Nonparametric Maximum Likelihood Methods for Mixture Models

Roger Koenker

UCL/cemmap

RSSB Annual Meeting: 18 October 2018

Joint work with Jiaying Gu (U. of Toronto)

A General Paradigm for Mixture Models

Suppose we begin confidently with a parametric model,

$$y_i \sim \varphi(y, \theta) \quad i = 1, \cdots, n,$$

but lose our nerve and admit there may be θ heterogeneity, so,

$$y_i \sim \varphi(y, \theta_i) \quad i = 1, \cdots, n.$$

If the θ_i's are generated iidly from the distribution F_0, this is the de Finetti mixture model and the y_i are exchangeable with density,

$$y_i \sim g(y) = \int \varphi(y, \theta) dF_0(\theta) \quad i = 1, \cdots, n.$$

The Average Man is Not Enough

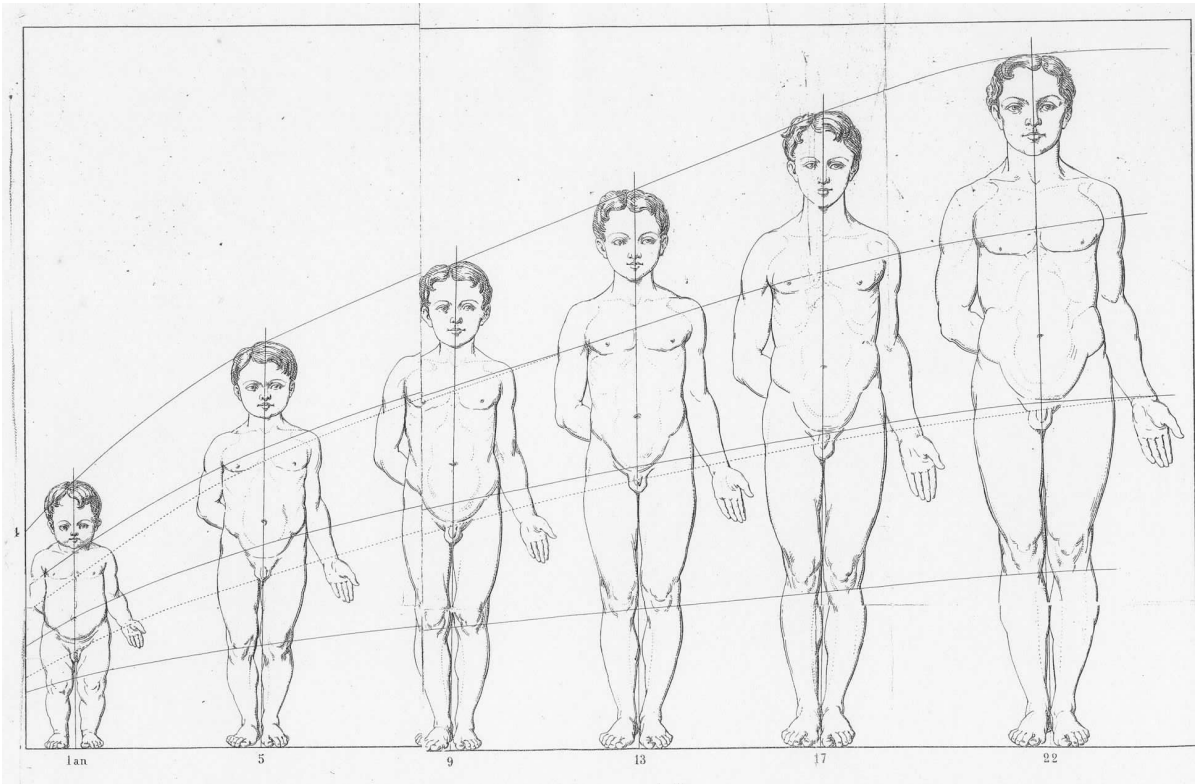


Figure: Source: Quetelet's (1871) Anthropométrie

The Average Man is Not Enough

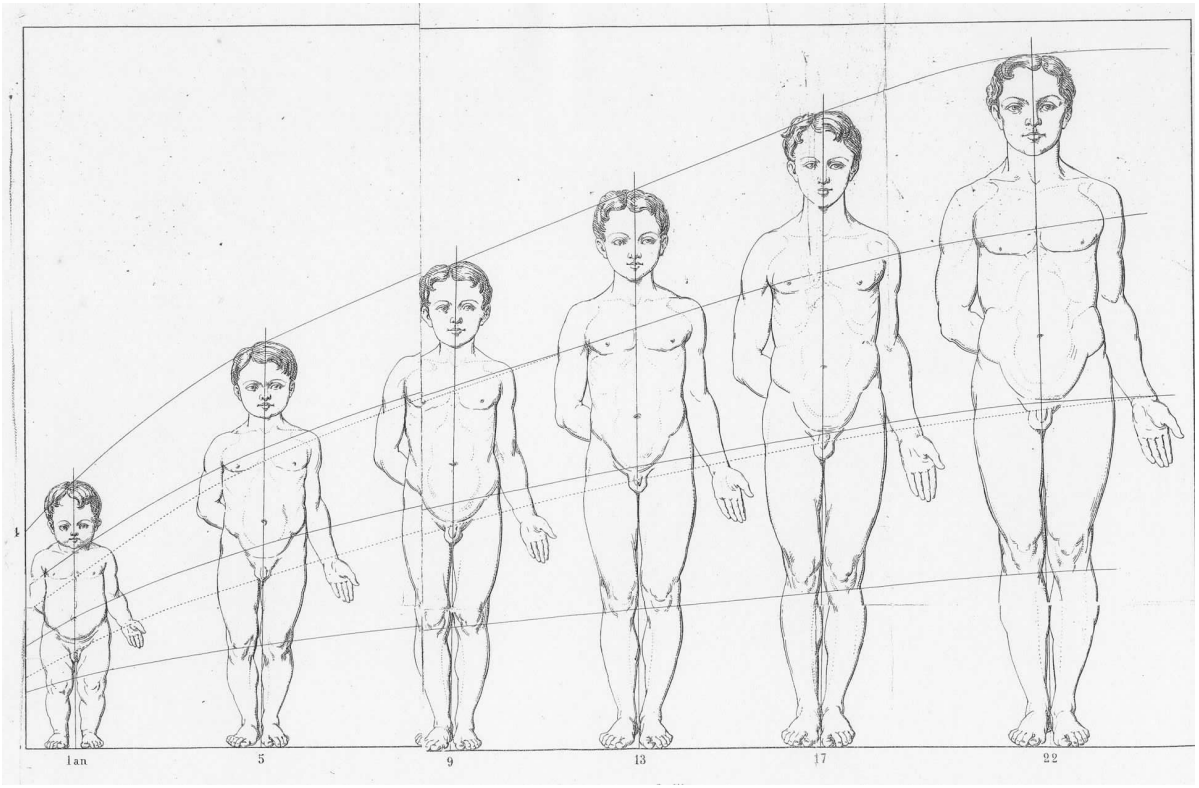


Figure: Source: Quetelet's (1871) Anthropométrie

One size does not fit all.

Some Examples

- **Robbins's Compound Decisions** Binary classification with Gaussian noise, φ is Gaussian,
- **Biased Thumbtack Flipping** Diaconis experiments with thumbtacks flipped on various surfaces, φ is binomial
- **Gaussian Sequence Model** Johnstone and Silverman simulation experiments meant to mimic genomic applications, φ is Gaussian.
