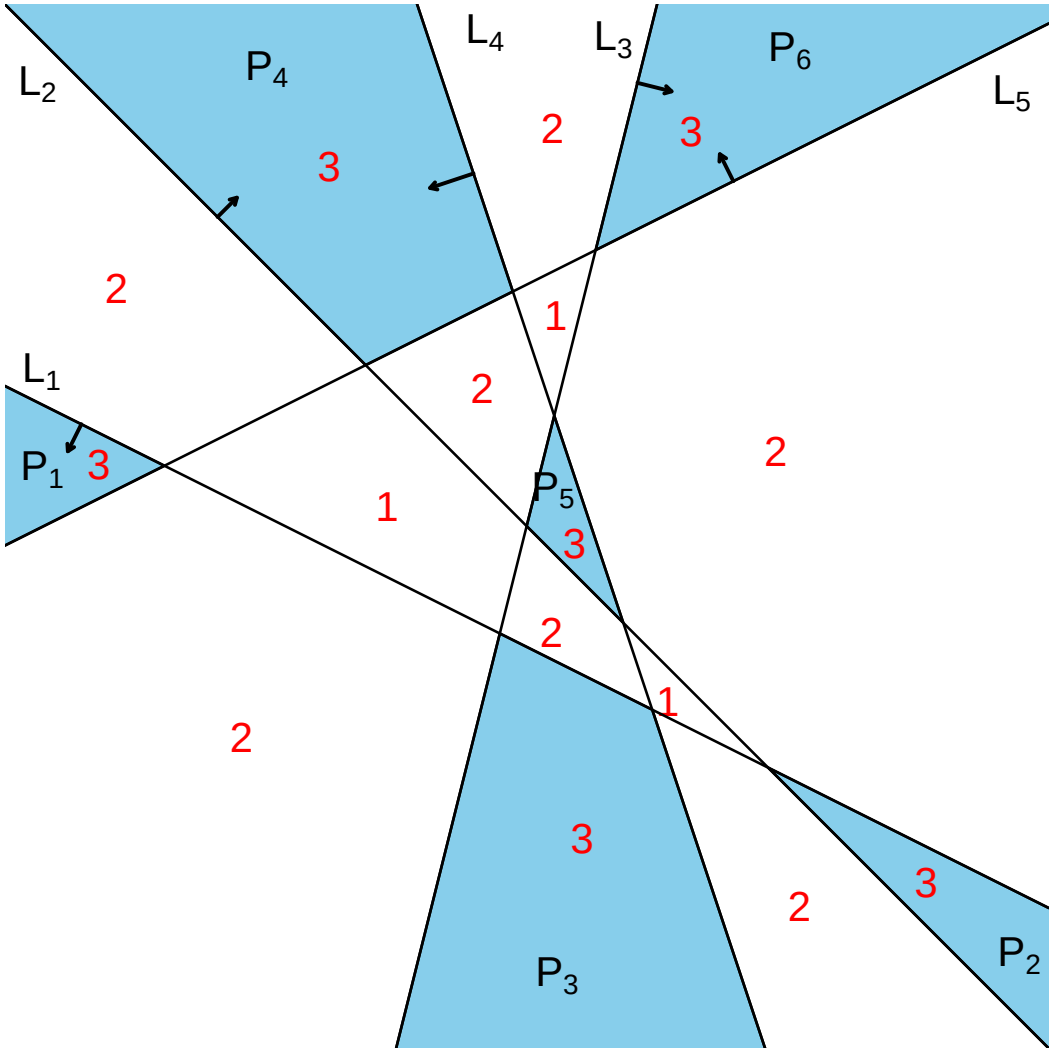
Over the last 50 years or so there has been considerable progress in algebraic and computational geometry on what is called “hyperplane arrangements”. Given n hyperplanes $H_i : i = 1, \cdots, n$ in $\mathbb{R}^d$, a first question might be: How many polytopes do they form? When the hyperplanes are in “general position” then the question has the following elegant answer:

$$M(n, d) = \sum_{k=0}^{d} \binom{n}{k}$$

This was apparently first proven by Buck (1943) and greatly elaborated in a MIT thesis by Zaslavsky (1975), titled “Facing up to Arrangements”. These results and subsequent work of others greatly facilitates the bookkeeping required to implement the NPMLE for this problem.

A (Pathological) Toy Example

Five observations and 16 polygons, of which six are locally maximal.



