Empirical Likelihood for Reinforcement Learning

Art Owen

Department of Statistics

Stanford University

EL for RL

EL is being used in off policy estimation (OPE) in RL

Duchi, Glynn, Namkoong (2021)

distributionally robust inference

Dai, Nachum, Chow, Li, Szepesvári, Schuurmans (2020)

off policy confidence intervals, CoinDICE

Kallus, Uehara (2019)

efficient & stable OPE

Karampatziakis, Langford, Mineiro (2020)

EL for contextual bandits

Today

Introduce EL + useful properties / tweaks

Empirical Likelihood

1) likelihood function

2) without any parametric family (e.g., Gaussian)

Like a bootstrap that doesn't resample

Uses

- quantify uncertainty
- adjust for biased sampling
- bring in outside information

Pairs well with

Bayes Lazar (2003)

Causal Inference Tan (2010)

Parametric likelihood

'the chance of getting the data you got'

$$x_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$$

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

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/(2\sigma^2)} \, \mathrm{d}x_i$$

$$\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/(2\sigma^2)}$$

Likelihood uses

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \theta)$ true value θ_0

Maximum likelihood estimate

$$\hat{\theta} = \arg\max_{\theta} L(\theta; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

Likelihood ratio inferences

$$-2\log\Big(\frac{L(\theta_0)}{L(\hat{\theta})}\Big) \overset{\text{d}}{\to} \chi^2_{(q)} \qquad q = \dim(\theta)$$

Bayes

$$\text{posterior}(\theta) \propto \text{prior}(\theta) \times L(\theta)$$

Where does $f(\boldsymbol{x}; \theta)$ come from?

experience, convenience, tradition

Nonparametric likelihood

$$L(F) = \prod_{i=1}^{n} F(\{\boldsymbol{x}_i\}) \qquad \text{for } \boldsymbol{x}_i \overset{\text{iid}}{\sim} F$$

Maximize $L(F)$ over **all** distributions

$$\widehat{F} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}$$

where $\delta_{\boldsymbol{x}}$ is a point mass at $\boldsymbol{x}$

NPMLE

The empirical distribution $\widehat{F}$ is the

Nonparametric Maximum Likelihood Estimate of F

Kiefer & Wolfowitz (1956)

Note that $L(F) = 0$ for any continuous F

NPMLEs

Generalize empirical distribution to non-IID settings

Kaplan-Meier	Right censored survival times
Lynden-Bell	Left truncated star brightness
Hartley-Rao	Sample survey data
Grenander	Monotone density for actuarial data

Likelihood ratios

Functional $T(F)$

mean(F), median(F), regression slope(F)

Likelihood ratio

$$R(F) = \frac{L(F)}{L(\widehat{F})}$$

Mimic the parametric case

Confidence set:

$$\Big\{T(F) \mid R(F) \geqslant c\Big\}$$

Reject H_0:$T(F) = \theta_0$ iff:

$$\max\{R(F) \mid T(F) = \theta_0\} < c$$

Still get

$-2 \log R(\cdot) \xrightarrow{\mathrm{d}} \chi^2_{(q)}$ (under conditions)

Empirical likelihood (short story)

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

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

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

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

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

[1] A longer story explains these choices

Monograph O (2001)

Empirical likelihood for the mean

$$T(F) = \mathbb{E}(\boldsymbol{x}; F) = \sum_{i=1}^{n} w_i \boldsymbol{x}_i$$

Confidence region is

$$C = \left\{ \sum_{i=1}^{n} w_i \boldsymbol{x}_i \mid w_i > 0, \sum_{i=1}^{n} w_i = 1, \prod_{i=1}^{n} n w_i \geqslant c \right\}$$

Multinomial

Multinomial on the n data points $\boldsymbol{x}_i \implies n-1$ parameters w_i.