- **Weibull Survival Model** Carey et al medfly experiments, φ is Weibull.
- **Binary Response with Random Coefficients** Modal choice for journey to work, φ is an indicator function.

Robbins (1951) Compound Decisions

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

Robbins (1951) Compound Decisions

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 then shows that this rule remains minimax for $n > 1$.
But isn't it foolish?

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 $p = 1/2$, repeatedly.

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 $p = 1/2$, repeatedly.
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 $p = 1/2$, repeatedly.
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,

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

This is a Bayes rule shrinkage adjustment. But we don't know p.

Hierarchical Bayes

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

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 (mea culpa!) Bayes rule,

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

Empirical Risk for Several Decision Rules

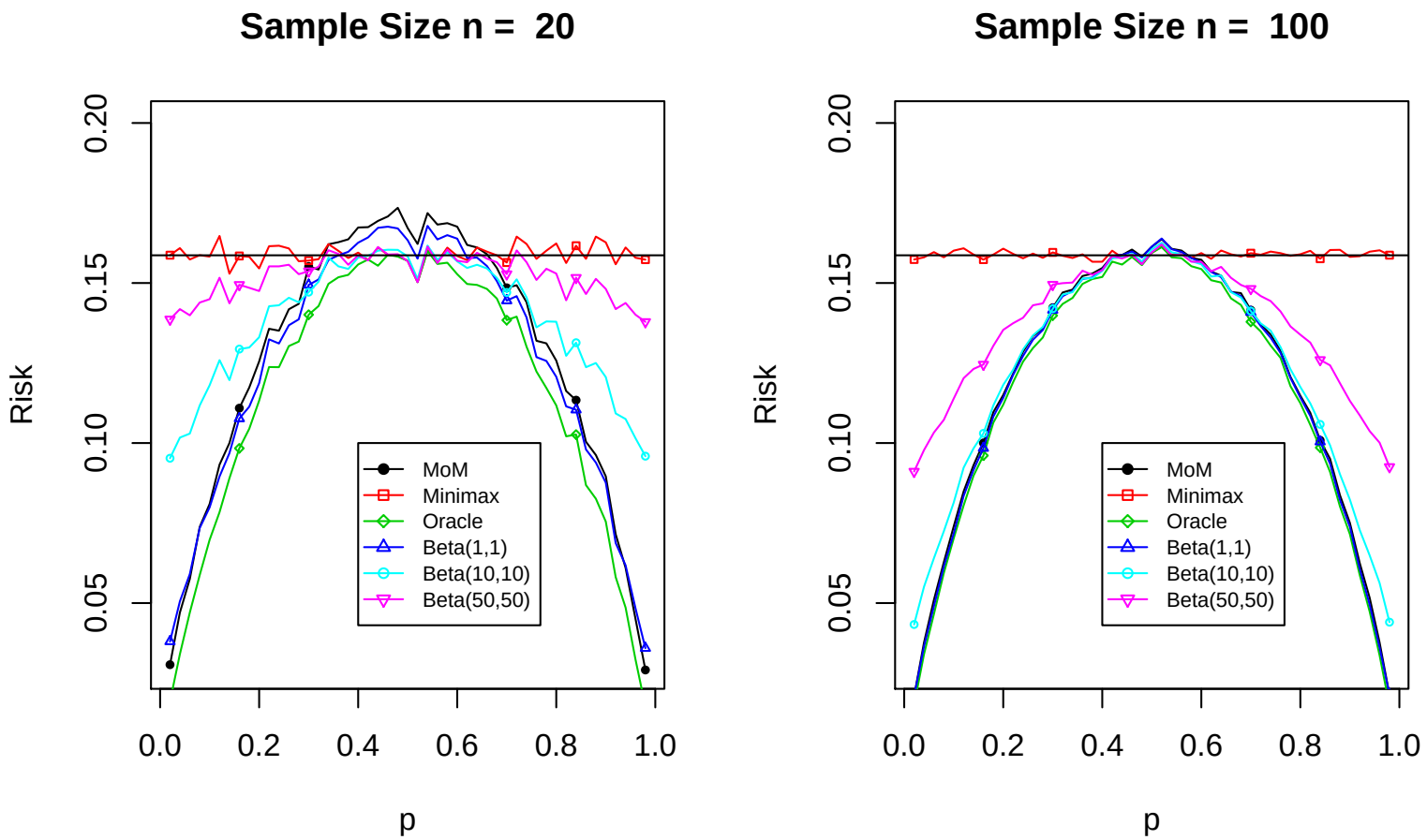


Figure: Mean absolute loss over 1000 replications.