The NPMLE: Three Equivalent Versions

Fix θ and let $F(z, v, \theta) = \{\eta | z^\top \eta - v + w^\top \theta \geqslant 0\}$. The NPMLE solves,

$$\max_{G \in \mathcal{G}} \sum_{i=1}^{n} y_i \log[\mathbb{P}_G(F(z_i, v_i, \theta))] + (1 - y_i) \log[1 - \mathbb{P}_G(F(z_i, v_i, \theta))].$$

Given locally maximal cells, $\{C_i, \cdots, C_{M^*}\}$, define a n by M^* matrix A with $A_{ij} = 1\{C_j \subset F(z_i, v_i, \theta)\}$ if $y_i = 1$ and $1 - 1\{C_j \subset F(z_i, v_i, \theta)\}$ if $y_i = 0$,

$$\min \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log g_i \mid g_i = \sum_j a_{ij} p_j, \sum_j p_j = 1, p_j \geqslant 0 \right\}$$

The dual problem is preferable since M^* is typically much larger than n,

$$\max \left\{ \sum_{i=1}^{n} \log \pi_i \mid \sum_{i=1}^{n} a_{ij} \pi_i \leqslant n \text{ for all } j \right\}$$

The NPMLE assigns mass $p_i = \pi_i$ to cell C_i for $i = 1, \cdots, M^*$. This convex optimization problem can be solved efficiently with Mosek, for example. Profile likelihood can then be optimized to obtain $\hat{\theta}_n$.

Identification and Asymptotics

Returning to our original model with profiled parameters θ_0 as well as F_0 to be estimated,

$$y_i = 1(x_i^\top \beta_i + w_i \theta_0 > 0).$$

Theorem

Under the following assumptions:

- *A1* *The random vectors* (x_i, w_i) *and* β_i *are independent and* $[X \dot{:} W]$ *has full column rank.*
- *A2* *The parameter space* Θ *is a compact subset of a Euclidean space and* $\theta_0 \in \Theta$. *The space* $\mathcal{F}$ *of probability distributions for* β_i *is supported on the* d*-dimensional unit sphere, and there exists a vector* $c \neq 0$ *such that* $\mathbb{P}_F(c^\top \beta_i > 0) = 1$ *for all* $F \in \mathcal{F}$.
- *A3* *The distribution of* $(z_i^\top, v_i)$ *is absolutely continuous on* $\mathbb{R}^d$ *and* $w_i^\top \theta_0$ *is absolutely continuous both possessing an everywhere positive density.*

the parameter (θ_0, F_0) *is identified, and the NPMLE is strongly consistent.*

The proof is very Wald-ish, as in Kiefer and Wolfowitz (1956).