Profile likelihood ratio

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

Confidence set

$$\Big\{\mu \mid \mathcal{R}(\mu) \geqslant c\Big\}$$

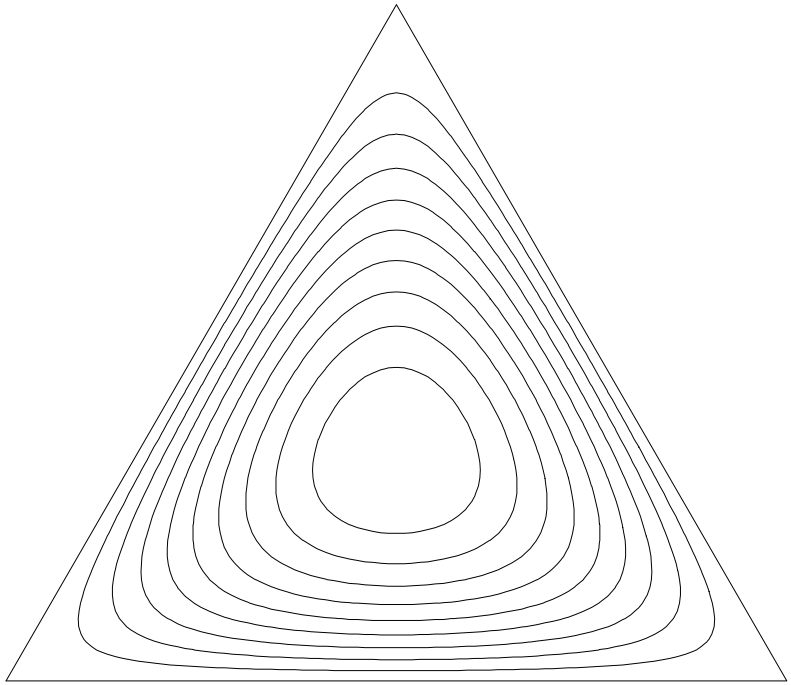
Hypothesis test

$$\text{Reject } \mathbb{E}(\boldsymbol{x}) = \mu \quad \Longleftrightarrow \quad \mathcal{R}(\mu) < c$$

Simplex

$$\triangle^{n-1} = \Big\{\boldsymbol{w} \in \mathbb{R}^n \mid w_i > 0, \sum_{i=1}^{n} w_i = 1\Big\}$$

Multinomial likelihood for $n = 3$



Contours of $\prod_i nw_i$ MLE at center of $\triangle^2$ LR$= i/10, i = 0, \ldots, 9$

Empirical likelihood theorem

If $\boldsymbol{x}_i \overset{\text{iid}}{\sim} F_0 \qquad \mu_0 = \mathbb{E}_{F_0}(\boldsymbol{x}) \qquad V_0 = \text{Var}_{F_0}(\boldsymbol{x})$

Then

$$-2 \log \mathcal{R}(\mu_0) \xrightarrow{\text{d}} \chi^2_{(q)} \qquad q = \text{rank}(V_0)$$

