Empirical Likelihood

Art B. Owen

Department of Statistics

Stanford University

David Arthur Sprott

Photo: Statistical Society of Canada

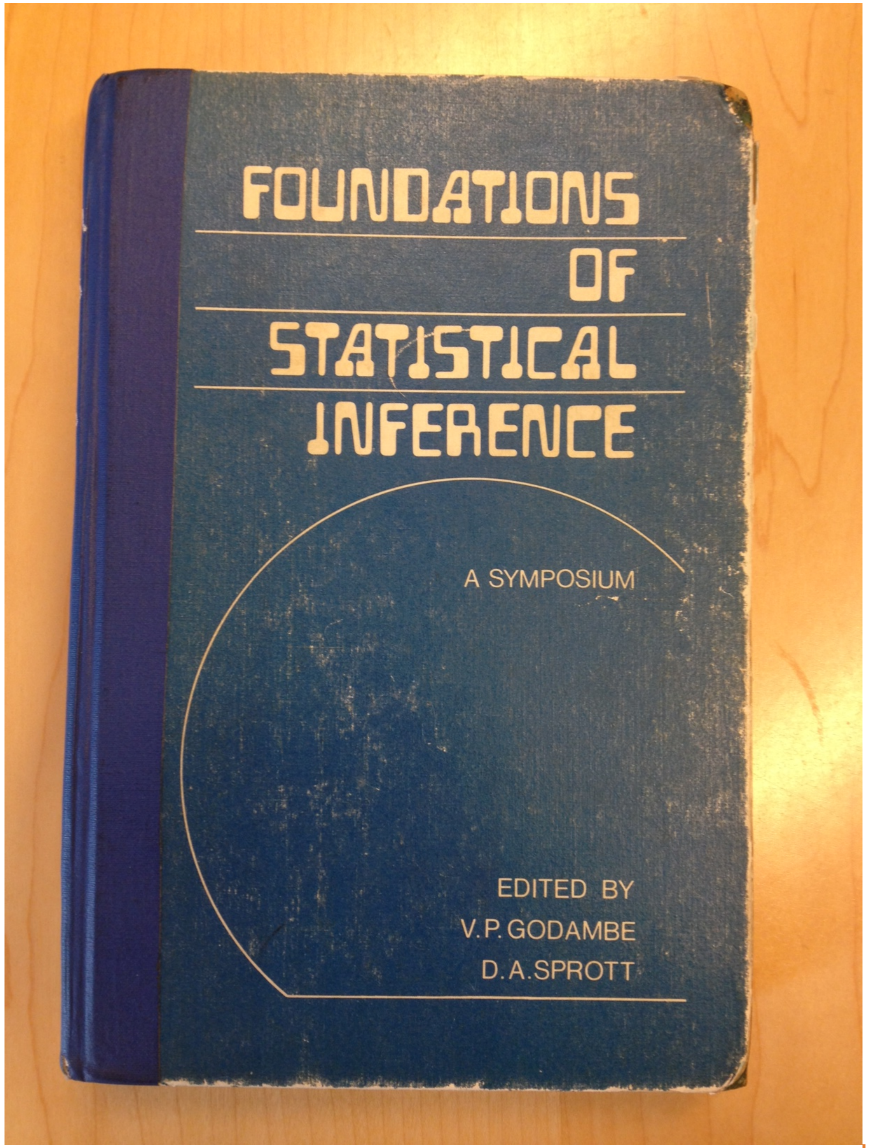
- Born Toronto 1930
- BA 1952, MA 1953, PhD 1955
  University of Toronto
- Founding Chair:
  Statistics & Actuarial Science
- Founding Dean:
  Faculty of Mathematics
- Likelihood researcher and advocate
- Teacher

David Sprott as a teacher

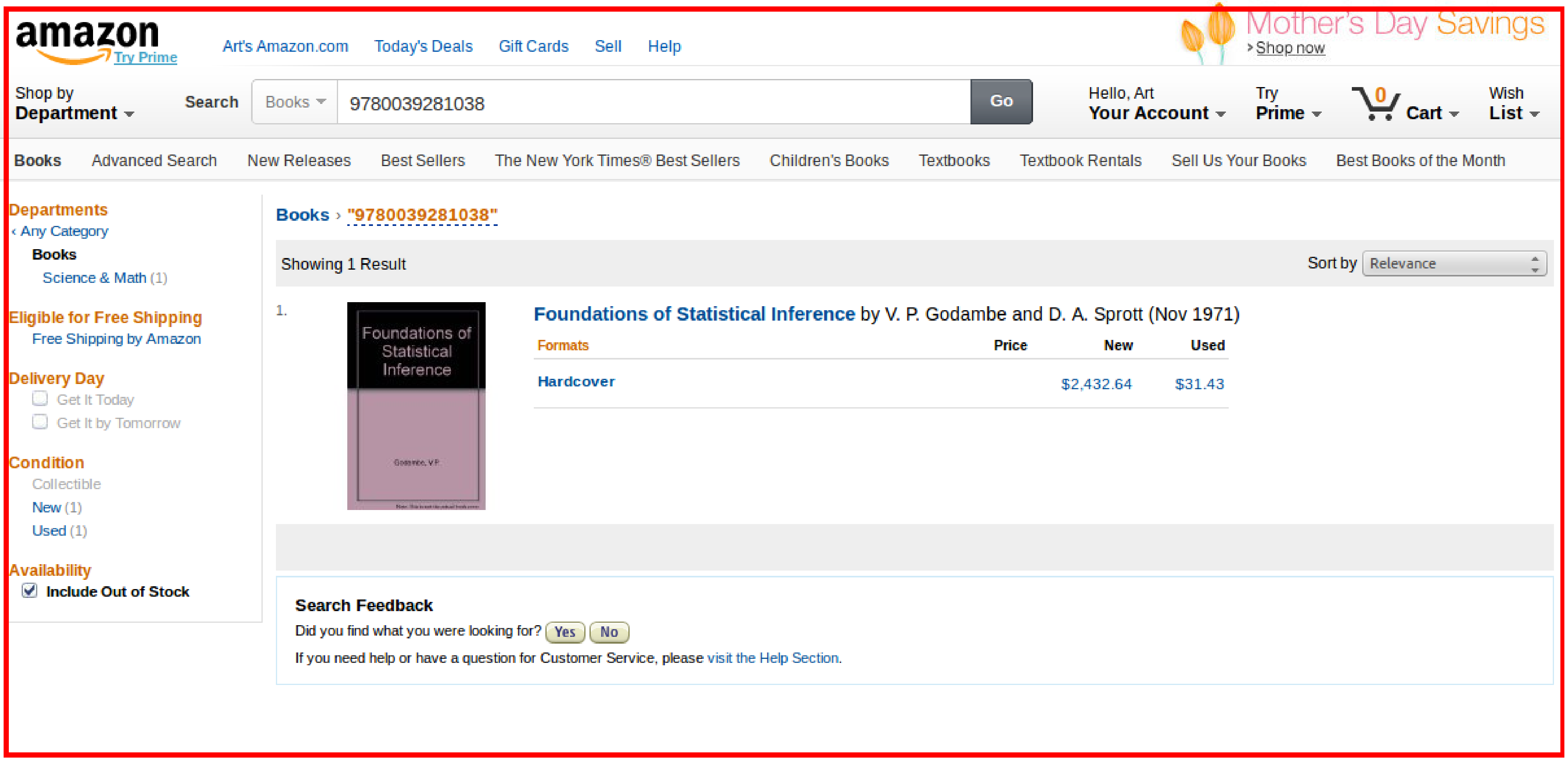
David Sprott did not make things easy for students. Instead he would stump us with really hard counter-intuitive puzzles. For instance one problem had us conditioning on an event of probability zero and getting contradictory answers. Those lessons stay with you.

Recently there has been much anguish about published findings that do not replicate. I don't think this outcome would have surprised him.

