Bayesian Empirical Likelihood:

a survey

Art B. Owen

Stanford University

These are the slides I presented at BNP 11 in Paris, a very fine meeting. I have added a few interstitial slides like this one to include comments I made while presenting the slides and to give some background thoughts in this area.

The session included a delightful presentation by Nils Lid Hjort on some not yet published theorems where he connects Bayesian empirical likelihood to Dirichlet process Bayesian nonparametrics.

I was pleased to see our session got a positive review on Xian's Og (written by our session chair, Christian Robert).

Also, the conference included two very strong posters by students working on Bayesian empirical likelihood. Frank Liu (Monash) has a nice way to robustify BETEL. Laura Turbatu (Geneva) took a close look at high order asymptotics to compare frequentist coverage of Bayes with various kinds of nonparametric priors.

The goal

Science driven prior $\times$ Data driven likelihood

This goal is not everybody's goal. But this goal means you don't have to make up a likelihood which could bring unwanted and consequential restrictions.

If the data deliver the likelihood and the prior is chosen flat or by previously known science, then what is there left for the statistician to do?

The goal

Science driven prior $\times$ Data driven likelihood

The ask

IID data and

an estimand defined by estimating equations.

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$$

$$\int m(\boldsymbol{x}, \theta)\, \mathrm{d}F(\boldsymbol{x}) = 0, \qquad \theta \in \mathbb{R}^p$$

We have IID $\boldsymbol{X}$ and an estimating equation. If you picture the plate diagram, it is one of the simplest possible ones. It covers an important set of problems, but Bayes includes much else as well.

Outline

1) Empirical Likelihood
2) Bayes EL Lazar (2003)
3) Exponential tilting, and ETEL Schennach (2005,2007)
4) ABC and EL Mengersen, Pudio, Robert (2013)
5) BEL with Hamiltonian MCMC Chaudhuri et al (2017)
6) Some other recent works

Notes

- Apologies about necessarily omitted papers.
- Motivation: bioinformatics.

In the bioinformatics motivation, the data are counts. They are assuredly not Poisson. So a common response is to model them as negative binomial. Of course they are not that either. This is work in progress.

SAMSI is having a year on QMC. The original slides touted the opening workshop. However the application deadline will almost surely have passed by the time you're reading this.

Empirical likelihood

For $\boldsymbol{X}_i \overset{\text{iid}}{\sim} F$ observed $\boldsymbol{X}_i = \boldsymbol{x}_i \in \mathbb{R}^d$, define

$$L(F) = \prod_{i=1}^{n} F(\{\boldsymbol{x}_i\}).$$

The nonparametric MLE (NPMLE) is

$$\widehat{F} \equiv \arg\max_F L(F) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}$$

There are many other NPMLEs.

E.g., Kaplan-Meier for right censored survival,

Log concave densities: Walther, Dumbgen, Rufibach, Cule, Samworth, Stewart $\cdots$

Basic notions

Model: $\boldsymbol{x}_i \overset{\text{iid}}{\sim} F$ True value: $F = F_0$ NPMLE: $\widehat{F}$

Estimand $T(F)$

E.g., population mean, median, regression coefficient etc.

True value: $\theta_0 = T(F_0)$ NPMLE: $\hat{\theta} = T(\widehat{F})$

Empirical likelihood ratios

	Point est.	Interval / Test
Parametric	$\hat{\theta} = \arg\max_\theta R(\theta)$	$-2 \log R(\theta_0) \to \chi^2$
Non-parametric	$T(\widehat{F})$	$***$

The purpose of empirical likelihood ratios is to fill in the $***$,

getting likelihood ratio confidence intervals and tests without parametric assumptions.

