Empirical Likelihood and Bayesian Nonparametrics

Nils Lid Hjort

Department of Mathematics, University of Oslo

Bayesian Nonparametrics XI, Paris, June 2017
(Belgirate '97, Reading '99, Ann Arbor '01, Roma '04, Jeju '06,
Cambridge '07, Torino '09, Veracruz '11,
Amsterdam '13, Raleigh '15)

FOCUS DRIVEN STATISTICAL INFERENCE WITH COMPLEX DATA

The main setting: inference for $\theta(F)$ with iid data

Suppose $y_1, \ldots, y_n$ are i.i.d. from F, and that a 'nonparametric estimand' $\theta = \theta(F)$ is identified via estimating equation:

$$\mathrm{E}_F\, m(Y, \theta) = \int m(y, \theta)\, \mathrm{d}F(y) = 0.$$

May have $m(y, \theta)$ and θ of dimension p.

Application 1: $m(y, \theta) = h(y) - \theta$ corresponds to $\theta = \mathrm{E}_F\, h(Y)$.

Application 2: $m_j(y, \theta_j) = I\{y \leq \theta_j\} - q_j$ corresponds to $\theta_j = F^{-1}(q_j)$, quantiles.

Application 3: For a given parametric family $f(y, \theta)$, let $m(y, \theta)$ be the score function. Then $\theta = \theta(F)$ is the least false parameter value, the minimiser of $\mathrm{KL}(f_{\mathrm{true}}, f_\theta)$.

Theory given in the talk extends to smooth functions of such θ (e.g. smooth functions of means, of quantiles, of ML estimators, etc.).

Three main stories

So, $y_1, \ldots, y_n$ from F; estimand $\theta = \theta(F)$ identified via $\mathrm{E}_F\, m(Y, \theta) = 0$. The (frequentist) nonparametric estimator $\widehat\theta$ solves

$$\int m(y, \widehat\theta)\,\mathrm{d}F_n(y) = n^{-1} \sum_{i=1}^{n} m(y_i, \widehat\theta) = 0.$$

I'll tell three stories:

- Basic frequentist story: $\sqrt{n}(\widehat\theta - \theta_0) \to_d Z$.
- Basic Dirichlet process story: with $\theta \,|\, \mathrm{data}$ via $F \,|\, \mathrm{data}$ and $\theta(F)$: $\sqrt{n}(\theta - \widehat\theta) \,|\, \mathrm{data} \to_d Z$; a Bernshteĭn–von Mises (BvM) theorem.
- Basic EL story: with $\theta \,|\, \mathrm{data}$ via $\pi(\theta)\,\mathrm{EL}_n(\theta)$, $\sqrt{n}(\theta - \widehat\theta) \,|\, \mathrm{data} \to_d Z$; another BvM theorem.

Notate bene: the same limit variable Z for each story.

Plan

A Story A: frequentist estimator and inference;
$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d Z$.

B Story B: starting with $F \sim \mathrm{Dir}(aF_0)$, there is a well-defined $\theta(F) \,|\, \mathrm{data}$, and a BvM theorem: $\sqrt{n}(\theta - \widehat{\theta}) \,|\, \mathrm{data} \to_d Z$.

C Story C: starting with just a prior on θ, and employing $\mathrm{EL}_n(\theta)$, bypassing the 'specify a full prior on F' step, we can study $\pi_n(\theta) \propto \pi(\theta)\,\mathrm{EL}_n(\theta)$, and it works: $\sqrt{n}(\theta - \widehat{\theta}) \,|\, \mathrm{data} \to_d Z$.

D Concluding remarks

Story A is 'classic frequentist'.

Story B is 'kosher': prior and posterior for $\theta(F)$ via full prior and posterior for F.

Story C is different, and bypasses the prior for F step. It's conceptually much simpler (give a prior only for the part you care about!).