Foundations book 1971



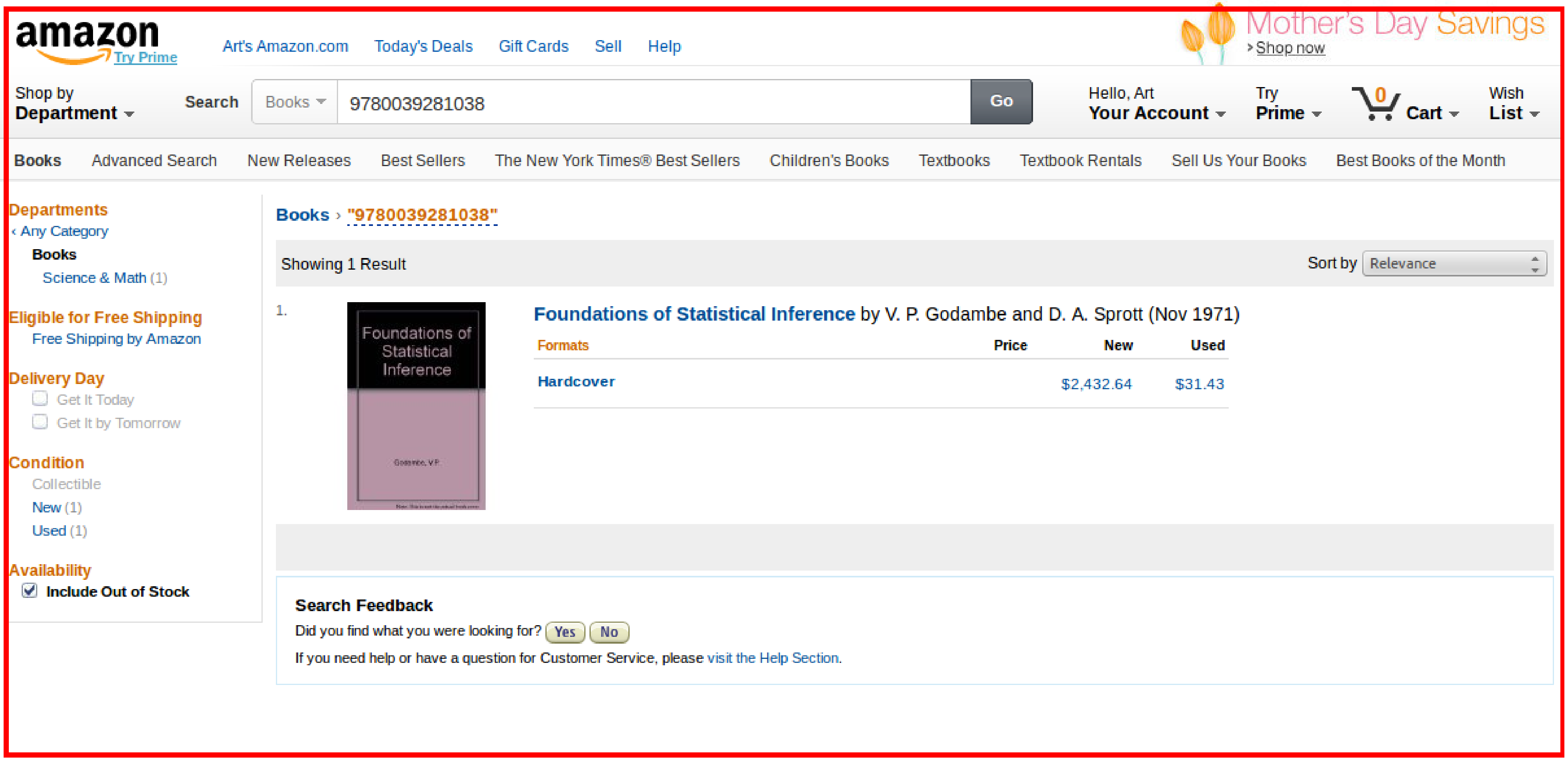
Foundations book 2014



Ad Feedback

Search powered by A9

Foundations book 2014



(Plus $3.99 shipping)

Ad Feedback

Search powered by A9

Waterloo to Stanford

At Waterloo I learned an approach to statistics that was based on thinking hard about what the problem meant $\cdots$ so that you could come up with the right likelihood.

When I got to Stanford, the emphasis was on doing everything nonparametrically. Use the computer instead of strong assumptions.

Empirical likelihood fits both. The spark was an exercise (#6 in Appendix 2) in the text book by Kalbfleisch and Prentice (1980), which points to Thomas and Grunkemeier (1979).

It ultimately ties back to estimating equations: Godambe & Thompson and Qin & Lawless.

Empirical likelihood provides:

- **likelihood** methods for inference, especially
  - tests, and
  - confidence regions,
- **without** assuming a parametric model for data
- **competitive** power even when parametric model holds

Like the bootstrap, but no resampling, and it picks the shape of confidence regions.

Parametric likelihoods

Data $X_1, X_2, \ldots, X_n$ have **known** distribution f_θ with **unknown** parameter θ

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) = f(x_1, \ldots, x_n; \theta)$$

For continuous data $\cdots$ use probability density function.

$f(\cdots;\cdot)$ known, $\quad \theta \in \Theta \subseteq \mathbb{R}^p$ unknown

Likelihood function

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta)$$

“Chance, under θ, of getting the data we did get”

Likelihood examples

$$X_i \sim \text{Poi}(\theta), \quad \theta \geq 0$$

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

A continuous example

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \quad x_i \text{ fixed}$$

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}$$

Likelihood inference

Maximum likelihood estimate

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x_1, \ldots, x_n)$$

Likelihood ratio inferences

$$-2\log(L(\theta_0)/L(\hat{\theta})) \to \chi^2_{(q)} \qquad \text{Wilks}$$

1) Reject $H_0 : \theta = \theta_0$ if

$$\frac{L(\theta_0)}{L(\hat{\theta})} < \exp\left(-\frac{1}{2}\chi^{2,1-\alpha}_{(q)}\right)$$

2) Confidence set for θ_0

$$\left\{\theta \mid \frac{L(\theta)}{L(\hat{\theta})} \geq \exp\left(-\frac{1}{2}\chi^{2,1-\alpha}_{(q)}\right)\right\}$$ e.g. 95% confidence if $\alpha = .05$

Statistical advantages

Typically $\cdots$ Neyman-Pearson, Cramer-Rao, . . .

1) $\hat{\theta}$ asymptotically normal

2) $\hat{\theta}$ asymptotically efficient

3) Likelihood ratio tests powerful

4) Likelihood ratio confidence regions small

A disadvantage

Problems with many parameters:

See Kalbfleisch & Sprott (1970) JRSS-B (with discussion)

Other likelihood advantages

- can model/undo data distortion: bias, censoring, truncation
- can combine data from different sources
- can factor in prior information
- obey range constraints: MLE of correlation in $[-1, 1]$
- transformation invariance
- data determined shape for $\{\theta \mid L(\theta) \geq rL(\hat{\theta})\}$

Unfortunately

We might not know a correct $f(\cdots;\theta)$

No reason to expect that new data belong to one of our favourite families

Wrong models sometimes work (e.g. Normal mean via CLT) and sometimes fail (e.g. Normal variance)

Nonparametric methods

Assume only $X_i \sim F$ where

- F is continuous, or,
- F is symmetric, or,
- F has a monotone density, or,
- F has log-concave density, or,
- $\cdots$ other believable, but big, family

Nonparametric usually means infinite dimensional parameter

Sometimes lose power (e.g. sign test), sometimes not

Nonparametric maximum likelihood

For $X_i \overset{\text{iid}}{\sim} F, \quad L(F) = \prod_{i=1}^{n} F(\{x_i\})$

The NPMLE is $\quad \widehat{F} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$

where δ_x is a point mass at x

Kiefer and Wolfowitz, 1956

Easy proof based on $\log(1+z) \leq z$

Other NPMLEs

NPMLEs are useful when we want the analogue of the empirical CDF for nonstandard settings.

Kaplan & Meier (1958)	Right censored survival times
Lynden-Bell (1971)	Left truncated star brightness
Hartley & Rao (1968)	Sample survey data
Grenander (1956)	Monotone density for actuarial data

Censoring and Truncation

The likelihood can be used to compensate for sampling distortions.

Censoring

All we know is that $X_i \in C_i$. For a patient that survived at least ≥ 438 days, $X_i \in [438, \infty]$.