I.e., a nonparametric Wilks' theorem

O (2000) Chapman & Hall, Monograph

Empirical likelihood ratio

Let $w_i = w_i(F) = F(\{\boldsymbol{x}_i\})$

Then $w_i \geqslant 0$ and $\sum_{i=1}^n w_i \leqslant 1$.

$$R(F) = \frac{R(F)}{R(\widehat{F})} = \prod_{i=1}^n \frac{w_i}{1/n} = \prod_{i=1}^n (nw_i)$$

Literal EL confidence regions

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

Profile likelihood

$$\mathcal{R}(\theta) = \max\{R(F) \mid T(F) = \theta\}$$

$$\text{Conf set} = \{\theta \mid \mathcal{R}(\theta) \geqslant c\}.$$

We still have to choose c

Usual EL

It is usual to force $\sum_i w_i = \sum_i F(\{\boldsymbol{x}_i\}) = 1$.

$R(F) \geqslant c$ already implies $1 - F(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) = O(1/n)$

It is a multinomial likelihood

Supported on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$

There are other choices

Empirical likelihood for the mean

$$T(F) = \int \boldsymbol{x} \, \mathrm{d}F(\boldsymbol{x})$$

Profile likelihood

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

Confidence set

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

To pick c

$$-2\log(\mathcal{R}(\mu_0)) \xrightarrow{\mathrm{d}} \chi^2_{(k)}, \quad k = \mathrm{rank}(\mathrm{Var}(\boldsymbol{x}))$$