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 F, so marginally,

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

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

Kiefer and Wolfowitz NPMLE's for Mixture Models

- Generic Problem

$$Y_i|\theta \sim \varphi(y|\theta), \quad \theta \sim F, \quad Y_i \sim g(y) = \int \varphi(y|\theta)\,dF(\theta)$$

$$\max_{F \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \log g(y_i) \mid g(y) = \int \varphi(y|\theta)\,dF(\theta) \right\}$$

- Generic Solution
  - Objective is strictly convex and constraints are polyhedral, so solutions are unique.
  - Constraints may be implemented on a fine grid, but solutions are discrete with only a few mass points.
  - Rather than impose a prior for F, we estimate it!

The Grouped Robbins Problem

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

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

where $g_i = g(y_{i1}, \cdots, g_{iJ})$, 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 dimensional simplex, $\mathcal{S}$.
The Diaconis thumbtack problem is very similar except φ is binomial rather than Gaussian.

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,

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

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

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

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

Needless [sic] and Haystacks

It is commonly assumed that F 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

Table: 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.

Bayesian Deconvolution

Tweedie's formula reveals that we don't really need to estimate the mixing distribution F to construct an estimate of the Bayes rule for the Gaussian sequence compound decision problem, all we need is a good estimate of the mixture density g. We have three options (at least):

- Classical deconvolution a la Stefanski and Carroll (1990),
- Efron's log-spline approach: let $\log f(y, \beta) = \sum_{j=1}^{p} \beta_j \psi_j(t) - \psi_0(\beta)$ so estimating dF is reduced to finding the MLE for β,
- Kiefer-Wolfowitz NPMLE by solving the convex program:

$$\max_f \{ \sum_{i=1}^{n} \log g(y_i) | g = Af, \ f \geqslant 0, 1_m^\top f = 1 \}$$

A Comparison for an Efron Simulation Setting

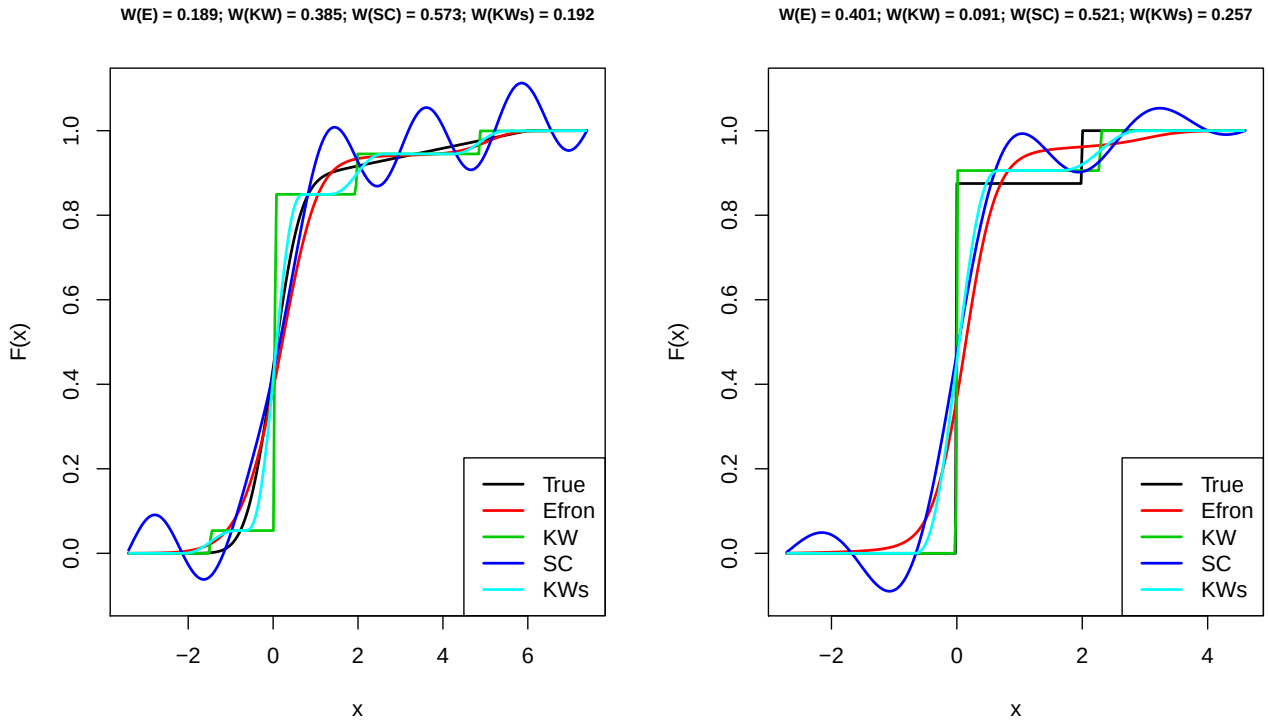


Figure: Comparison of Mixing Distribution Estimators: Left panel is smooth target estimators, right panel is discrete target estimators. Wasserstein (L1) errors reported in the panel headings.