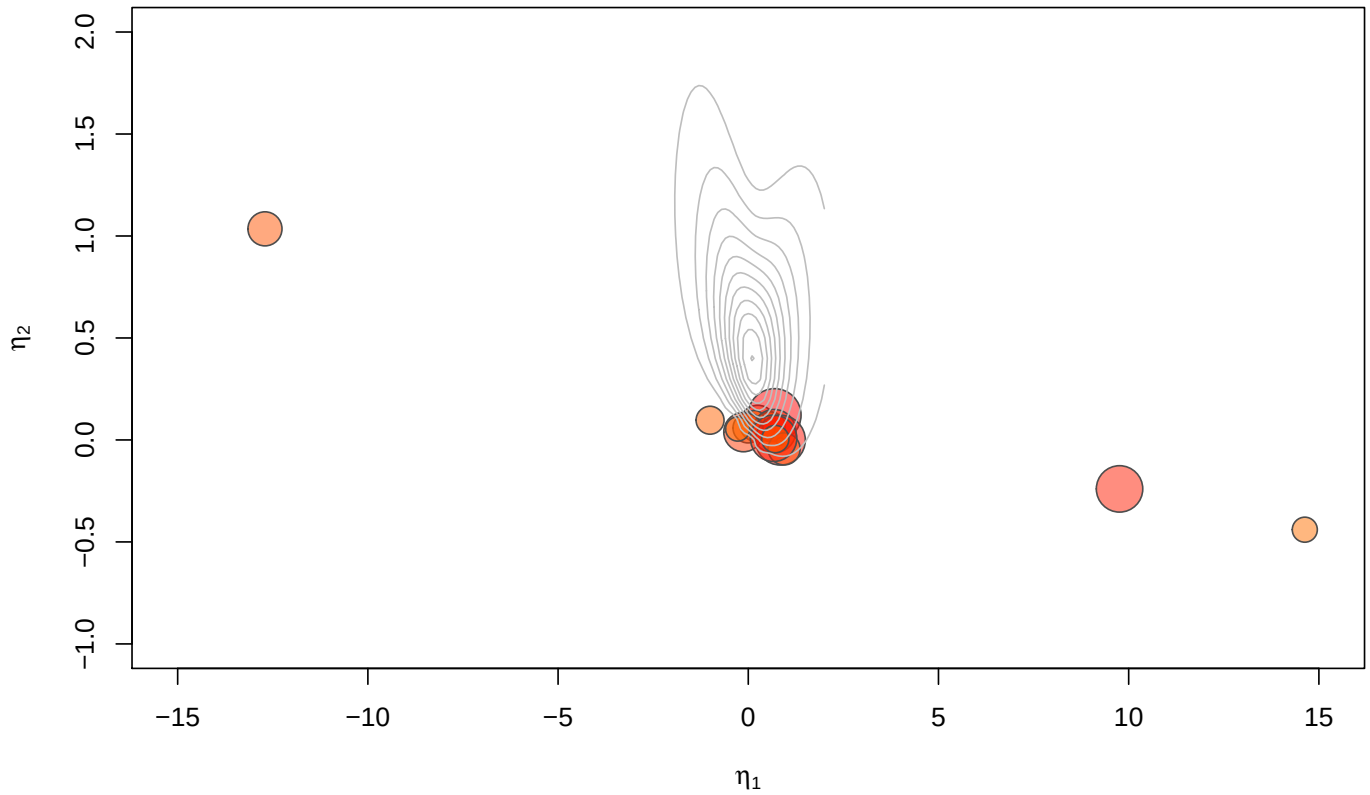
Journey to work in Washington DC

- Either by automobile, $y = 1$, or public transit, $y = 0$,
- There are two relevant covariates:
  - difference in commuting time, z, in minutes, and
  - difference in commuting cost, v, in dollars per trip.
- The coefficient on v is normalized to equal 1
- There are no other covariates, w, with fixed coefficients.
- Observations are stratefied by the number of cars, k, owned by the household, and analysis is conditional on k.

Cars	0	1	2	>2
n	79	355	316	92
% Transit	78	14	4	0

Subsample of Commuters Owning One Car

Grey contours are estimated density of Gautier and Kitamura's (2013) deconvolution estimator, pink circles depict mass points of NPMLE.



The Unlikelihood of Quantile Regression

To conclude, I should say a few words about quantile regression. It is a ridiculously simple idea, instead of minimizing sums of squared errors from a linear predictor, why not minimize sums of absolute errors, or asymmetrically weighted absolute errors,

$$\hat{\beta}(1/2) = \text{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} |y_i - x_i^\top b|$$

$$\hat{\beta}(\tau) = \text{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$. Rather than a global conditional mean model we have a local conditional quantile model.

A QRious likelihood

While viewing the usual quantile regression objective for a single quantile as a global model seems like a very bad idea, there has also been increasing interest in treating an ensemble of QR models as something that could function more like a likelihood. When we write,

$$\hat{\beta}(\tau) = \text{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b),$$

it is tempting to see it as a estimate of the entire conditional quantile function. Indeed if we were to consider the weighted version,

$$\hat{\beta}(\tau) = \text{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} w_i \rho_\tau(y_i - x_i^\top b),$$

and let $w_i = f_i(y_i|x_i)$, the conditional density of the response, we have an efficient parametric model for each conditional quantile function. Imposing some smoothness on the resulting quantile regression process makes it look a lot like penalized likelihood.

The Unlikelihood of Quantile Regression

Again, we have unlikely likelihoods for distributional objects without making distributional assumptions:

- Weights can be estimated via the relation $\partial Q(\tau|x)/\partial\tau = 1/f(Q(\tau|x))$, Wei and Carroll (2009), Yang and He (2012), Feng, Chen and He (2015), Arellano, Blundell and Bonhomme (2017), Yang, Narisetty and He (2018), ...
- Simulation from the model is easy, via $Q(U_i|x_i)$, suggesting a possible connections to approximate Bayesian computation,
- Linearity (in parameters) assumptions and the data magically produce a likelihood-like object,
- Extensions to multivariate response via optimal transport, Wei (2009), Chernozhukov, Galichon, Hallin and Henry (2017), Carlier, Chernozhukov and Galichon (2016)
- Local likelihood variants enable one to focus exclusively on tail behavior when desirable: Wang, Li and He (2012)

Some Concluding Slogans

- Think globally, estimate locally.
- “Every parameter would like to grow up to be a distribution.”
- Mixtures aren’t “like tequila” [Wasserman] and shouldn’t be avoided.
- There is never “an effect,” there is always a distribution of effects.