Same as parametric limit

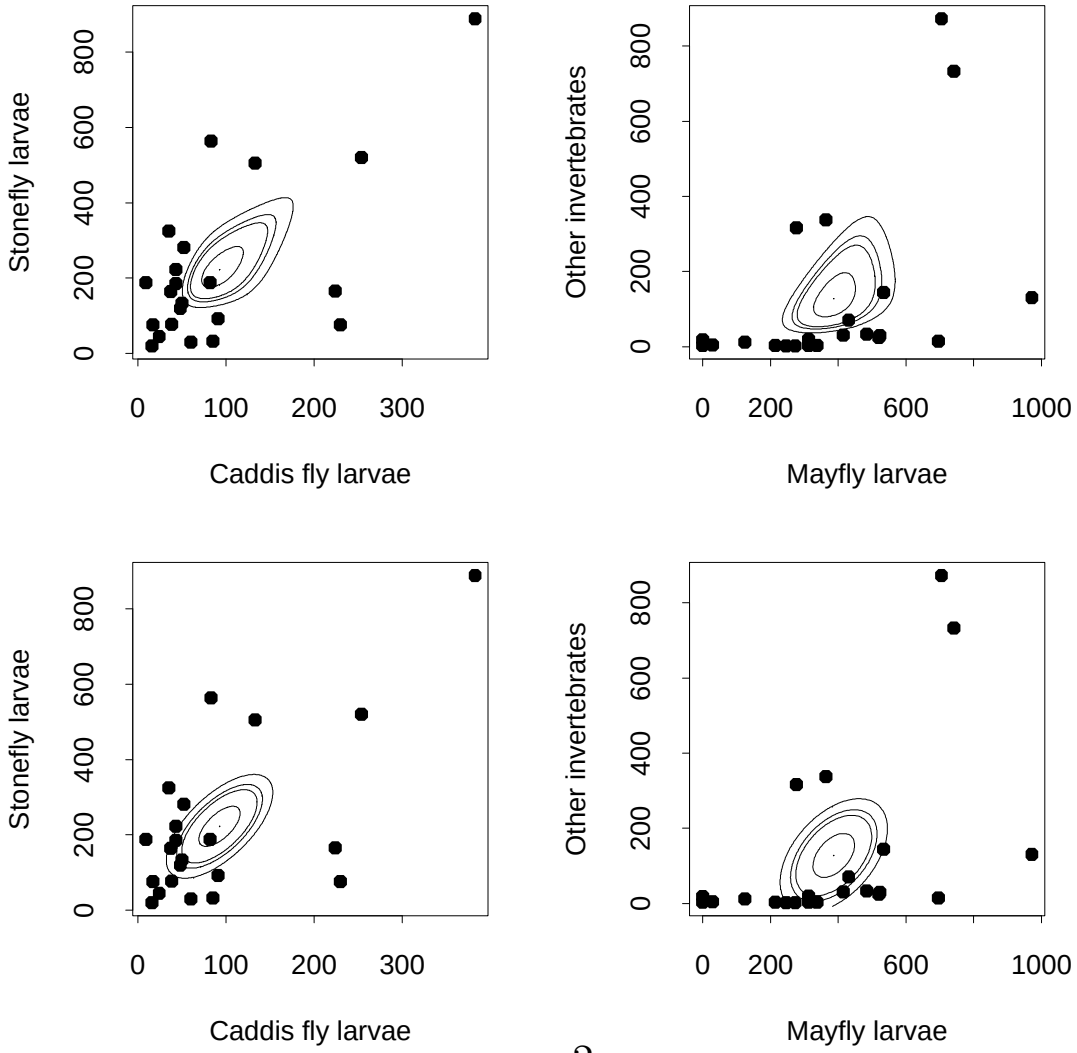
No apparent penalty for using $n - 1$ parameters.

Dipper, Cinclus cinclus

Eats larvae of Mayflies, Stoneflies, Caddis flies, other

Photo By Mark Medcalf. From Wikipedia. Lic under CC by 2.0. https://commons.wikimedia.org/w/index.php?curid=15739681

Dipper diet means



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

Data from 22 rivers in Wales: Iles (1993)

Hall (1990): region shapes are 'second order correct'

Nonparametric inferences

- Asymptotically valid:

$$\Pr(\mu_0 \in C) = 1 - \alpha + O\Big(\frac{1}{n}\Big)$$

- Like using a bootstrap
- Replace sampling by optimization

Computing EL for the mean

Start with the convex hull:

$$\mathcal{H} = \mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \left\{ \sum_{i=1}^{n} w_i \boldsymbol{x}_i \;\middle|\; \boldsymbol{w} \in \triangle^{n-1} \right\}$$

$$\mu \notin \mathcal{H} \implies \mathcal{R}(\mu) = 0 \implies \log \mathcal{R}(\mu) = -\infty$$

If $\mu \in \mathcal{H}$ then $\mathcal{R}(\mu) > 0$

and easily computable

Convex hull constraint

Awkward for large d:

there are fixes

The Lagrangian

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

Set $\dfrac{\partial G}{\partial w_i} = 0$

A little algebra gives $\gamma = -n$

Then

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

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

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

We have to find $\lambda \in \mathbb{R}^d$ to get $\boldsymbol{w} \in \mathbb{R}^n$

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 $\log R$

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

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

Newton step $=$ weighted least squares

Self-concordant convex version O (2013)

Bartlett correction

DiCiccio, Hall, Romano (1991)

Replace $\chi^{2,1-\alpha}$ by $\left(1+\frac{a}{n}\right)\chi^{2,1-\alpha}$ for carefully chosen a

and get coverage $1-\alpha+O\left(\frac{1}{n^2}\right)$

same as for parametric likelihoods

Sets in slowly

Estimating equations

"means $\rightarrow$ anything"