If observed exactly, then $C_i = \{X_i\}$. Conditional on C_i

$$L(F) = \prod_{i=1}^{n} F(C_i)$$

Truncation

X_i only observed if $X_i \in T_i$. E.g.: star only seen if it is bright enough.

$$L(F) = \prod_{i=1}^{n} \frac{F(\{X_i\})}{F(T_i)} \quad \text{or} \quad \prod_{i=1}^{n} \frac{F(C_i \cap T_i)}{F(T_i)}$$

Monotone & unimodal

Grenander (1956) $X \in [0, \infty)$ density f non-decreasing NPMLE $\hat{F}$ is 'least concave majorant of the ECDF'

piece-wise linear density

Log concave

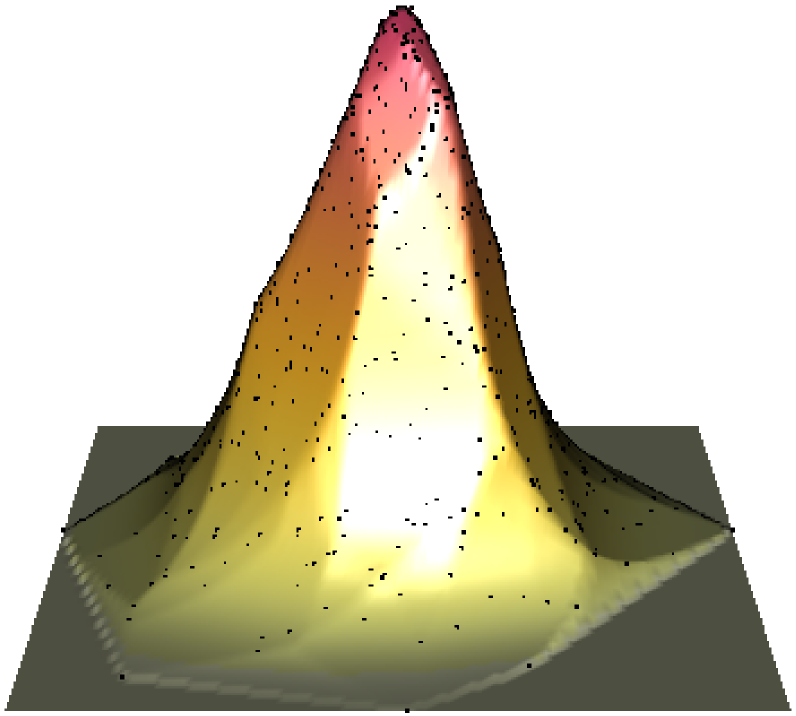
Recent work Samworth, Cule, Walther, Dumbgen $\cdots$

$\log f(\boldsymbol{x})$ concave on $\mathbb{R}^d$

MLE computable for small d

No bandwidth to select

A log concave MLE



Downloaded January 2014 from

`http://www.statslab.cam.ac.uk/Statistics/activities/CSI_RS2.png`

Empirical likelihood (short story)

Let $w_i = F(\{X_i\})$ the probability under F of getting **exactly** X_i.

We assume[1] that $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$, then

$$L(F) = \prod_{i=1}^n w_i \qquad \text{Likelihood}$$

$$L(\hat{F}) = \prod_{i=1}^n (1/n) \qquad \text{Maximized likelihood}$$

$$R(F) = \prod_{i=1}^n n w_i \qquad \text{Empirical likelihood ratio}$$

[1] A longer story explains these choices

Empirical likelihood for the mean

Confidence region is

$$C_{r,n} = \Big\{ \sum_{i=1}^{n} w_i \boldsymbol{x}_i \mid w_i > 0, \sum_{i=1}^{n} w_i = 1, \prod_{i=1}^{n} n w_i \ge r \Big\}$$

Profile likelihood

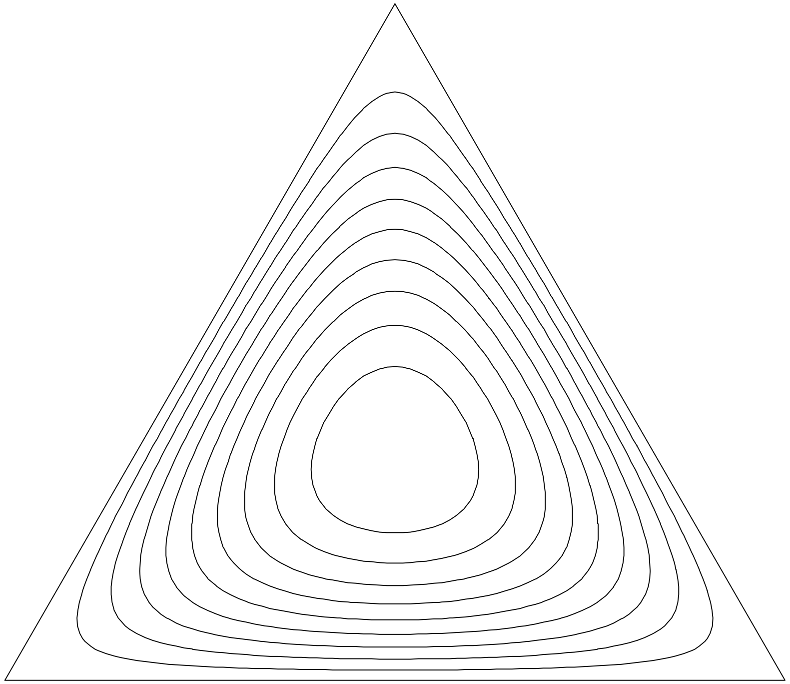
$$\mathcal{R}(\mu) = \sup \Big\{ \prod_{i=1}^{n} n w_i \mid w_i > 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i \boldsymbol{x}_i = \mu \Big\}$$

$$C_{r,n} = \{ \mu \mid \mathcal{R}(\mu) \ge r \}$$

Multinomial

We have a multinomial on the n data points X_i, hence $n-1$ parameters

Multinomial likelihood for $n = 3$



Contours of $\prod_i nw_i$ MLE at center LR$= i/10, i = 0, \ldots, 9$

Empirical likelihood theorem

Suppose that $\boldsymbol{X}_i \sim F_0$ are IID in $\mathbb{R}^d$

$\mu_0 = \int \boldsymbol{x} dF_0(\boldsymbol{x})$

$V_0 = \int (\boldsymbol{x} - \mu_0)(\boldsymbol{x} - \mu_0)^T dF_0(\boldsymbol{x})$ finite

rank$(V_0) = q > 0$

Then as $n \to \infty$

$$-2 \log \mathcal{R}(\mu_0) \to \chi^2_{(q)}$$

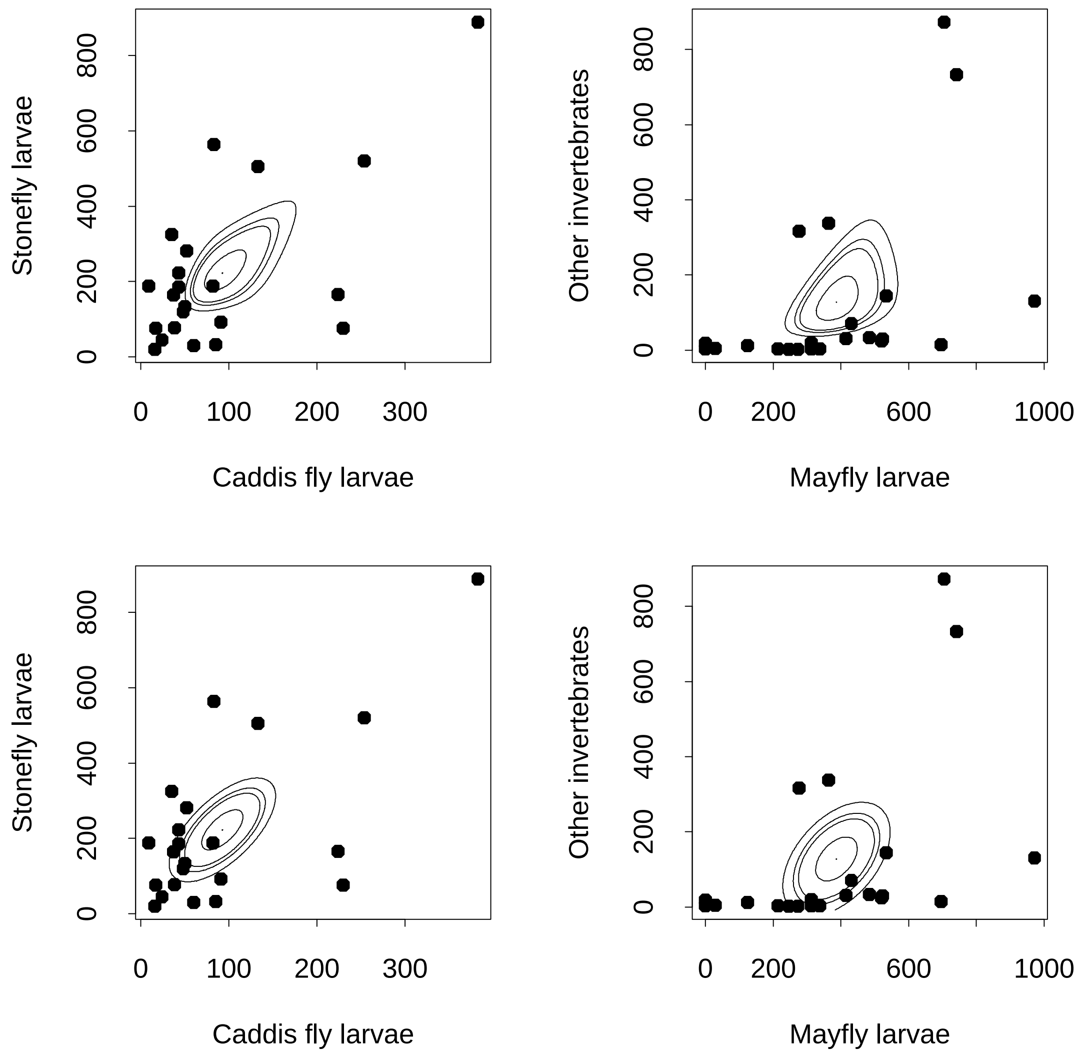
same as parametric limit

No apparent penalty for using $n-1$ parameters.