Story A: Frequentist estimator $\widehat\theta$ and its behaviour

Consider the $p \times 1$ random function

$$U_n(\theta) = \int m(y,\theta)\,\mathrm{d}F_n(y) = n^{-1}\sum_{i=1}^n m(y_i,\theta).$$

Again: $\widehat\theta$ solves $U_n(\theta) = 0$, and aims at θ_0, solving $\int m(y,\theta)\,\mathrm{d}F_{\mathrm{true}}(y) = 0$.

Application 1: $m(y,\theta) = h(y) - \theta$: then $\widehat\theta = \bar h = n^{-1}\sum_{i=1}^n h(y_i)$.

Application 2: $m_j(y,\theta_j) = I\{y \le \theta_j\} - q_j$: then $\widehat\theta_j = F_n^{-1}(q_j)$, empirical quantiles.

Application 3: $m(y,\theta)$ the score function from given parametric family: then $\widehat\theta$ is the ML estimator.

There's a Theorem A saying that $\sqrt{n}(\widehat\theta - \theta_0) \to_d Z$, a zero-mean multinormal.

(still: limit distribution for $\widehat{\theta}$)

In a bit of detail: First, $\sqrt{n}U_n(\theta_0) \to_d U \sim \mathrm{N}_p(0, K)$, by CLT. If $m^*(y, \theta) = \partial m(y, \theta)/\partial\theta$ exists, we have

$$J_n(\theta_0) = -n^{-1} \sum_{i=1}^{n} m^*(y_i, \theta_0) \to_{\mathrm{pr}} J.$$

Theorem A: Under reasonable regularity conditions,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d Z = J^{-1}U \sim \mathrm{N}_p(0, J^{-1}KJ^{-1}).$$

Works fine for mean functionals (and smooth functions thereof); not directly for quantiles, but with some additional efforts the result holds there too.

The basic technical ingredient is that the functional $\theta(F)$, defined as solution to $\int m(y, \theta)\,\mathrm{d}F(y) = 0$, is Hadamard differentiable, with an influence function. It will have the form $J^{-1}m(y, \theta_0)$.

Story B: Dirichlet process prior

Among the standard nonparametric priors for a cdf is the Dirichlet process. We have

$$F \sim \mathrm{Dir}(aF_0) \Longrightarrow F \,|\, \mathrm{data} \sim \mathrm{Dir}(aF_0 + \sum_{i=1}^{n} \delta(y_i)),$$

and the data take over in $aF_0 + nF_n$. There's a functional BvM theorem here: if data are iid from F_{true}, then

$$\sqrt{n}(F_n - F_{\mathrm{true}}) \to_d W^0(F_{\mathrm{true}}(\cdot)),$$
$$\sqrt{n}(F - F_n) \,|\, \mathrm{data} \to_d W^0(F_{\mathrm{true}}(\cdot)).$$

Incidentally, a very special case of this is the classical connection

$$\sqrt{n}(\widehat{p} - p_{\mathrm{true}}) \to_d \mathrm{N}(0, p_{\mathrm{true}}(1 - p_{\mathrm{true}})),$$
$$\sqrt{n}(p - \widehat{p}) \,|\, \mathrm{data} \to_d \mathrm{N}(0, p_{\mathrm{true}}(1 - p_{\mathrm{true}})),$$

with both binomial and Beta tending to the same normal; cf. the first results in such directions, by Bernshteĭn (1917) and von Mises (1931).

The functional BvM theorem above leads with $\widehat{\theta} = \theta(F_n)$ and $\theta = \theta(F)$ and the functional delta method to this result:

Theorem B: Assume data $y_1, \ldots, y_n$ are iid from $F_{\rm true}$. As long as $\theta = \theta(F)$ is Hadamard smooth (whence having an influence function, which will have the form $J^{-1}m(y, \theta_0)$),

$$\sqrt{n}(\widehat{\theta} - \theta_{\rm true}) \to_d Z,$$
$$\sqrt{n}(\theta - \widehat{\theta}) \,|\, {\rm data} \to_d Z,$$

with Z the inherited delta method functional operating on the limit process $W^0(F_{\rm true}(\cdot))$.

So, the Dirichlet process prior is very fine: the nonparametric Bayesian agrees with the nonparametric frequentist, for large n.