A Weibull Mixture Problem

Carey et al (1992) studied survival times for 1.2 million medflies and reached several surprising conclusions:

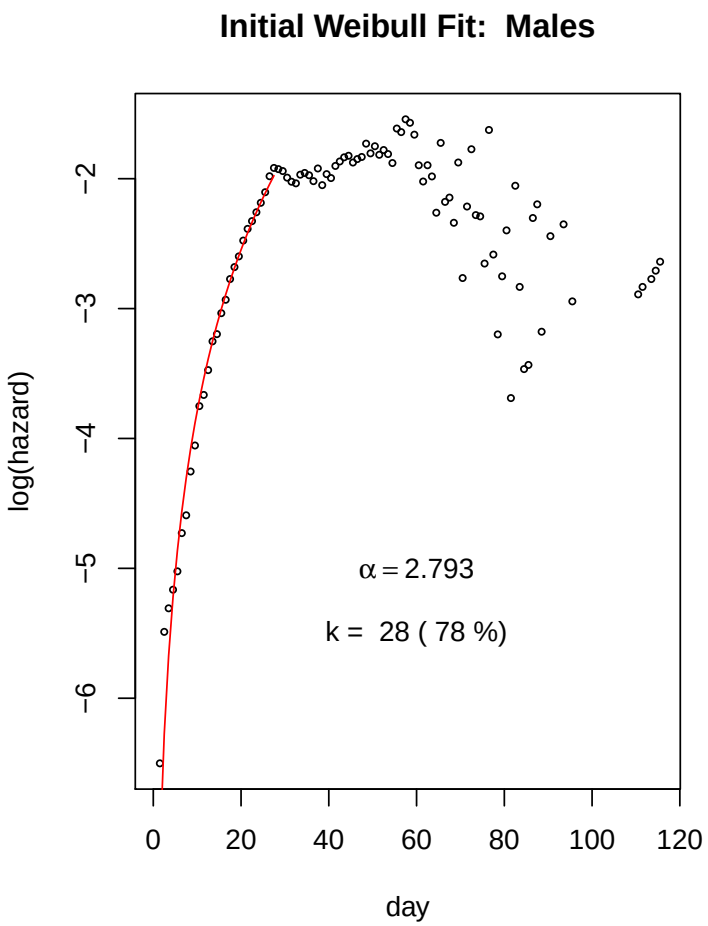
- Mortality (hazard) rates declined after age 60 days,
- Extremely long right tail with some flies living until age 200 days,
- Gender cross-over in mortality rates with males more frail after age 25.

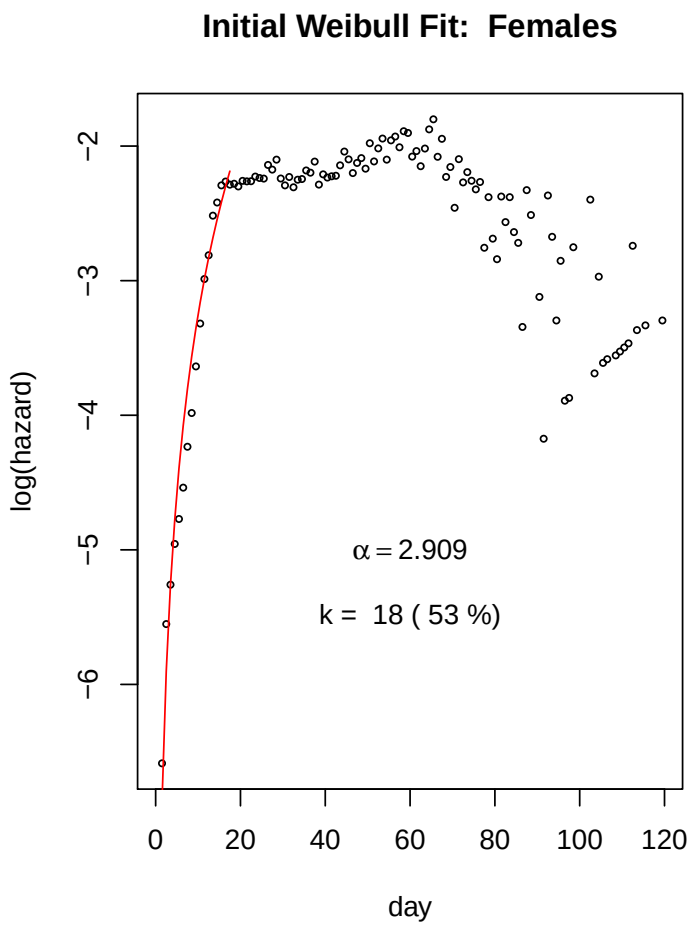
Weibull scale mixture model:

$$\max_{F \in \mathcal{F}} \{ \sum_{i=1}^{n} g(y_i) \mid g(y) = \int \varphi(y, \theta) dF(\theta)$$

Weibull shape parameter poses an interesting profile likelihood problem.