$$c = \exp\left(-\frac{1}{2}\chi^{2,1-\alpha}_{(k)}\right)$$

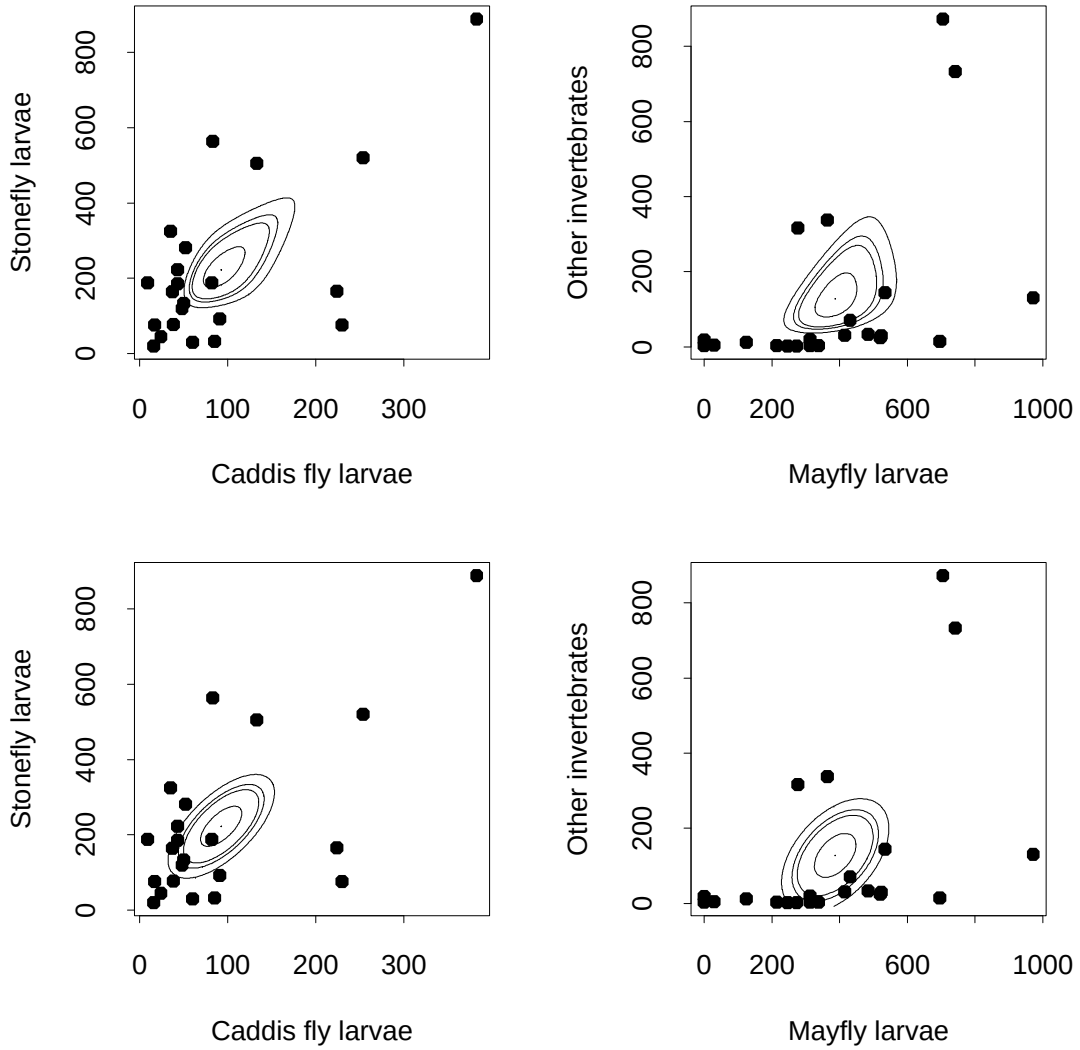
Dipper, Cinclus cinclus

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 Iles (1993) 22 sites in Wales.

Hall (1990) quantifies correctness of shape.

Computing EL for the mean

Easy by convex duality.

The dual is self-concordant

$\Longrightarrow$ truncated Newton assured of convergence.

O (2013) Canadian Journal of Statistics

R code online

Estimating equations

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

Examples

$m(\boldsymbol{x}, \theta)$	Estimand θ
$\boldsymbol{x} - \theta$	$\mathbb{E}(\boldsymbol{x})$
$1_{x<\theta} - 0.5$	median(x)
$(\boldsymbol{x} - \theta) \times 1_{\boldsymbol{z} \in A}$	$\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{z} \in A)$
$\boldsymbol{x}(y - \boldsymbol{x}^{\mathsf{T}}\beta)$	regression
$\frac{\partial}{\partial \theta} \log f(\boldsymbol{x}, \theta)$	MLE

Note that some of these estimands can be well defined without making any assumptions about the likelihood. For instance, we could well want a median without wanting to assume a double exponential distribution or any other whose MLE is the median.

Similarly, least squares estimates are interpretable as sample based counterparts to population least squares quantities.

Other cases, like logistic regression, make it harder to interpret what you get if the motivating parametric model does not hold. That said, you might still prefer reliable to unreliable uncertainty quantification for such an estimand.

EL properties

- High power Kitamura (2001), Lazar & Mykland (1998)

  Comparable to a true parametric model (when there is one)

- Bartlett correctable DiCiccio, Hall & Romano (1991)

  Coverage errors $O(1/n) \to O(1/n^2)$

Challenges

- Convex hull issue, fixable by pseudo-data,

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

- Computing $\max\{\mathcal{R}(\theta_1, \ldots, \theta_p) \mid \theta_j = \theta_{j0}\}$ can be hard

  O (2000) used expensive NPSOL solver

Possibility

Sampling might work more easily than optimization

Overdetermined estimating equations

$$\theta \in \mathbb{R}^p, \quad m(\boldsymbol{x}, \theta) \in \mathbb{R}^q, \quad q > p$$

q equations in $p < q$ unknowns: $\frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{x}_i, \theta) = 0$

Popular in econometrics

E.g., generalized method of moments (GMM) Hansen (1982)

E.g., regression through $(0, 0)$

$$\mathbb{E}(y - x\beta) = \mathbb{E}(x(y - x\beta)) = 0$$

Misspecification

Sometimes no $\theta \in \mathbb{R}^p$ gives $\mathbb{E}(m(\boldsymbol{x}, \theta)) = 0 \in \mathbb{R}^q$

EL for overdetermined

Get n more multinomial parameters $w_1, \ldots, w_n$

NPMLE

$$\max_{\theta, \boldsymbol{w}} \prod_{i=1}^{n} w_i \quad \text{subject to} \quad \sum_{i=1}^{n} w_i m(\boldsymbol{x}_i, \theta) = 0, \; \sum_{i=1}^{n} w_i = 1, \; w_i > 0$$

Then solve $\sum_i \hat{w}_i m(\boldsymbol{x}_i, \theta) = 0$ for $\hat{\theta}$

Newey & Smith (2004): same variance & less bias than GMM

Bayesian EL

$$p(\theta \mid \boldsymbol{x}) \propto p(\theta)\mathcal{R}(\theta \mid \boldsymbol{x})$$

Lazar (2003) justifies by a Bayesian CLT, like Boos & Monahan (1992)

Also Hjort (Today!)