In order for an echo of the prior $F \sim {\rm Dir}(aF_0)$ to be audible, one needs $a \doteq c\sqrt{n}$ (giving a certain bias, but no change in variance) or $a \doteq cn$ (changing also the variance).

Connections to bootstrapping

In practice, one doesn't work out the exact distribution of $\theta(F) \mid \mathrm{data}$, but arrives at the posterior via easy sampling $F \sim \mathrm{Dir}(aF_0 + nF_n)$.

When a small and/or n moderate: very close to drawing $\theta(F)$ via F having weights

$$(p_1, \dots, p_n) \sim \mathrm{Dir}(1, \dots, 1)$$

at the observed data points $y_1, \dots, y_n$ – which is Bayesian bootstrapping.

There is also a natural informative Bayesian bootstrap, which samples from $F \sim \mathrm{Dir}(aF_0 + nF_n)$ rather than from $F \sim \mathrm{Dir}(\delta(x_1) + \cdots + \delta(x_n))$.

These Bayesian bootstrapping schemes are also close enough cousins to Efron's (1979) classic (frequentist) nonparametric bootstrapping to make them large-sample equivalent.

Tentative argument, or point of view: A Bayesian using a 'fully nonparametric prior' for F which does not satisfy the BvM theorem, for smooth functionals $\theta = \theta(F)$, is a determined / strange / insistent / idiosyncratic / individualistic nonparametric Bayesian.

Yes, I might say this in Supreme Court of Statistics – but it depends on the intentions (built-in or implied) of the priors in question. 'Fully nonparametric' (starting from data iid from F, only) is different from 'nonparametric, but with additional constraints or desiderata'.

Of course $\exists$ lots o' clever and valuable priors for F that do not lead to BvM-approved posteriors. But these (typically) employ something extra in their constructions (whether explicitly stated or not).

Story C: Doing Bayes with the Empirical Likelihood

For data $y_1, \ldots, y_n$, with focus on $\theta(F)$ defined as solution to $\int m(y, \theta)\,\mathrm{d}F(y) = 0$, the empirical likelihood $\mathrm{EL}_n(\theta)$ is

$$\max\Big\{\prod_{i=1}^n (nw_i)\colon\ \sum_{i=1}^n w_i = 1, \mathrm{each}\ w_i > 0, \sum_{i=1}^n w_i m(y_i, \theta) = 0\Big\}$$

(Art Owen, Stanford statistics seminar, October 1985).

The Basic EL Theorem says that with $\mathrm{EL}_n(\theta) = \exp\{-\frac{1}{2}A_n(\theta)\}$, then $A_n(\theta_0) \to_d \chi^2_p$, at true value $\theta_0 = \theta(F_{\mathrm{true}})$ (quite a bit more: Hjort, McKeague, Van Keilegom, Annals 2009).

The Bold Nonparametric Bayesian proposal is to bypass setting up a full prior for F, and go directly to

$$\pi_n(\theta) = (1/k)\pi_0(\theta)\,\mathrm{EL}_n(\theta) = \frac{\pi_0(\theta)\exp\{-\frac{1}{2}A_n(\theta)\}}{\int \pi_0(\theta')\exp\{-\frac{1}{2}A_n(\theta')\}\,\mathrm{d}\theta'}.$$

So, the proposal is to go from prior $\pi_0(\theta)$ (for $\theta = \theta(F)$ alone, no need to go via F) to the pseudo-posterior $\pi_n(\theta) \propto \pi_0(\theta)\,\mathrm{EL}_n(\theta)$.

(a) It's not kosher.

(b) It's not clear if it works.

(c) But I'll demonstrate that it does – in the BvM sense. For $\sqrt{n}(\theta - \widehat{\theta})$, given data, there is for moderate to large n no significant difference between
- genuine posterior from Dirichlet process prior;
- pseudo-posterior from EL;
- Bayesian (non-informative or informative) bootstrapping.

(d) Is this enough?