Gender Specific Baseline Weibull Model Estimation





NPMLE Gender Specific Estimated Hazard Rates

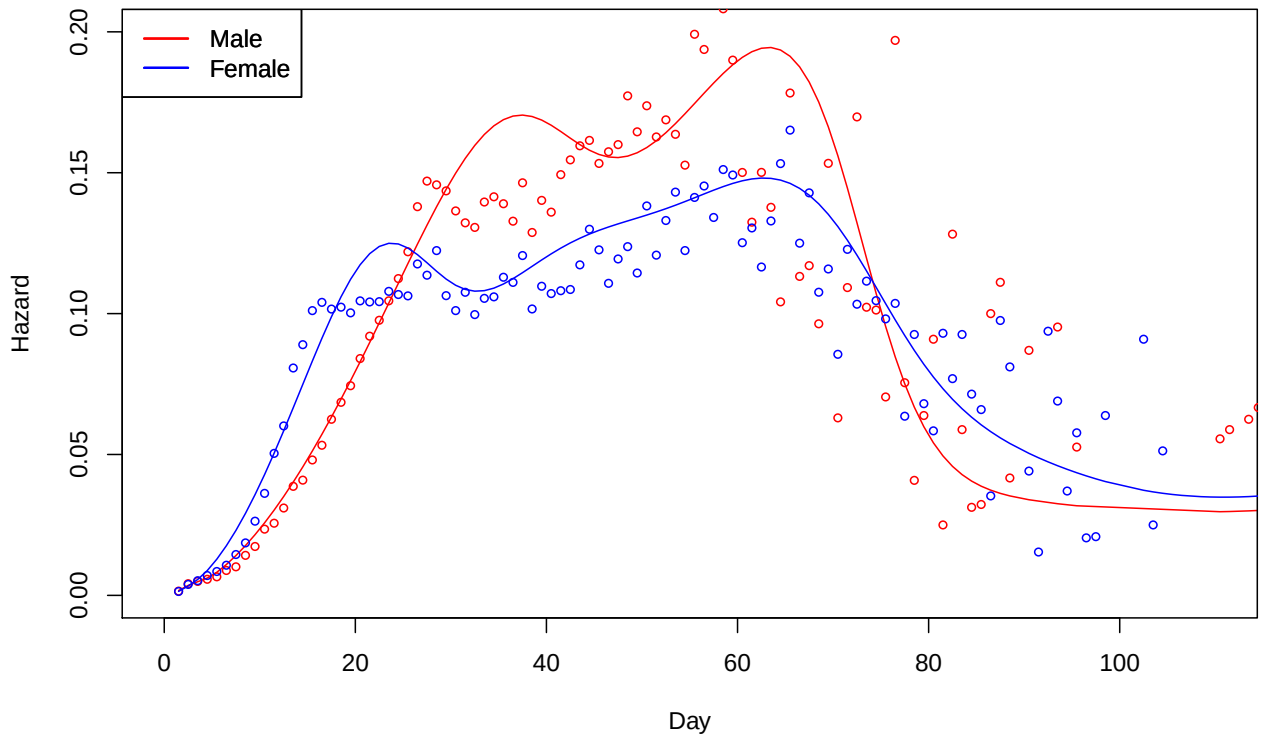


Figure: Gender Specific Hazard Functions for the Weibull Mixture Model: Raw daily mortality rates are plotted in red for males and blue for females, superimposed are the estimated hazard functions for the Weibull mixture models.

Random Coefficient Binary Response

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,

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

The random coefficients β_i are drawn independently of x_i and w_i and iidly from a distribution F_0. We will assume that $x_i = (1, z_i^\top, -v_i)^\top$ and will need to normalize β_i since it is only identified up to scale. It is convenient to normalize by setting one coefficient equal to one. Our objective is to estimate the pair (θ_0, F_0).

Random Coefficient Binary Response

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,

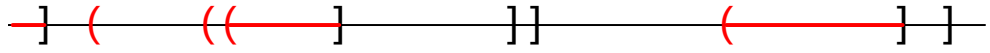
$$y_i = 1(x_i^\top \beta_i + w_i \theta_0 > 0).$$

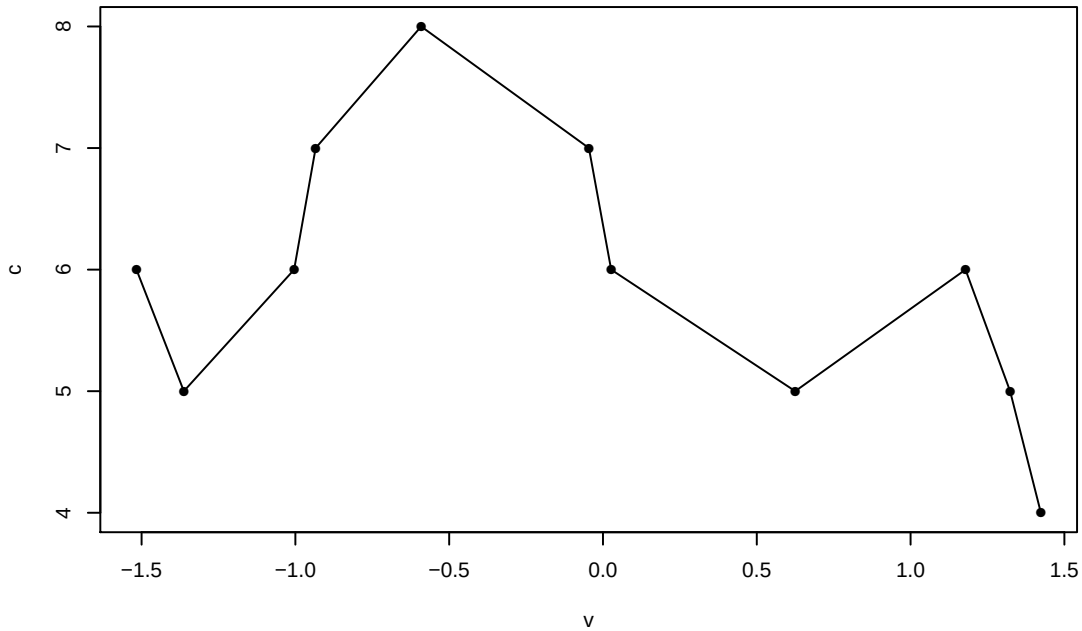
The random coefficients β_i are drawn independently of x_i and w_i and iidly from a distribution F_0. We will assume that $x_i = (1, z_i^\top, -v_i)^\top$ and will need to normalize β_i since it is only identified up to scale. It is convenient to normalize by setting one coefficient equal to one. Our objective is to estimate the pair (θ_0, F_0). The simplest setting is univariate with no z_i or w_i and $\beta_i = (\eta_i, 1)^\top$, so,