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

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

Examples

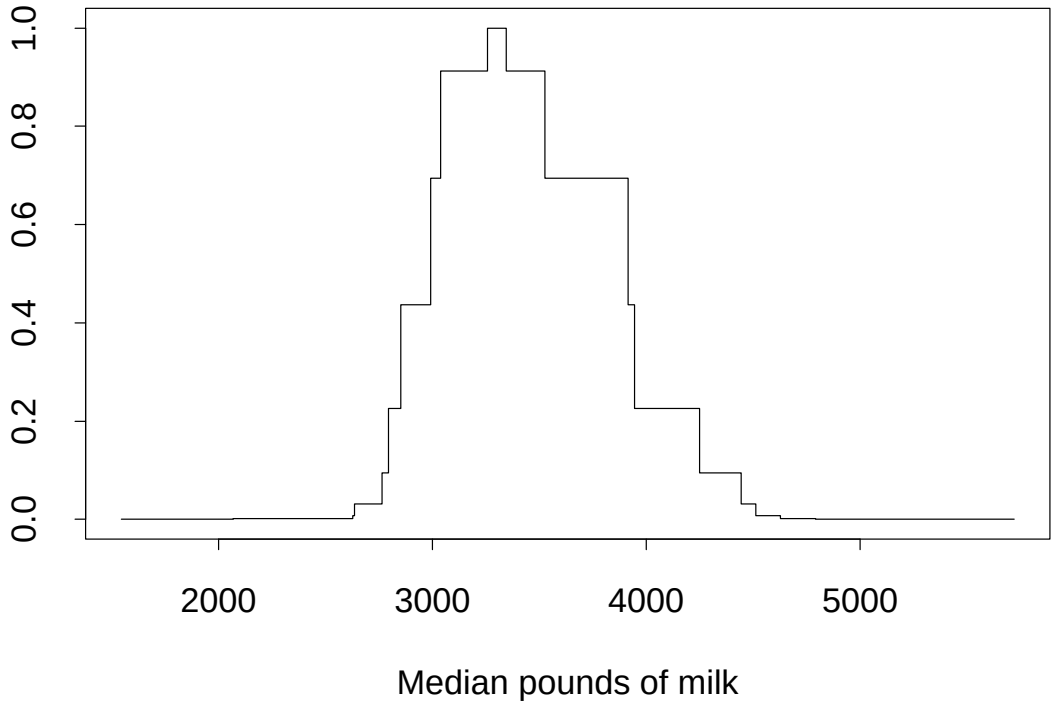
$m(\boldsymbol{x}, \theta)$	Statistic θ
$\boldsymbol{x} - \theta$	Mean
$1_{\boldsymbol{x} \in A} - \theta$	$\Pr(\boldsymbol{x} \in A)$
$1_{x \leqslant \theta} - \frac{1}{2}$	Median
$\frac{\partial}{\partial \theta} \log(f(\boldsymbol{x}; \theta))$	MLE as if $\boldsymbol{x} \sim f$
$(y - \boldsymbol{x}^\mathsf{T}\theta)\boldsymbol{x}$	Linear regression
$(\boldsymbol{x} - \theta)1_{\boldsymbol{x} \in A}$	$\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{x} \in A)$

Long history: Godambe (1960), Huber, White, Thompson

Empirical likelihood for a median

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

Reweight data so that half the weight is below m



LR is constant between observations

For α-quantile: $\mathbb{E}(1_{x \leqslant \theta} - \alpha) = 0$

Using estimating equations

$$\mathcal{R}(\theta) = \max\left\{\prod_{i=1}^{n}(nw_i) \mid \boldsymbol{w} \in \Delta^{n-1}, \sum_{i=1}^{n} w_i m(\boldsymbol{x}_i, \theta) = 0\right\}$$

Sampling bias

If $\boldsymbol{x}_i \sim F$ observed with probability proportional to $\omega(\boldsymbol{x})$ then use

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

Small $\omega(\boldsymbol{x})$ are problematic

Side information

$(\boldsymbol{x}_i, \boldsymbol{y}_i) \overset{\text{iid}}{\sim} F$

We want $\mathbb{E}(\boldsymbol{y}) \equiv \mu_{\boldsymbol{y}0}$, and we **know** $\mathbb{E}(\boldsymbol{x}) = \mu_{\boldsymbol{x}0}$.