Consider the pseudo-posterior, proportional to

$$\pi_0(\theta)\,\mathrm{EL}_n(\theta) = \pi_0(\theta)\exp\{-\tfrac{1}{2}A_n(\theta)\},$$

where $A_n(\theta) = -2\log \mathrm{EL}_n(\theta)$. The basic start theorem about EL technology is the nonparametric Wilks theorem,

$$A_n(\theta_0) \to_d \chi^2_p \quad \text{at } \theta_0 = \theta(F_{\rm true}).$$

This is (already) splendid, and enough to do testing, confidence regions, nonparametric confidence curves (as in Schweder and Hjort, *Confidence, Likelihood, Probability*, 2016), etc.

For analysing the $\pi_n(\theta)$ we need more, however. The posterior density of $Z_n = \sqrt{n}(\theta - \widehat\theta)$ is

$$q_n(z) \propto \pi_0(\widehat\theta + z/\sqrt{n})\exp\{-\tfrac{1}{2}A_n(\widehat\theta + z/\sqrt{n})\}.$$

Recall, from Story B: with the bona fide BNP approach, as with the Dirichlet:

$$\sqrt{n}\{\theta - \theta(F_n)\} \,|\, \mathrm{data} \to_d Z = J^{-1}U \sim \mathrm{N}_p(0, J^{-1}KJ^{-1}).$$

So for the posterior of the EL based $Z_n = \sqrt{n}(\theta - \widehat{\theta})$ we should hope for

$$\begin{aligned} q_n(z) \;&\propto\; \pi_0(\widehat{\theta} + z/\sqrt{n}) \exp\{-A_n(\widehat{\theta} + z/\sqrt{n})\} \quad \text{[this we know]} \\ &\to_d\; c\exp(-\tfrac{1}{2} z^{\mathrm{t}}\, JK^{-1}J\, z). \quad \text{[this is the hope]} \end{aligned}$$

Indeed, for $A_n(\theta) = -2\log \mathrm{EL}_n(\theta)$, there's a Theorem C saying that under decent conditions,

$$A_n(\widehat{\theta} + z/\sqrt{n}) \to_{\mathrm{pr}} z^{\mathrm{t}}\, JK^{-1}J\, z.$$

A tougher version, with L_1 convergence: with probability 1,

$$\int |q_n(z) - q_0(z)| \,\mathrm{d}z \to_{\mathrm{pr}} 0$$

where $q_0(z) \propto \exp(-\frac{1}{2} z^{\mathrm{t}}\, JK^{-1}J\, z)$ is the density of $\mathrm{N}_p(0, J^{-1}KJ^{-1})$.

Essence of proof for Theorem C

So: with $A_n(\theta) = -2\log \mathrm{EL}_n(\theta)$, where the 'usual EL theorems' are about $A_n(\theta_0)$, we now need to understand $A_n(\widehat\theta + z/\sqrt{n})$. We can prove that as long as $\|\theta - \theta_0\| \le c/\sqrt{n}$,

$$A_n(\theta) = V_n(\theta)^{\mathrm{t}} K_n(\theta)^{-1} V_n(\theta) + \varepsilon_n(\theta),$$

with

$$V_n(\theta) = n^{-1/2}\sum_{i=1}^n m(y_i,\theta),$$

$$K_n(\theta) = n^{-1}\sum_{i=1}^n m(y_i,\theta)m(y_i,\theta)^{\mathrm{t}},$$

and $\varepsilon_n(\theta)$ uniformly to zero in probability (HMV, Annals 2009, but used there for different purposes). Hence

$$A_n(\widehat\theta+z/\sqrt{n}) = V_n(\widehat\theta+z/\sqrt{n})^{\mathrm{t}} K_n(\widehat\theta+z/\sqrt{n})^{-1} V_n(\widehat\theta+z/\sqrt{n}) + o_{\mathrm{pr}}(1).$$

Recall, detail from Story A:

$$J_n(\theta_0) = -n^{-1} \sum_{i=1}^{n} m^*(y_i, \theta_0) \to_{\rm pr} J.$$

First lemma: Under mild conditions,

$$V_n(\widehat\theta + z/\sqrt{n}) = n^{-1/2} \sum_{i=1}^{n} m(y_i, \widehat\theta + z/\sqrt{n}) \to_{\rm pr} -Jz.$$