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

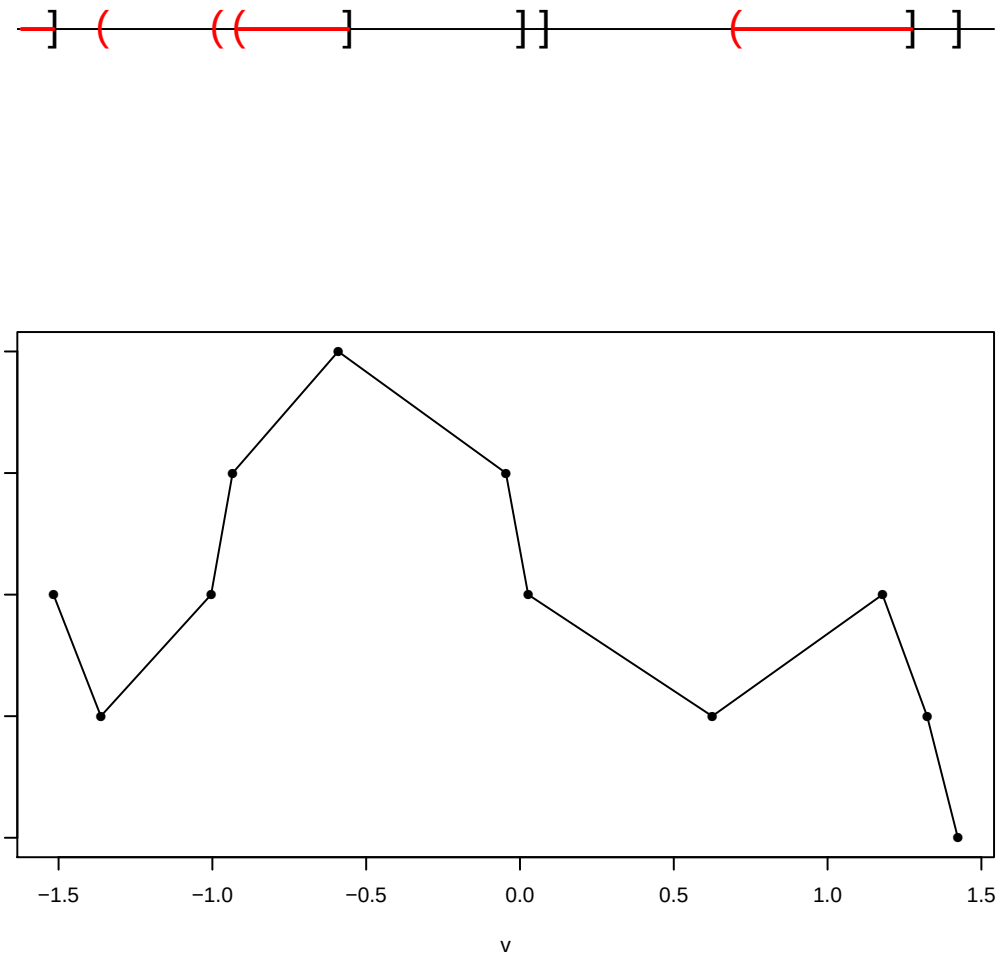
In econometrics this is called the Cosslett (1983) model. In biostatistics it is called the current status model, and has been studied extensively notably by Groeneboom and Jongbloed. The v_i's are inspection times and we observe only the binary indicator of the onset of a disease.

Only Locally Maximal Intervals “Count”





Only Locally Maximal Intervals “Count”



Theorem Only intervals with locally maximal intersections contain points that may have positive mass in the NPMLE solution. Within the optimizing intervals how mass is allocated is arbitrary.

Nonparametric Maximum Likelihood for Bivariate F_η

When the random parameter η is bivariate we have half spaces instead of half lines and polygons instead of intervals so arrangements become more complicated. Our binary response is generated as,

$$\mathbb{P}(y_i = 1|z_i, v_i) = \mathbb{P}(\eta_{1i} + z_i\eta_{2i} + v_i \geqslant 0).$$

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 F_η, so the log likelihood is,

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

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 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? The question isn't quite well posed, however unless the hyperplanes are in “general position” so any subset of size $k \leqslant d$ has normals that are linearly independent, 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 elaborated by Zaslavsky (1975).

In $\mathbb{R}^2$ How Many Interesting Polygons Are There?

When the hyperplanes are lines in $\mathbb{R}^2$ in general position, there are $\binom{n}{2} + n + 1 = \mathcal{O}(n^2)$ polygons. But only locally maximal polygons are interesting; how do we find those? The naive answer is that we count the number of half-space intersections for each polygon and ignore any that have neighbours with larger counts. This works great up to about $n = 10$.

- Suppose all the lines are oriented in the same way so above is 1 and below is 0,
- Then we can just count intersections of the halfspaces for each polygon,
- For our binary responses, we just flip the sign of the coefficients for the $y_i = 0$ lines,
- Given the counts for each polygon, we delete polygons that are not locally maximal,
- Polygons are then represented by any interior point.

Manski's Maximum Score Estimator: A Digression

The maximum score estimator is looking for the globally maximal polygon:

- So maximum score is a bounds estimator *avant la lettre*,
- Each locally maximal polygon constitutes a region in which the search for a global maximum may become marooned,
- The piecewise constant function of counts on the polygons can be viewed as a likelihood surface, and therefore serves as a "sort-of, kind-of" posterior for the parameter F_η,
- The NPMLE for F_0 assigns positive mass to some of these locally maximal polygons, but not all, so in Bayesian terminology it is a (somewhat curious) MAP estimator.

Incremental Cell Enumeration

Rada and Černý (2018), refining prior proposals of Avis and Fukuda (1996), have proposed an algorithm for cell (polytope) enumeration with running time proportional to the number of cells. Given a hyperplane arrangement, $\mathcal{H} = \{H_1, \cdots, H_n\}$ it proceeds by adding one hyperplane at a time, finding the newly created cells and their associated interior points.