Is that Bayesian enough?

Cressie-Read likelihoods

There are numerous alternative nonparametric likelihoods.

Chang & Mukerjee (2008) give conditions for them to have a very accurate BCLT for very general priors.

EL is uniquely able to do so.

Some others will work with a flat prior.

Also

Poster by Turbatu (yesterday)

Least favorable families

$$\Pr_{\mathrm{EL}}(\boldsymbol{x}_i) = w_i = \frac{1}{n}\frac{1}{1+\lambda(\theta)^{\mathsf{T}} m(\boldsymbol{x}_i,\theta)}$$

p-dimensional $\theta \implies$ p-dimensional $\lambda(\theta)$.

There is a least favorable p-dimensional family for θ.

Any other family makes the inference artificially easy.

EL family is asymptotically the LFF DiCiccio & Romano (1990)

The connection

$p(\theta)$ induces a prior on the LFF, and

$\mathcal{R}(\theta \mid \boldsymbol{x}) \doteq$ likelihood on LFF

$\implies p(\theta) \times \mathcal{R}(\theta \mid \boldsymbol{x}) \doteq$ posterior on LFF

$=$ posterior on θ

To my knowledge nobody has tracked down exactly how the approximations work out for the least favorable family motivation of Bayesian EL.

Exponential tilting (ET)

Schennach (2005)

To get an empirical probability model, for

$$\theta \in \Theta \subset \mathbb{R}^p, \qquad \text{partition} \quad \Theta = \bigcup_{\ell=1}^{N} \Theta_\ell$$

$p(\theta)$ is multinomial on N disjoint cells.

Now let $N \to \infty$ for universal approximation.

Ultimately

$$p(\theta \mid \boldsymbol{x}) \propto p(\theta) \times \prod_{i=1}^{n} w_i^*(\theta), \quad \text{where}$$

$$w_i^* \quad \text{maximize} \quad -\sum_{i=1}^{n} w_i \log(w_i)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} w_i^* \boldsymbol{x}_i = \theta, \quad \sum_i w_i = 1$$

Exponential tilting EL

Schennach (2007) Form parametric family via exponential tilting

$$w_i^*(\theta) = \frac{e^{\lambda(\theta)^\mathsf{T} m(\boldsymbol{x}_i,\theta)}}{\sum_{j=1}^n e^{\lambda(\theta)^\mathsf{T} m(\boldsymbol{x}_j,\theta)}}, \quad \text{where}$$

$$0 = \sum_{i=1}^n e^{\lambda(\theta)^\mathsf{T} m(\boldsymbol{x}_i,\theta)} m(\boldsymbol{x}_i,\theta)$$

Likelihood

$$L(\theta) = \prod_{i=1}^n n w_i^*(\theta)$$

ETEL

$$\hat{\theta}_{\mathrm{ETEL}} = \arg\max_\theta L(\theta)$$

Works better on misspecified overdetermined models,

i.e., no $\theta \in \mathbb{R}^p$ makes $\mathbb{E}(m(\boldsymbol{x},\theta)) = 0 \in \mathbb{R}^q$

Reason: at least there is an estimand.

If we know that no θ can make $\mathbb{E}(m(\boldsymbol{x}, \theta)) = 0$ then I'm not sure why we still want an estimand. It seems like we should instead change the choice of m.

BETEL

Chib, Shin, Simoni (2016,2017)

Use a slack parameter V_k.

$V_k \neq 0$ for non-truthfulness of moment condition k.