Dipper, Cinclus cinclus

Eats larvae of Mayflies, Stoneflies, Caddis flies, other

Dipper diet means



Top row shows EL; bottom Hotelling's T^2 ellipses

Data from Iles

Computing EL for the mean

Start with the convex hull:

$$\mathcal{H} = \mathcal{H}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \Bigl\{ \sum_{i=1}^n w_i \boldsymbol{x}_i \mid w_i \ge 0, \sum_{i=1}^n w_i = 1 \Bigr\}$$

$\mu \notin \mathcal{H} \implies \log \mathcal{R}(\mu) = -\infty$

If $\mu \in \mathcal{H}$ then $\mathcal{R}(\mu) < \infty$

and we will compute it via Lagrange multipliers

Lagrange multipliers

$$G = \sum_{i=1}^{n} \log(nw_i) - n\lambda^{\mathsf{T}}\Big(\sum_{i=1}^{n} w_i(\boldsymbol{x}_i - \mu)\Big) + \gamma\Big(\sum_{i=1}^{n} w_i - 1\Big)$$

$$\frac{\partial}{\partial w_i} G = \frac{1}{w_i} - n\lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu) + \gamma = 0$$

$$\sum_i w_i \frac{\partial}{\partial w_i} G = n + \gamma = 0 \quad \Longrightarrow \quad \gamma = -n$$

Solving,

$$w_i = \frac{1}{n}\frac{1}{1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)}$$

Where $\lambda = \lambda(\mu)$ solves

$$0 = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i - \mu}{1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)}$$

reciprocal tilting

Convex duality

$$\text{Let} \quad \mathbb{L}(\lambda) \equiv -\sum_{i=1}^{n} \log(1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)) = \log R(F)$$

$$\frac{\partial \mathbb{L}}{\partial \lambda} = -\sum_{i=1}^{n} \frac{\boldsymbol{x}_i - \mu}{1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu)}$$

Minimizing $\mathbb{L}$ sets gradient to 0 and maximizes $\log R$

$$\frac{\partial^2 \mathbb{L}}{\partial \lambda \partial \lambda^{\mathsf{T}}} = \sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \mu)(\boldsymbol{x}_i - \mu)^{\mathsf{T}}}{(1 + \lambda^{\mathsf{T}}(\boldsymbol{x}_i - \mu))^2}$$

$\mathbb{L}$ is convex and d dimensional $\implies$ easy optimization

Recently: self-concordant convex version O (2013)

Why χ^2?

$$-2\log(\mathcal{R}(\mu)) \doteq n(\bar{\boldsymbol{x}} - \mu_0)^{\mathsf{T}} S^{-1}(\bar{\boldsymbol{x}} - \mu_0)$$

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \mu)(\boldsymbol{x}_i - \mu)^{\mathsf{T}}$$

Taylor expansion plus central limit theorem

Hall shows that the shape of the EL confidence regions is a meaningful improvement over the ellipsoids from Hotelling's T^2

Coverage errors

1) $\Pr(\mu_0 \in C_{r,n}) = 1 - \alpha + O\big(\frac{1}{n}\big)$ as $n \to \infty$ Hall

2) One-sided errors of $O\big(\frac{1}{\sqrt{n}}\big)$ cancel

3) Bartlett correction DiCiccio, Hall, Romano
Replace $\chi^{2,1-\alpha}$ by $\big(1 + \frac{a}{n}\big)\chi^{2,1-\alpha}$ for carefully chosen a
and get coverage errors $O\big(\frac{1}{n^2}\big)$

same as for parametric likelihoods

Power

Suppose $X_i \in \mathbb{R}$ with $\mathbb{E}(X) = \mu$ and Var$(X) = \sigma^2 > 0$. Then

$$-2\log(\mathcal{R}(\mu_0 + \tau\sigma_0 n^{-1/2})) \to \chi^2_{(1)}(\tau^2)$$

noncentral χ^2 and so

$$\text{power } = \Pr(\chi^2_{(1)}(\tau^2) \geq \chi^{2,1-\alpha}_{(1)}),$$

same as in parametric setting

Finer print

When a parametric model holds, we may use it to generate an MLE of $\hat{\theta}$. EL inferences for that estimate are also as efficient as ones based on parametric likelihood, to a second order analysis in Lazar and Mykland (1998)

Calibrating empirical likelihood

Plain $\chi^{2,1-\alpha}$	undercovers
$F_{d,n-d}^{1-\alpha}$	is a bit better
Bartlett correction	asymptotics slow to take hold
Bootstrap	seems to work best

Bootstrap calibration

Resample the data to estimate the distribution of $-2\log\mathcal{R}(\mu_0)$ by that of $-2\log\mathcal{R}^*(\bar{\boldsymbol{x}})$