- Given any cell we can associate a sign vector, $s_k \in \mathbb{R}^n$ that reveals whether each of the n hyperplanes lie above or below interior points of the cell. Let S be an n by M matrix with columns composed of these vectors.
- The algorithm proceeds by sequentially updating these sign vectors as hyperplanes are added.
- To verify the validity of the new sign vectors and find an interior point for each new cell we need to solve a linear program.

Keeping up with the Joneses

Given the matrix of sign vectors, S, for the full sample as produced by the AIE algorithm, it is easy to determine neighbours for each cell.

- Cells C_j and C_k are neighbours if and only if their sign vectors differ in exactly one coordinate.
- Define cell counts, for each cell by replacing all -1's in their sign vector by 0's, and summing.
- Eliminate from consideration any cells with neighbours that exceed their own cell counts
- This typically reduces the number of candidate cells from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ in bivariate problems.
- The interior points of the remaining cells constitute potential support points of the NPMLE.

The NPMLE: 3 Equivalent Versions

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

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

Given locally maximal cells, $\{C_i, \cdots, C_{M^*}\}$, define a n by M^* matrix A with $A_{ij} = 1\{C_j \subset G(z_i, v_i, \theta)\}$ if y_i and $1 - 1\{C_j \subset G(z_i, v_i, \theta)\}$ 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 (1948) like, as in Kiefer and Wolfowitz (1956) and recently elaborated in a review paper by Chen (2017).

Some Simulation Comparisons

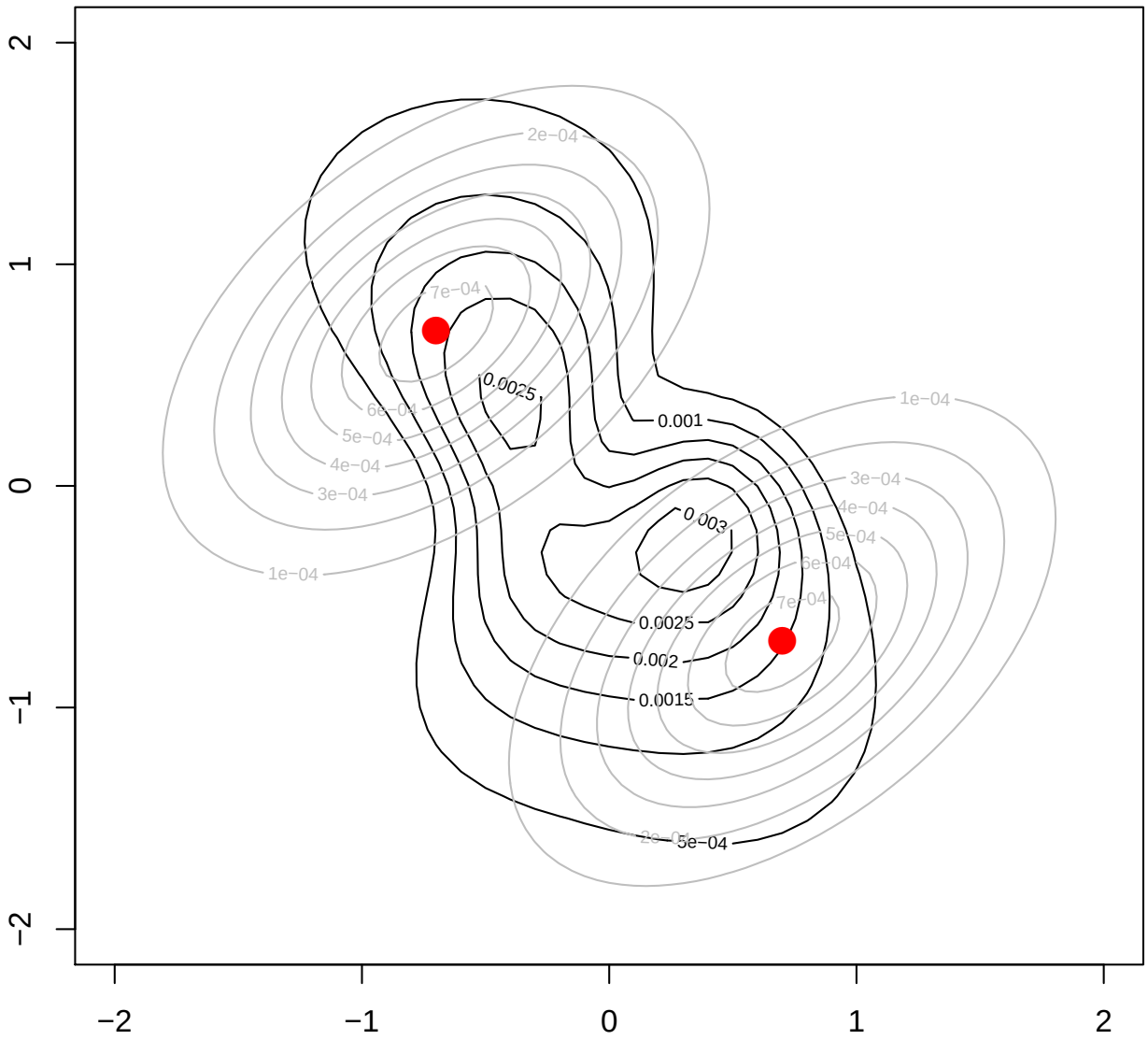
Gautier and Kitamura (2013) have proposed an elegant Fourier-Laplace deconvolution approach to estimation of F_η for the bivariate problem. We will compare performance of our NPMLE approach to theirs.