Some Concluding Slogans

- Think globally, estimate locally.
- “Every parameter would like to grow up to be a distribution.”
- Mixtures aren’t “like tequila” [Wasserman] and shouldn’t be avoided.
- There is never “an effect,” there is always a distribution of effects.

Virtual T-shirts available in the virtual lobby.

Selected References I

ARELLANO, M., R. BLUNDELL, AND S. BONHOMME (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85(3), 693–734.

CARLIER, G., V. CHERNOZHUKOV, AND A. GALICHON (2016): "Vector quantile regression: An optimal transport approach," *Ann. Statist.*, 44, 1165–1192.

CHERNOZHUKOV, V., A. GALICHON, M. HALLIN, AND M. HENRY (2017): "Monge–Kantorovich depth, quantiles, ranks and signs," *Annals of Statistics*, 45, 223–256.

EFRON, B. (2016): "Empirical Bayes deconvolution estimates," *Biometrika*, 103, 1–20.

——— (2019): "Bayes, Oracle Bayes and Empirical Bayes," *Statistical Science*, 34, 177–201.

FENG, Y., C. Y., AND X. HE (2015): "Bayesian quantile regression with approximate likelihood," *Bernoulli*, 21, 832–850.

GAUTIER, E., AND Y. KITAMURA (2013): "Nonparametric estimation in random coefficients binary choice models," *Econometrica*, 81, 581–607.

GRENANDER, U. (1956): "On the theory of mortality measurement," *Scandinavian Acturial Journal*, 39, 125–153.

GU, J., AND R. KOENKER (2020): "Random Coefficient Binary Response," *Journal of the American Statistical Association*, forthcoming.

Selected References II

GUVENEN, F., F. KARAHAN, S. OZKAN, AND J. SONG (2019): "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Dynamics?," preprint.

HAN, Q., AND J. A. WELLNER (2016): "Approximation and estimation of s -concave densities via Rényi divergences," *Annals of Statistics*, 44, 1332–1359.

KIEFER, J., AND J. WOLFOWITZ (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., AND J. GU (2019): "Comment: Minimalist G-Modeling," *Statistical Science*, 34, 209–213.

KOENKER, R., AND I. MIZERA (2010): "Quasi-concave density estimation," *Annals of Statistics*, 38, 2998–3027.

KOENKER, R., AND I. MIZERA (2014): "Convex Optimization, Shape Constraints, Compound Decisions and Empirical Bayes Rules," *J. of Am. Stat. Assoc.*, 109, 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," *Annals of Statistics*, 11, 86—-94.

ROBBINS, H. (1950): "A Generalization of the Method of Maximum Likelihood; Estimating a Mixing Distribution (Abstract)," *The Annals of Mathematical Statistics*, 21, 314–315.

Selected References III

——— (1956): "An Empirical Bayes Approach to Statistics," in *Proceedings of the Third Berkeley Symposium*, vol. I, pp. 157–163. University of California Press: Berkeley.

STIGLER, S. (1975): "The transition from point to distribution estimation," *Bulletin of the International Statistical Institute*, 46, 332–340.

WANG, H. X., D. LI, AND X. HE (2012): "Estimation of High Conditional Quantiles for Heavy-tailed Distributions," *Journal of the American Statistical Association*, 107, 1453–1464.

WEI, Y. (2008): "An Approach to Multivariate Covariate-Dependent Quantile Contours With Application to Bivariate Conditional Growth Charts," *Journal of the American Statistical Association*, 103, 397–409.

WEI, Y., AND R. J. CARROLL (2009): "Quantile Regression With Measurement Error," *Journal of the American Statistical Association*, 104, 1129–1143.

YANG, X., N. NARISETTY, AND X. HE (2018): "A new approach to censored quantile regression estimation," *Journal of Computational and Graphical Statistics*, 27, 417–425.

YANG, Y., AND X. HE (2012): "Bayesian empirical likelihood for quantile regression," *Annals of Statistics*, 40, 1102–1131.

ZASLAVSKY, T. (1975): *Facing up to arrangements: Formulas for partitioning space by hyperplanes*, vol. 154 of *Memoirs of the AMS*. American Mathematical Society.