Results

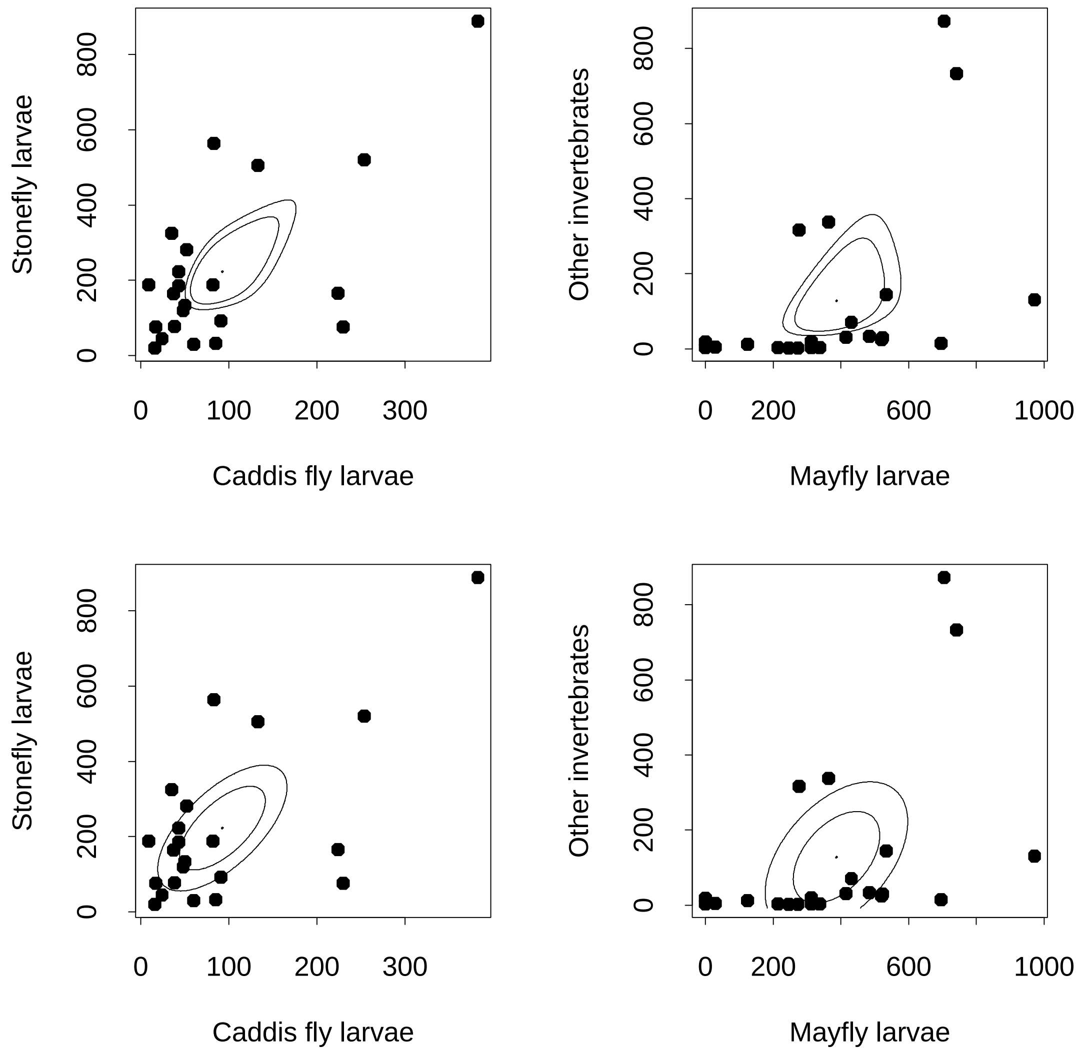
Coverage error $O(n^{-2})$

Same error rate as bootstrapping the bootstrap

Sets in faster than Bartlett correction

Need further adjustments for one-sided inference

Bootstrap (and χ^2) calibrated Dipper regions



Euclidean log likelihood

$-\sum_{i=1}^n \log(nw_i)$ is a 'distance' of w from $(1/n, \ldots, 1/n)$.

Replace loglik by

$$\ell_E = -\frac{1}{2}\sum_{i=1}^n (nw_i - 1)^2$$

Then $-2\ell_E \to \chi^2_{(q)}$ too

Reduces to Hotelling's T^2 for the mean O. (1990)

Reduces to Huber-White covariance for regression

Reduces to continuous updating GMM Kitamura

Quadratic approx to EL, like Wald test is to parametric likelihood

Exponential empirical likelihood

Replace $-\sum_{i=1}^{n}\log(nw_i)$ by

$$\text{KL} = \sum_{i=1}^{n} w_i \log(nw_i)$$

relates to entropy and exponential tilting

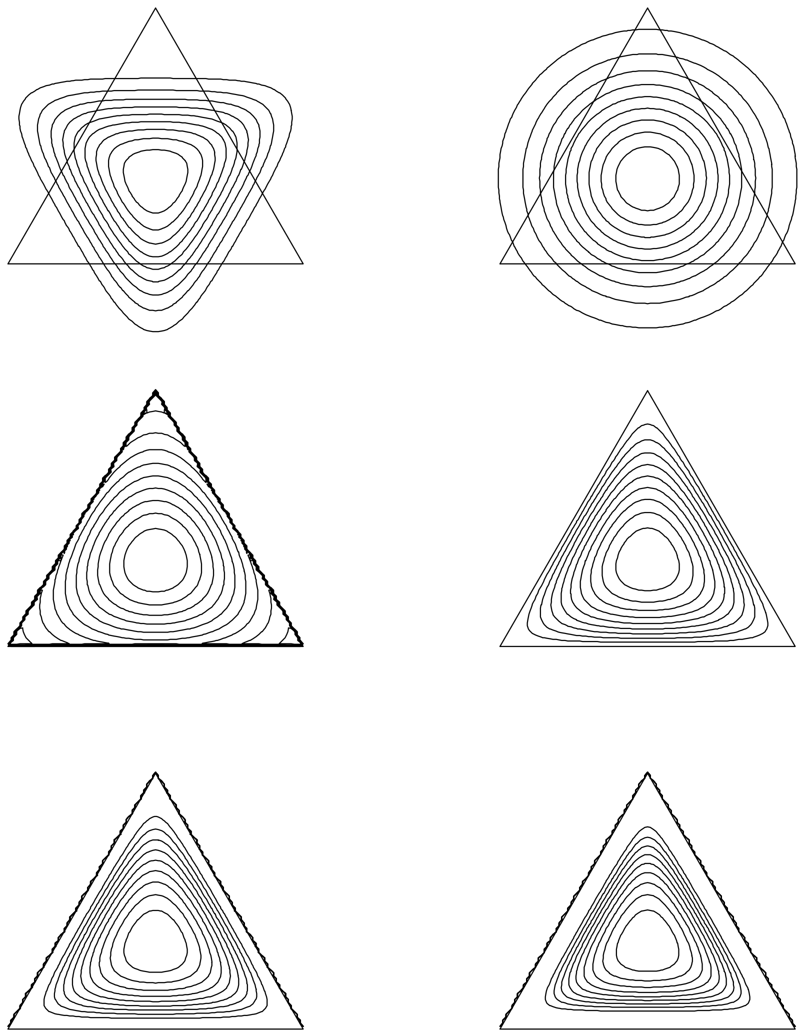
Hellinger distance

$$\sum_{i=1}^{n}(w_i^{1/2} - n^{-1/2})^2$$

Renyi, Cressie-Read

$$\frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{n}((nw_i)^{-\lambda} - 1)$$

Renyi-Cressie-Read contours



Top to bottom, left to right, λ: -5 -2 0 1 2/3 3/2

Estimating equations

More powerful and general than smooth functions

Define θ via $\mathbb{E}(m(\boldsymbol{X},\theta)) = 0$

Define $\hat{\theta}$ via $\frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i,\hat{\theta}) = 0$

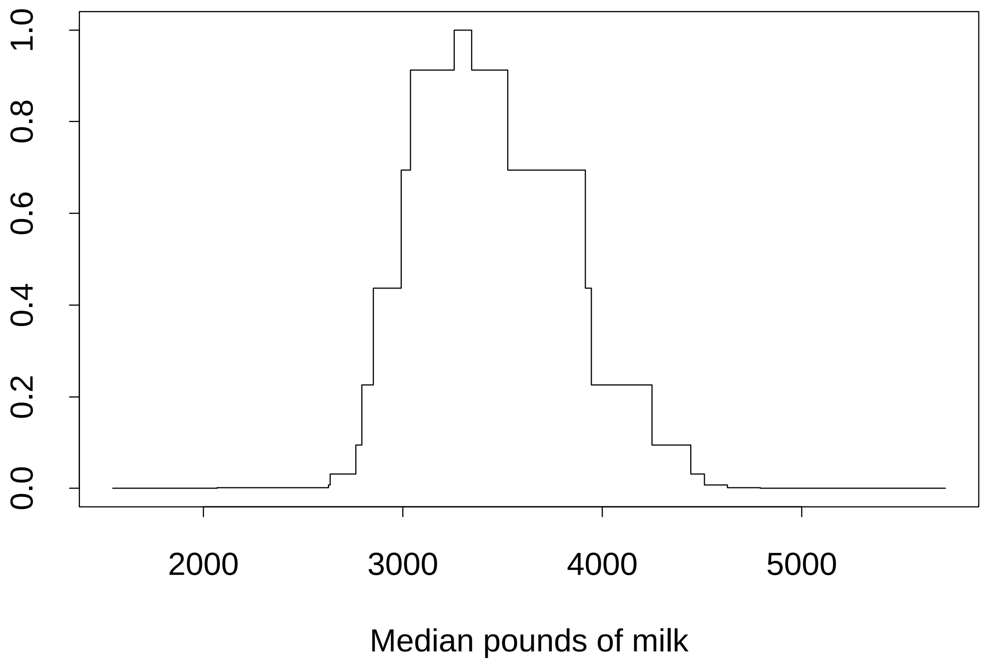
Usually $\dim(m) = \dim(\theta)$

Basic examples:

$m(\boldsymbol{X},\theta)$	Statistic
$\boldsymbol{X} - \theta$	Mean
$1_{\boldsymbol{X}\in A} - \theta$	Probability of set A
$1_{X\le\theta} - \frac{1}{2}$	Median
$\frac{\partial}{\partial\theta}\log(f(\boldsymbol{X};\theta))$	MLE under f

$$-2\log\mathcal{R}(\theta_0) \to \chi^2_{\text{Rank}(\text{Var}(m(\boldsymbol{X},\theta_0)))}$$

Empirical likelihood for a median



LR is constant between observations

$\mathbb{E}(1_{X \leq m} - 1/2) = 0$

α-quantile: $\mathbb{E}(1_{X \leq \theta} - \alpha) = 0$

Nuisance parameters

Sometimes we cannot write $\mathbb{E}(m(\boldsymbol{X}, \theta)) = 0$ directly, but can by introducing a few extra (nuisance) parameters,

$$\mathbb{E}(m(\boldsymbol{X}, \theta, \nu)) = 0$$

where θ is of interest and ν is the nuisance. IE, we expand the parameter vector from θ to (θ, ν).