Enforce the knowledge

Add the constraint:

$$\sum_{i=1}^{n} w_i(\boldsymbol{x}_i - \mu_{\boldsymbol{x}}) = 0$$

Get sharper inferences.

The result

$$\mathcal{R}_{\boldsymbol{y}|\boldsymbol{x}}(\mu_{\boldsymbol{y}} \,|\, \mu_{\boldsymbol{x}}) \equiv \mathcal{R}_{\boldsymbol{x},\boldsymbol{y}}(\mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}) / \mathcal{R}_{\boldsymbol{x}}(\mu_{\boldsymbol{x}})$$

$$-2 \log \mathcal{R}_{\boldsymbol{y}|\boldsymbol{x}}(\mu_{\boldsymbol{y}0} \mid \mu_{\boldsymbol{x}0}) \to \chi^2_{(p)}$$

O (1991)

Maximum EL estimates

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

Maximize $\prod_i w_i$ with $\sum_i w_i \boldsymbol{x}_i = \mu_{\boldsymbol{x}0}$

Maximum Empirical Likelihood Estimate (MELE)

Weight $\boldsymbol{y}$ with the $\boldsymbol{x}$-optimal weights:

$$\widetilde{\mu}_y = \sum_{i=1}^{n} w_i \boldsymbol{y}_i \approx \bar{\boldsymbol{y}} - \Sigma_{yx}\Sigma_{xx}^{-1}(\bar{\boldsymbol{x}} - \mu_{\boldsymbol{x}0})$$

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

Using known mean reduces variance when $\boldsymbol{y}$ correlated with $\boldsymbol{x}$

O (1991)

Overdetermined equations

E.g.: "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

Nobel Laureate Hansen (1982)

Minimal example

Regression through origin:

$$y_i = \beta x_i + \varepsilon_i$$
$$\mathbb{E}(y_i - \beta x_i) = 0 = \mathbb{E}(x_i(y_i - \beta x_i))$$

One parameter and two constraints

MELE

Maybe **no** θ has

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

Relax to $\boldsymbol{w} \in \triangle^{n-1}$ and solve

$$\sum_{i=1}^{n} w_i m(\boldsymbol{x}_i, \theta) = 0 \tag{1}$$

$$(\boldsymbol{w}^*, \tilde{\theta}) = \arg\max_{\boldsymbol{w},\theta} \prod_{i=1}^{n} n w_i \quad \text{s.t.} \quad \boldsymbol{w} \in \triangle^{n-1} \quad \text{and} \quad (1)$$

vs GMM

$\tilde{\theta}$ has same variance, less bias

Newey, Smith (2004)

Some MELEs

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

Qin and Lawless (1993)

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

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

$$-2\log\mathcal{R}(\widetilde{\theta}) \to \chi^2_{(q)}$$ tests the constraint

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

Next

As many of

- Bayes
- power
- escaping the hull
- bootstrap of EL

as time permits

Bayesian empirical likelihood

$$p(\theta \,|\, \text{data}) \propto \pi(\theta) \times \mathcal{R}(\theta)$$

Science $\Longrightarrow$ prior

Data $\Longrightarrow$ likelihood

First study Lazar (2003)

Exponential tilting EL Schennach (2007)

Approximate Bayesian Computation: Mengersen, Pudio, Robert (2013)

Bayesian EL + Hamiltonian MCMC Chaudhuri, Mondal, Yin (2017)

Justification

Subtle: ask me later

Power

Large deviations power is about

$$\limsup_{n\to\infty} \frac{1}{n} \log\Big(\Pr(\text{test accepts false } H_0)\Big), \quad \text{and}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log\Big(\Pr(\text{test rejects true } H_0)\Big)$$

for fixed distributions of $\boldsymbol{x}_i$

Kitamura (2001) finds that EL tests dominate any other test at any $P \notin H_0$

Generalizes multinomial result of Hoeffding (1965)

Standard $1/\sqrt{n}$ asymptotics

Lazar and Mykland (1998)