At most $m - p$ nonzero V_k.

Metropolized Independence Sampling from a t distribution.

Asymptotic model selection

Use log marginal likelihood.

If there is a true model then it will be selected over any false one.

Among true models: most equations holding wins.

Robust BETEL

poster by Liu (yesterday)

EL with ABC

Mengersen, Pudlo & Robert (2013)

Basic EL-ABC

Sample $\theta_i \sim p(\theta)$

Compute $w_i = \mathcal{R}(\theta_i \mid \boldsymbol{x})$

Normalize the weights

Advanced EL-ABC

Use adaptive multiple importance sampling (AMIS)

Algorithm has 5 loops

Max depth 3

Hamiltonian MCMC

Chaudhuri, Mondal & Yin (2017)

Using $p(\theta)\mathcal{R}(\theta \mid \boldsymbol{x})$ they report that:

- Gibbs is not applicable
- Random walk Metropolis has problems with irregular support of $\mathcal{R}(\theta \mid \boldsymbol{x})$

Their solution

Use gradients, Hamiltonian MCMC.

Gradient gets steep just where you need it.

Posterior moments

Vexler, Tao & Hutson (2014)

For $x_i \in \mathbb{R}$

$$\frac{\int_{x_{(1)}}^{x_{(n)}} \theta^k \, \mathcal{R}(\theta \mid \boldsymbol{x}) p(\theta) \, \mathrm{d}\theta}{\int_{x_{(1)}}^{x_{(n)}} \mathcal{R}(\theta \mid \boldsymbol{x}) p(\theta) \, \mathrm{d}\theta}$$

Nonparametric James-Stein

For $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik}) \in \mathbb{R}^K$, get $\mathcal{R}_j(\theta_j) = \mathcal{R}(\theta_j \mid x_{1j}, \ldots, x_{nj})$

$$\hat{\theta}_j = \frac{\int \theta_j \, \mathcal{R}_j(\theta) \varphi\left(\frac{\theta_j - \theta_0}{\sigma_*}\right) \mathrm{d}\theta_j}{\int \mathcal{R}_j(\theta) \varphi((\theta_j - \theta_0)/\sigma_*) \, \mathrm{d}\theta_j}$$

$$\sigma_*^2 = \arg\max_{\sigma^2} \sum_{j=1}^{K} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \int \mathcal{R}_j(\theta_j) e^{-\theta_j^2/(2\sigma^2)} \, \mathrm{d}\theta_j\right)$$

Quantiles

Vexler, Yu & Lazar (2017)

EL Bayes factors for quantile $\theta \equiv Q^\alpha = Q_0$ vs $Q^\alpha \neq Q_0$

They overcome a technical challenge handling $\int_{H_A} \mathcal{R}(\theta \mid \boldsymbol{x}) p(\theta)\, \mathrm{d}\theta$ because $\mathcal{R}$ has jump discontinuities.

Also two sample Bayes factors.

If scientific or regulalatory interest is in a specific quantile then the user does not have to pick a whole parametric family.

Quantile regression

$$\Pr(Y \leqslant \alpha + \beta x \mid x) = \tau, \quad \text{e.g., } \tau = 0.9$$

Estimating equations

$$0 = \mathbb{E}\big(1_{y \leqslant \alpha+\beta x} - \tau\big)$$
$$0 = \mathbb{E}\big(x\big(1_{y \leqslant \alpha+\beta x} - \tau\big)\big)$$

Lancaster & Jun (2010)

BETEL for quantiles similar to Jeffrey's prior

Yang & He (2012)

BCLT accounting for non smoothness of estimating equations

Shrinking priors pool β for multiple τ

Both use MCMC

Summary

We can do Bayes without requiring a parametric family for the observations.

It is the “other nonparametric Bayes”.

Thanks

- Luke Bornn, invitation
- Sanjay Chaudhuri, many references
- Scientific and organizing committees
- Especially Judith Rousseau
- NSF DMS-1521145, DMS-1407397