Profile likelihood

$$\mathcal{R}(\theta, \nu) = \max\Bigl\{\prod_{i=1}^{n} nw_i \mid w_i \geq 0, \sum_{i=1}^{n} w_i, \sum_{i=1}^{n} w_i m(\boldsymbol{x}_i, \theta, \nu)\Bigr\}$$

$$\mathcal{R}(\theta) = \max_{\nu} \mathcal{R}(\theta, \nu) \equiv \text{profile empirical likelihood}$$

The first optimization is simple. The second may be difficult.

Typically $-2 \log \mathcal{R}(\theta_0) \to \chi^2_{(\dim(\theta))}$

Example: correlation

Suppose we are interested in $\rho = \mathrm{Corr}(X, Y)$. Then,

$$0 = \mathbb{E}(X - \mu_x)$$

$$0 = \mathbb{E}(Y - \mu_y)$$

$$0 = \mathbb{E}((X - \mu_x)^2 - \sigma_x^2)$$

$$0 = \mathbb{E}((Y - \mu_y)^2 - \sigma_y^2)$$

$$0 = \mathbb{E}((X - \mu_x)(Y - \mu_y) - \rho\sigma_x\sigma_y)$$

Parameter and nuisance

$\theta = (\rho)$ and $\nu = (\mu_x, \mu_y, \sigma_x, \sigma_y)$
$\mathbb{E}(m(\boldsymbol{X}, \theta, \nu)) = 0 = \frac{1}{n}\sum_{i=1}^{n} m(X_i, \hat{\theta}, \hat{\nu})$
$m(\cdot)$ has the five components above

Huber's robust M-estimate

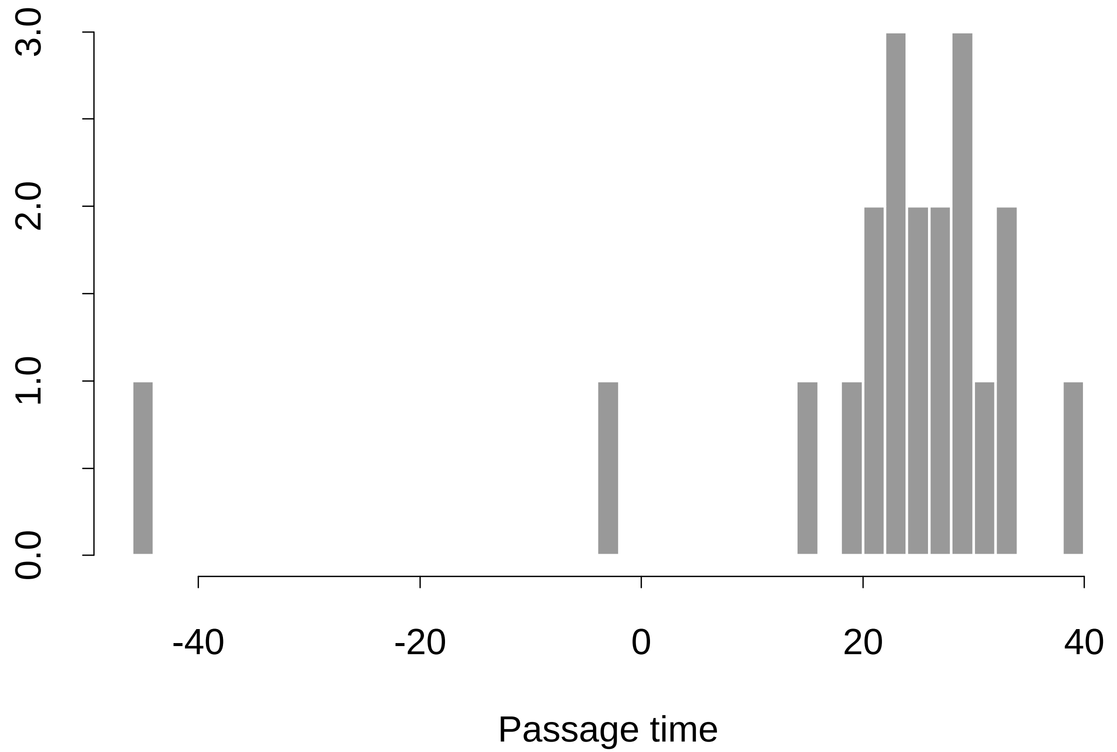
$$0 = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{x_i-\mu}{\sigma}\right) \qquad 0 = \frac{1}{n}\sum_{i=1}^{n}\left[\psi\left(\frac{x_i-\mu}{\sigma}\right)^2 - 1\right]$$

Like mean for small obs, median for outliers

$$\psi(z) = \begin{cases} z, & |z| \leq 1.35 \\ 1.35\,\mathsf{sign}(z), & |z| \geq 1.35. \end{cases}$$

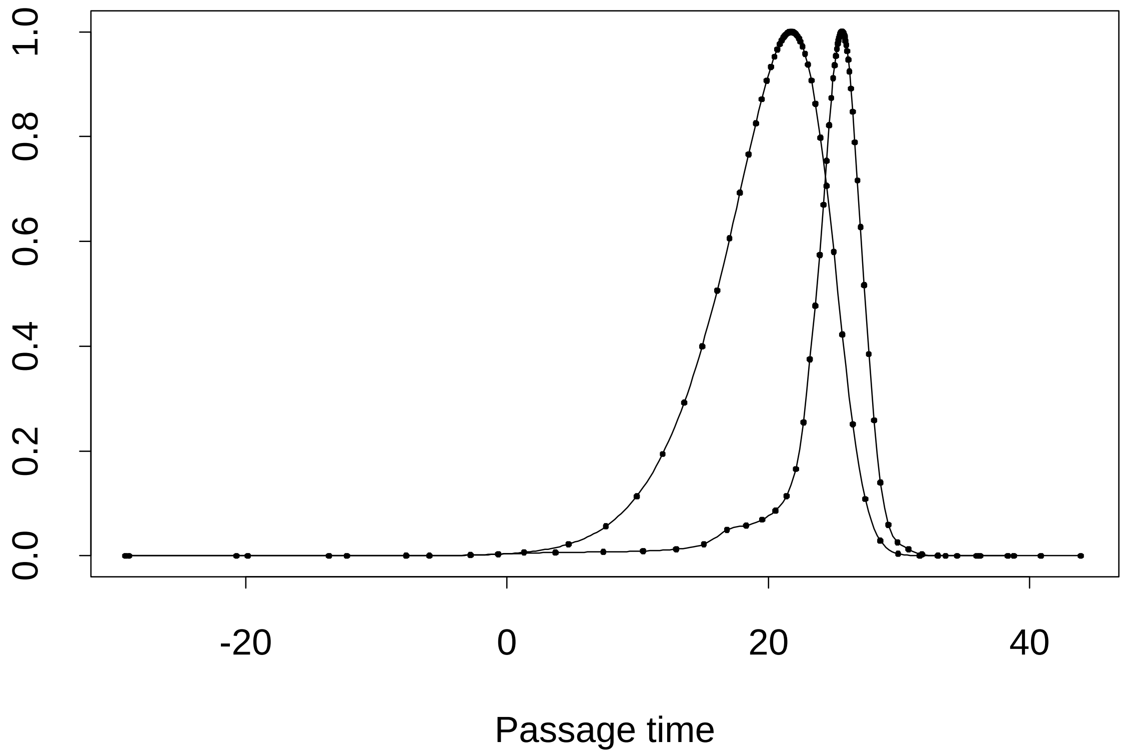
$$\mathcal{R}(\mu) = \max_{\sigma}\max\left\{\prod_{i=1}^{n} nw_i \mid 0 \leq w_i, \sum_i w_i = 1, \sum_i w_i\psi\left(\frac{x_i-\mu}{\sigma}\right) = 0,\right.$$
$$\left.\sum_i w_i\left[\psi\left(\frac{x_i-\mu}{\sigma}\right)^2 - 1\right] = 0\right\}$$

Newcomb's passage times of light



From Stigler

EL for mean and Huber's location



Curve for the mean is much more skewed by the outlier.

Robust statistic slightly skewed.

Side information

Maybe we know some relevant expectations.