Parametric and empirical LR tests match power to second order

Two big challenges

The convex hull

$\mathcal{R}(\mu) = 0$ for μ outside the convex hull

$\implies$ all confidence regions nested within convex hull

problematic for large d

Profiling

It can be hard to compute

$$\max_{\theta_2} \mathcal{R}(\theta_1, \theta_2)$$

That can also be hard for parametric likelihoods

There one usually uses

$$\log(\ell(\theta)) \approx \log \ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^{\mathsf{T}} \mathrm{Hessian}(\hat{\theta})(\theta - \hat{\theta})$$

Euclidean log likelihood

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

Replace loglik by

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

Allows $w_i < 0 \implies$ out of hull

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

Reduces to

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

Huber-White covariance for regression

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

Allows $w_i = 0$ but not $w_i < 0$

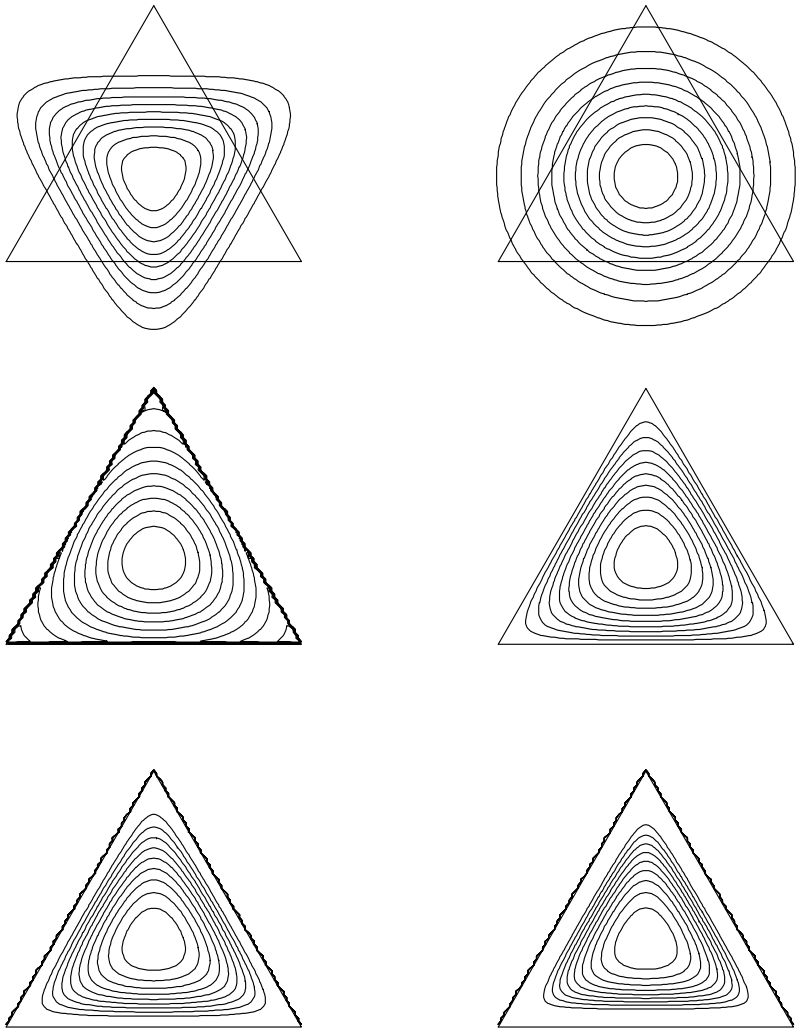
Hellinger distance

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

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

Additional points

We can add an unobserved point $\boldsymbol{x}_0 = \boldsymbol{x}_0(\mu)$ to get

$$\mu \in \mathcal{H}(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

Retain χ^2 distribution

Sometimes use two points

Chen, Variyath, Abraham (2008), Emerson & O (2009) , Tsao & Wu (2013/14)