We adopt the same simulation setup used by Gautier and Kitamura: Data is generated with (z_i, ν_i)'s drawn iidly from the standard bivariate Gaussian distribution, η_i drawn either from:

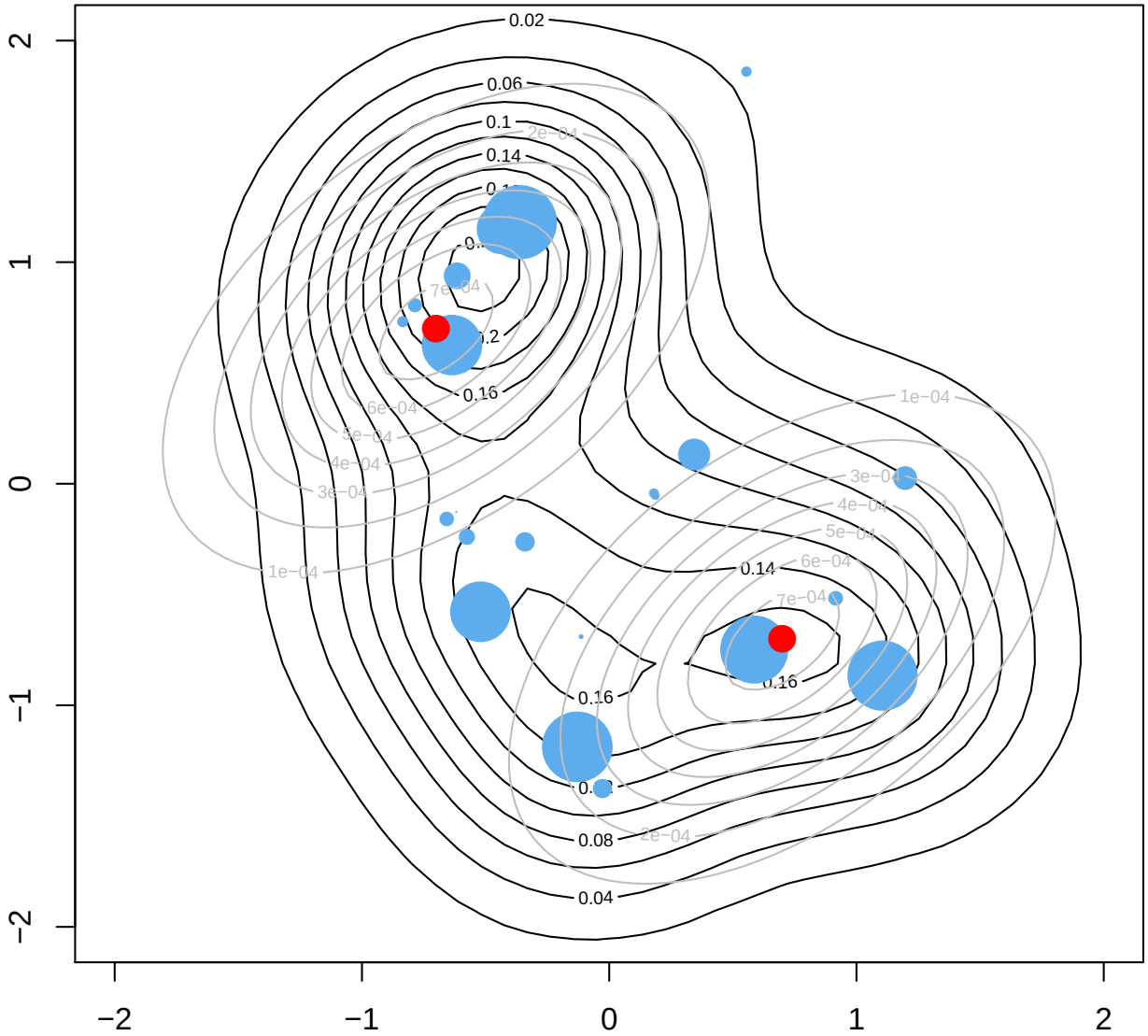
- With equal probability from the two points, $\{(0.7, -0.7), (-0.7, 0.7)\}$, or
- From bimodal correlated Gaussians with separated modes at these points,

Gautier-Kitamura Contours for their Bimodal DGP

Default Tuning Parameters, n = 500



NPMLE for the Gautier-Kitamura Bimodal DGP



One Picture is Worth 1000 Simulations

It is dangerous to infer too much from a single realization so we ran a small scale (500 replications) simulation experiment to evaluate out-of-sample predictive performance of the methods. Sample size $n = 500$

	GK	NPMLE	NPMLEs	Logit
MAE	0.1333	0.0868	0.1274	0.1753
RMSE	0.1705	0.1576	0.1726	0.2150

Table: Bivariate Point Mass Simulation Setting: Mean Absolute and Root Mean Squared Errors of Predicted Probabilities

	GK	NPMLE	NPMLEs	Logit
MAE	0.1288	0.0592	0.0475	0.0709
RMSE	0.1440	0.0748	0.0594	0.0896

Table: Bivariate Gaussian Simulation Setting: Mean Absolute and Root Mean Squared Errors of Predicted Probabilities

Some References

GU, J., AND R. KOENKER (2016): "On a Problem of Robbins," *International Statistical Review*, 84, 224–244.

——— (2018): *RCBR: An R Package for Binary Response with Random Coefficients.* Coming soon to your favorite CRAN mirror.

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 (2013): "Frailty, Profile Likelihood and Medfly Mortality," in *Contemporary Developments in Statistical Theory: A Festschrift for Hira Lal Koul*, ed. by S. Lahiri, A. Schick, A. Sengupta, and T. Sriram. Springer.

KOENKER, R., AND J. GU (2017): "REBayes: An R Package for Empirical Bayes Mixture Methods," *Journal of Statistical Software*, 82, 1–26.

KOENKER, R., AND I. MIZERA (2014): "Convex Optimization, Shape Constraints, Compound Decisions and Empirical Bayes Rules," *J. of Am. Stat. Assoc.*, 109, 674–685.

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