For example, we want $\mathbb{E}(\boldsymbol{Y})$, we know $\mathbb{E}(\boldsymbol{X})$, and we observe $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ $i = 1, \ldots, n$

Then we can restrict our model to $w_i = F(\{\boldsymbol{x}_i, \boldsymbol{y}_i\})$ with

$$\sum_{i=1}^{n} w_i(\boldsymbol{x}_i - \mathbb{E}(\boldsymbol{X})) = 0.$$

The result

$$-2 \log \mathcal{R}_{Y|X}(\mu_y \mid \mu_{x0}) \to \chi^2_{(p)}$$

Maximum E. L. estimates

$$\text{Var}\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

MELE $$\widetilde{\mu}_y = \sum_{i=1}^{n} w_i \boldsymbol{y}_i \doteq \bar{\boldsymbol{Y}} - \Sigma_{yx}\Sigma_{xx}^{-1}(\bar{\boldsymbol{X}} - \mu_{x0})$$

$$n\,\text{Var}(\widetilde{\mu}_y) \doteq \Sigma_{y|x} \equiv \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

Using known mean reduces variance when $\boldsymbol{Y}$ correlated with $\boldsymbol{X}$

General side information

Can be incorporated via estimating equations

Known parameter	Estimating equation
mean	$\boldsymbol{X} - \mu_x$
α quantile	$1_{X \leq Q} - \alpha$
$\Pr(\boldsymbol{X} \in A \mid B)$	$(1_{\boldsymbol{X} \in A} - \rho)1_B$
$\mathbb{E}(\boldsymbol{X} \mid B)$	$(\boldsymbol{X} - \mu)1_B$

Qin has a nice example of Y vs X regression where $E(Y)$ is known

Maximum empirical likelihood estimates

Hartley & Rao	1968	means & finite population setting
O.	1991	means IID sampling
Qin & Lawless	1993	estimating eqns IID

Overdetermined equations

"10 equations in 5 unknowns:"

$$\mathbb{E}(m(\boldsymbol{X},\theta)) = 0, \quad \dim(m) > \dim(\theta)$$

Popular in econometrics, e.g. Generalized Method of Moments Hansen

Approaches:

1) Drop $\dim(m) - \dim(\theta)$ equations

2) Replace $m(\boldsymbol{X},\theta)$ by $m(\boldsymbol{X},\theta)A(\theta)$ where
   A a $\dim(m) \times \dim(\theta)$ matrix (IE pick $\dim(\theta)$ linear comb. of m)

3) GMM: estimate the optimal A

4) MELE: $\widetilde{\theta} = \arg\max_\theta \max_{w_i} \prod_i nw_i$ st $\sum_{i=1}^n w_i m(\boldsymbol{x}_i,\theta) = 0$

MELE has same asymptotic variance as using optimal $A(\theta)$

Bias scales more favorably with dimensions for MELE than for $\hat{A}$ methods

Newey, Smith, Kitamura

Qin and Lawless result

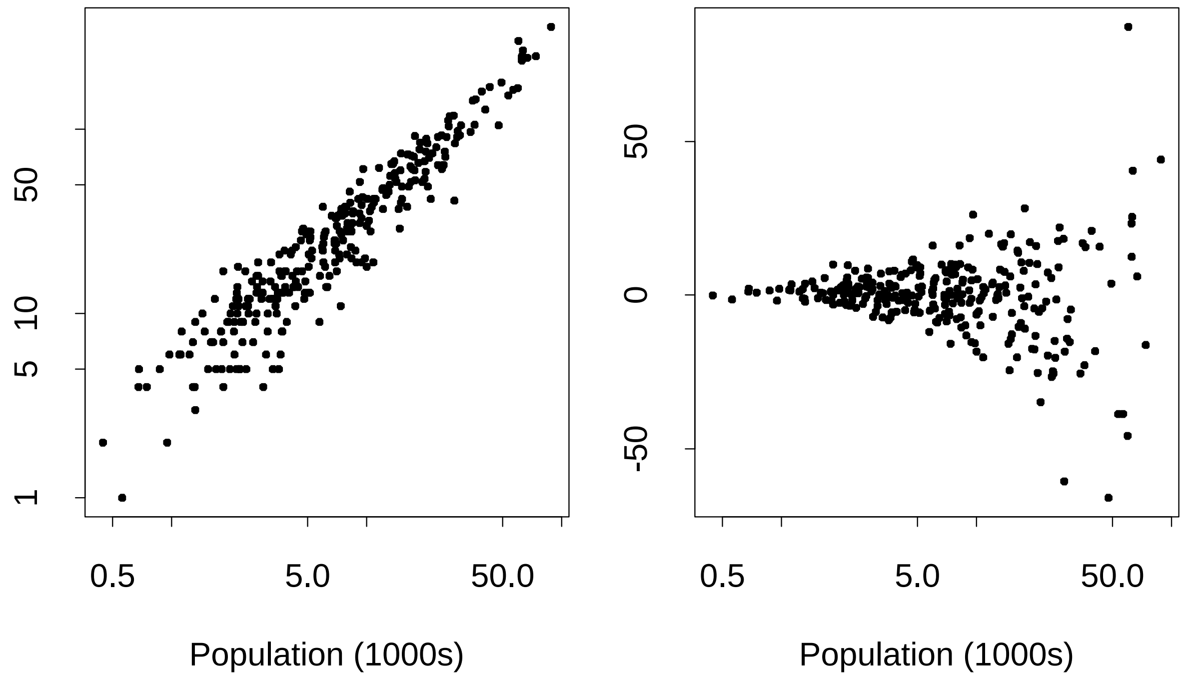
$\dim(m) = p + q \geq p = \dim(\theta)$ MELE $\widetilde{\theta}$

$$-2\log(\mathcal{R}(\theta_0)/\mathcal{R}(\widetilde{\theta})) \to \chi^2_{(p)}$$ conf regions for θ_0

$$-2\log \mathcal{R}(\widetilde{\theta}) \to \chi^2_{(q)}$$ goodness of fit tests when $q > 0$

Uses only differentiability, moment, identifiability and non-degeneracy conditions, no parametric assumptions.

Cancer deaths vs population, by county



Nearly linear regression nonconstant residual variance

Royall via Rice

Estimating equations for regression

$$\mathbb{E}(\boldsymbol{X}^{\mathsf{T}}(Y - \boldsymbol{X}^{\mathsf{T}}\beta)) = 0, \qquad \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^{\mathsf{T}}\hat{\beta})\boldsymbol{x}_i = 0$$

$$\mathcal{R}(\beta) = \max\Big\{\prod_{i=1}^{n} nw_i \mid \sum_{i=1}^{n} w_i \boldsymbol{Z}_i(\beta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\Big\}$$

$$\boldsymbol{Z}_i(\beta) = (Y_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)\boldsymbol{x}_i$$

need $\mathbb{E}(\|\boldsymbol{Z}\|^2) \leq \mathbb{E}\Big(\|\boldsymbol{X}\|^2(Y - \boldsymbol{X}^{\mathsf{T}}\beta)^2\Big) < \infty$

Don't need:

normality, constant variance, exact linearity

For cancer data

$P_i =$ population of i'th county in 1000s

$C_i =$ cancer deaths of i'th county in 20 years

$C_i \doteq \beta_0 + \beta_1 P_i$

$\hat{\beta}_1 = 3.58 \quad \Longrightarrow \quad 3.58/20 = 0.18$ deaths per thousand per year

$\hat{\beta}_0 = -0.53$ near zero, as we'd expect

Regression through the origin

$$C_i \doteq \beta_1 P_i$$

Residuals should have mean zero and be orthogonal to P_i

We want two equations in one unknown β_1

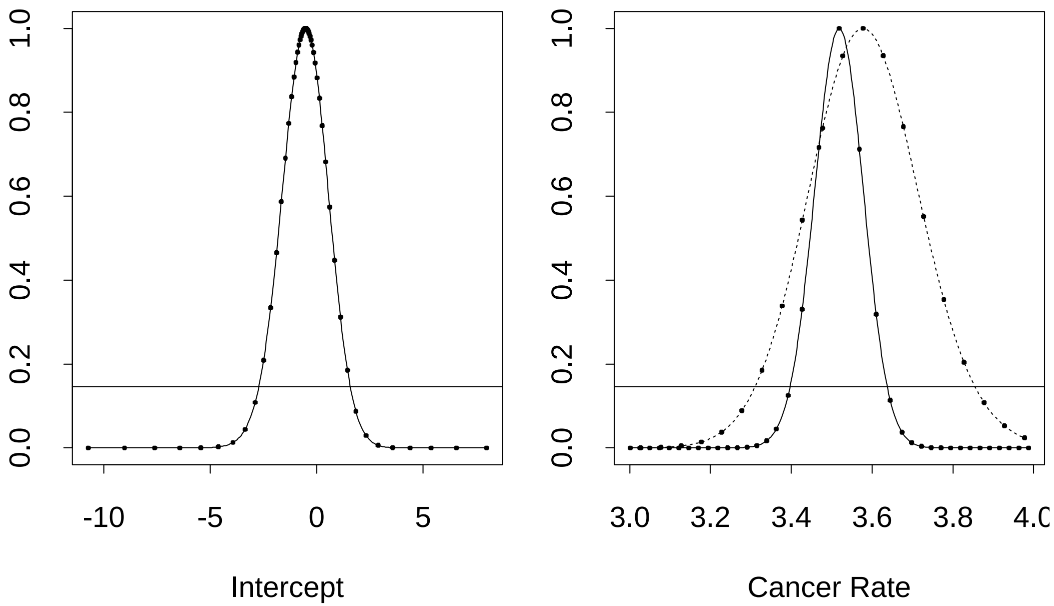
Equivalently, side information $\beta_0 = 0$