Space warping

M. Tsao (2013) maps $\mu \in \mathbb{R}^d \to \tilde{\mu} \in \mathcal{H}$

$$\widetilde{\mathcal{R}}(\mu) = \mathcal{R}(\tilde{\mu})$$

Bootstrap calibration

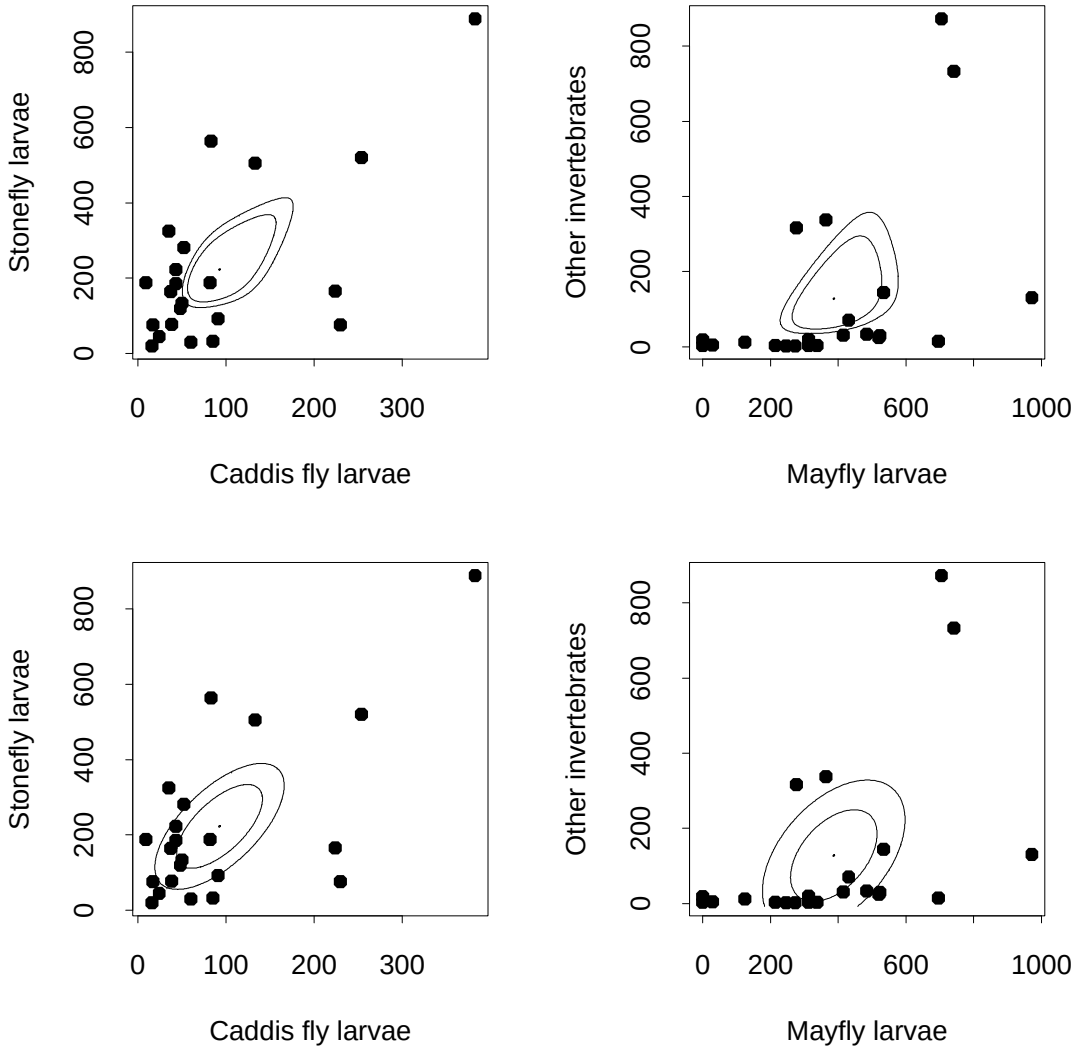
Replace $\chi^2_{(q)}$ by a bootstrap

We want	distn of $-2\log\mathcal{R}(\mu_0)$	for $\boldsymbol{x}_i \overset{\text{iid}}{\sim} F$
We use	distn of $-2\log\mathcal{R}(\bar{\boldsymbol{x}})$	for $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \widehat{F}$

Coverage is $1-\alpha+O(n^{-2})$

like bootstrapping the bootstrap

Bootstrap (and χ^2) calibrated Dipper regions



Thanks

- Csaba Szepesvari and Lin Yang
- NSF for long term support
- ICML organizers