Second lemma: Under mild conditions,

$$K_n(\widehat\theta + z/\sqrt{n}) = n^{-1} \sum_{i=1}^{n} m(y_i, \widehat\theta + z/\sqrt{n}) m(y_i, \widehat\theta + z/\sqrt{n})^{\rm t} \to_{\rm pr} K.$$

So we're done (under mild conditions):

$$A_n(\widehat\theta + z/\sqrt{n}) \to_{\rm pr} z^{\rm t} J K^{-1} J\, z.$$

The L_1 convergence involves further details (uniformity over compacta, low probability far away from home).

Conclusions and remarks

I've told three stories, with Theorem A (classic nonparametric frequentist); Theorem B (the Dirichlet process satisfies the Bernshteĭn–von Mises stamp of approval); Theorem C (the EL approach works, with precisely the same stamp of approval).

The basic start idea, to dare to use $\pi_n(\theta) \propto \pi(\theta)\,\mathrm{EL}_n(\theta)$ is mentioned in Owen (2001) and to some moderate extent worked with in Lazar (Biometrika 2003) – essentially in the restricted context of a 1-dimensional mean and without a very clear result.

Is it kosher, coherent, bona fide? Strict Bayes would still require a full prior for F, before we can put up a clear $\pi(\theta(F)\,|\,\mathrm{data})$. But $\mathrm{EL}_n(\theta)$ can be seen as the data likelihood put through a least favourable family. And it works – your Le Monde and L'Humanité readers won't see the difference between two analyses, one with full Dirichlet, the other with EL.

Methods and results of my talk can be extended to regression models, survival data models, etc.

There are even Theorems D and E, saying that profiling works (for frequentist EL) and integrating out works (for Bayes EL). Art Owen seminar, Stanford, October 1985, showed a χ^2_p theorem for means. But what about e.g. $\sigma = \mathrm{stdev}\, Y$? There is no single estimating equation for σ.

One solution (as I suggested at the seminar, after Art's talk): Do $\mathrm{EL}_n(\mu_1, \mu_2)$ for 1st and 2nd moments, then do profiling to get EL intervals for $\sigma = (\mu_2 - \mu_1^2)^{1/2}$. Theorem D says this works.

Back to Paris, June 2017: Can do

$$\pi_n(\mu_1, \mu_2) \propto \pi_0(\mu_1, \mu_2)\, \mathrm{EL}_n(\mu_1, \mu_2)$$

and then read of $\pi(\sigma \,|\, \mathrm{data})$. Theorem E says this works in the Bernshteĭn–von Mises sense.

Another talk (another time)

Thanks for attentively listening to Nils Talk X, about Stories A, B, C, with Theorems A, B, C (and D and E). I can also give a Nils Talk X', about Stories A', B', C', with Theorems A', B', C' – about survival data with censoring.

Instead of cdf F and empirical cdf F_n and the Dirichlet, this will be about cumulative hazard function A, the Nelson–Aalen estimator A_n, and the Beta process. Also, the limits involve a $W(\sigma(t)^2)$, a time-scaled Brownian motion, rather than a Brownian bridge.

Efron's (1979) bootstrap would here be replaced by the weird bootstrap, and the Dirichlet process dictated posterior sampling with that of the Beta process. There is a hazard rate world analogue of Rubin's (1981) bootstrap as well as the Empirical Likelihood – and there are Bernshteĭn–von Mises theorems.

References

Hjort, N.L., McKeague, I.W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics*.

Hjort, N.L., McKeague, I.W., and Van Keilegom, I. (2017). Hybrid combinations of parametric and empirical likelihoods. Submitted [today!, June 28].

Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika*. [Has essentially the same start idea as in this talk (I discovered after preparing it), but tentative result reached only for one-dimensional mean, and with a bit limited discussion.]

Owen, A. (1985). Nonparametric likelihood ratio intervals. Technical report, Laboratory for Computational Statistics, Department of Statistics, Stanford University.

Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood, Probability.* Cambridge University Press.