Least squares regression through origin does not solve both equations

$$\text{MELE } \widetilde{\beta}_1 = \arg\max_{\beta_1} \mathcal{R}(\beta_1)$$

$$\mathcal{R}(\beta_1) = \max\Big\{\prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i(C_i - P_i\beta_1) = 0,$$
$$\sum_{i=1}^n w_i P_i (C_i - P_i\beta_1) = 0, \sum_{i=1}^n w_i = 1, w_i \geq 0\Big\}$$

Regression parameters



Intercept nearly 0, MELE smaller than MLE

CI based on conditional empirical likelihood

Constraint narrows CI for slope by over half

Variance modelling

Working model $Y \sim \mathcal{N}(\boldsymbol{x}^{\mathsf{T}}\beta, e^{2\boldsymbol{z}^{\mathsf{T}}\gamma})$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)\, e^{-2\boldsymbol{z}_i^{\mathsf{T}}\gamma} \qquad (\text{weight} \propto 1/\text{var})$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i\Big(1 - \exp(-2\boldsymbol{z}_i^{\mathsf{T}}\gamma)(y_i - \boldsymbol{x}_i^{\mathsf{T}}\beta)^2\Big)$$

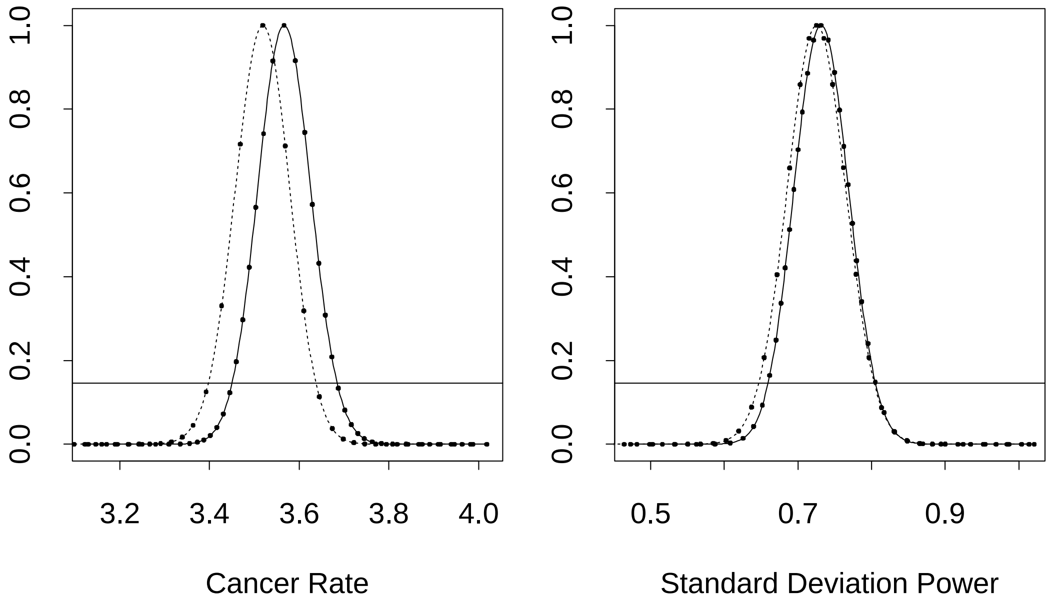
For cancer data

$$\boldsymbol{x}_i = (1, P_i)^{\mathsf{T}} \quad \boldsymbol{z}_i = (1, \log(P_i))^{\mathsf{T}}$$

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 P_i \quad \sqrt{\text{Var}(Y_i)} = \exp(\gamma_0 + \gamma_1 \log(P_i)) = e^{\gamma_0} P_i^{\gamma_1}$$

and $\beta_0 = 0$

Heteroscedastic model



Left: solid curve accounts for nonconstant variance

Right: solid curve forces $\beta_0 = 0$, and,

rules out $\gamma_1 = 1/2$ (Poisson) and $\gamma_1 = 1$ (Gamma)

Bayesian connection

- Use an informative prior on θ
- multiply by an empirical likelihood
- reverses usual non-informative paradigm

See Lazar (2003) also Rao & Wu (2010) (survey sampling)

MELEs for finite population sampling

1) use side information

   (a) population means, totals, sizes

   (b) stratum means, totals, sizes

2) take unequal sampling probabilities

3) use non-negative observation weights

Hartley & Rao, Chen & Qin, Chen & Sitter

More finite population results

χ^2 limits	$-2\left(1-\frac{n}{N}\right)\mathcal{R}(\mu) \to \chi^2$	Zhong & Rao
EL variance ests	via pairwise inclusion probabilities	Sitter & Wu
Multiple samples	varying distortions	Zhong, Chen, & Rao

EL confidence bands

Kolmogorov-Smirnov bands are too wide in the tails.

They are based on dist'n of $\max_x |\hat{F}(x) - F(x)|$

Equal width is not appropriate. Bands should narrow near the tails. Should also become skewed, e.g., to avoid 0.01 ± 0.03.

Replace by $\max$ of binomial likelihood ratio and get some large deviations optimality Berk & Jones

Recent work extends to censored data survivor function Matthews (2013)

Curve estimation problems

$$\mu(x) \equiv \mathbb{E}(Y \mid X = x) \quad \text{smooth}$$

- Estimate μ by kernel method
- Get confidence set for $\mu(x)$
- $x \in \mathbb{R}$, $y \in \mathbb{R}^2 \implies$ confidence tube
- $x \in \mathbb{R}^2$, $y \in \mathbb{R} \implies$ confidence sandwich

Have to contend with bias and pointwise vs simultaneous

Similar confidence sets for densities

Hall & O

Computation

$$\begin{aligned} \log \mathcal{R}(\theta) &= \max_{\nu} \log \mathcal{R}(\theta, \nu) \\ &= \max_{\nu} \min_{\lambda} \mathbb{L}(\theta, \nu, \lambda), \quad \text{where,} \end{aligned}$$

$$\mathbb{L}(\theta, \nu, \lambda) = -\sum_{i=1}^{n} \log\bigl(1 + \lambda^{\mathsf{T}} m(x_i, \theta, \nu)\bigr)$$

Inner and outer optimizations $\ll n$ dimensional

Used NPSOL, expensive and not public domain (but it works)

Algorithmic strategies

Newton's method to solve for a saddlepoint:

$$0 = \frac{\partial}{\partial \nu} \mathbb{L}(\theta, \nu, \lambda)$$
$$0 = \frac{\partial}{\partial \lambda} \mathbb{L}(\theta, \nu, \lambda)$$

Progress towards a saddle-point is more difficult to define than progress towards a mode.

Newton's method to solve

$$\max_{\nu} \mathcal{R}(\theta, \nu)$$

deriving gradient and Hessian from $\mathbb{L}(\theta, \nu, \lambda)$

These methods usually work well around the MLE.

As $n \to \infty$ the region where they work grows.

Next: research directions

Two main challenges for empirical likelihood are

1) escaping the convex hull

2) profiling out nuisance parameters

Lots of progress on problem 1

Chen, Variyath & Abraham (2008) Emerson & O (2009) Liu & Chen (2010) Tsao & Wu (2013)

Problem 2 is also difficult for parametric likelihoods; usually we just make a second order Taylor approximation to the log likelihood around the MLE.

Biconvex optimization methods needed for problem 2.

Thanks

1) Grace Yi
2) Christiane Lemieux
3) Marg Feeney & Anthea Dunne
4) NSF
5) David Sprott and my other Waterloo teachers