Subjective Bayesian and Likelihood-Based Inference as Weapons of Mass Destruction

James Robins

·Counterexamples not examples in the old days

·The world has changed from when computation was not possible

·This talk is a transcription of the Glen Beck show of Nov 14, 2013.

·Dateline: Washington DC, Federal Drug Administration, Nov 14, 2013

·Today President Obama named AWF Edwards, aka "the likelihood kid", and James, aka "Jimmie the Bayesian ", Savage as special FDA czars to chair the **special panel on the ongoing crisis in the analysis of randomized drug trials.**

·With their appointment, the death panels we warned you, the American people, about back in 2009 have been realized.

·However, in 2009, we dramtically underestimated the number of Americans that would be sent to their deaths by the president's panels.

·With the passage of the personalized-medicine equipoise bill of Nov. 12, 2012, likelihoodism and subjective Bayesianism have been transformed from relatively harmless academic curiousities into weapons of mass destruction

·The FDA death panel will murder thousands of innocent Americans each and every year.

· We must overthrow the czars and replace with either

(i) frequentist with knowledge of "locally semiparametric efficient estimation "

(ii) or its fellow travelers - Bayesians engaged in frequentist pursuit

s

Some background:

·The personalized-medicine equipoise bill of Nov. 12, 2011 required all drug companies who run randomized trials

(1) to collect a data on all relevant baseline predictors X of the response Y that might influence whether therapy was good for you and

(2) randomize to binary experimental treatment D with with probability (propensity score)

$$\rho\left(X\right)=pr\left(D=1|X\right)$$

based FDA expert prior opinion as to how likely the treatment D=1 was to be advantageous vs disadvantageous at joint covariate level X compared to standard treatment $D = 0$, subject to $0.1 \leq \pi(X) \leq 0.9$ (so we can explore).

·The baseline predictors must include the patients last ten years of laboratory tests and the MRA levels of 20,000 genes

·Due to the major world financial crisis that began in August 2013, it no longer was possible to measure any of these X variables on patients before selecting treatmentnor was it possible to finance any more trials.

·Fortunately between Nov. 12, 2011 and August 2013, 50,00 trials of all sorts of drugs had been run using the $2 trillion a year allocated by the Obama administration to discover clinical effective treatments.

·Thus for each of the 50,000 trials, the relevant question is whether on average the effect of treatment was beneficial or harmful. That is we wanted to estimate for each trial the average treatment effect ψ

$$\psi = E\left[Y_1\right] - E[Y_0]$$

and

·license the drug if $E\,[Y_1] - E[Y_0] > 0$ and not otherwise

· could use a small pos number

·Rough estimates were that 10,000 of the treatments are beneficial, 20,000 harmful, and 20,000 make no difference.

·How should we estimate the trial specific effect ψ?

·AWF says use the likelihood;

·Jimmie the Bayesian says use Bayes with subjective prior

as he believes "There is no Bayes but Bayes"

·They will cause many deaths but there is a simple approach that will do well

Simplifications

1.Nobody is interested in combing results across trials as the 50,000 conditions were too different.

2. Those entering the trials for a drug can be considered a random sample of those elgible so selction effects are to be ignored.

3. individuals in a trial can be assumed iid given some parameters (DiFinetti if you like)

· Following close relation to

Godambe and Thompson’s (1976) criticism of likelihood based inference in the context of finite population inference from sample survey data

· “ancillarity paradoxes” of Brown (1990) and Foster and George (1996)

The model for a given trial

n iid copies of $O = (Y, D, X)$ with n say about 1000.

Parameter of interest ψ identified by randomization of D within levels of X :

$$\begin{aligned}
E\left[Y_1 - Y_0 | X\right] &= E[Y|D=1, X] - E[Y|D=0, X] = \psi\left(X\right) \\
\psi &= \int \psi\left(X\right) f\left(X\right) dX = E\left[Y_1\right] - E\left[Y_0\right] \\
E\left[Y_1\right] &= \int E[Y|D=1, X] f\left(X\right) dX, \\
E\left[Y_0\right] &= \int E[Y|D=0, X] f\left(X\right) dX,
\end{aligned}$$

·Likelihood:

$$\begin{aligned}
& \mathcal{L}_1(\theta)\,\mathcal{L}_2(\gamma) \\
\mathcal{L}_1(\theta) &= \prod_{i=1}^{n} L_{1i}(\theta) \\
L_1(\theta) &= f(Y|D,X;\theta_1)\, f(X;\theta_2)
\end{aligned}$$

·$f(Y|D,X;\theta_1)$ is the set of all conditonal densities as θ_1 varies over Θ_1.

·$f(X;\theta_2)$ is the set of all densities as θ_2 varies over Θ_2, the set of densities dominated by Lesbegue measure on R^p.

$$\begin{aligned} & \mathcal{L}_1(\theta)\,\mathcal{L}_2(\gamma) \\ \mathcal{L}_2(\gamma) &= \prod_{i=1}^{n} L_{2i}(\gamma) \\ L_2 &= \rho(X;\gamma)^D \{1-\rho(X;\gamma)\}^{1-D} \\ \rho(X;\gamma) &= pr(D=1|X;\gamma) \end{aligned}$$

·$\rho(x;\gamma)$ are all functions of x bounded between .1 and .9 as γ varies over Γ

·Our goal is to estimate the parameter

$$\psi(\theta^*) = \int \{E[Y|D=1,X;\theta_1^*] - E[Y|D=0,X;\theta_1^*\}\, f(X;\theta_2^*)\, dX.$$

·$\psi(\theta)$ only is a function of parameters occurring in the $\mathcal{L}_1(\theta)$part of the likelihood.

$\cdot$The model is infinite dimensional because the set Θ for example cannot be put into a smooth, one-to-one correspondence with a finite dimensional Euclidean space.

·Assume a Bayesian with independent priors for now $\pi(\theta,\gamma) = \pi(\theta)\pi(\gamma)$ so posterior

$$\pi(\theta|\mathbf{O}) \propto \mathcal{L}_1(\theta)\pi(\theta)$$

·A consequence of prior and likelihood factoring into a $\theta - part$ and a $\gamma - part$ because of prior independence.

·Hence the posterior for θ and any $\psi(\theta)$ is the same whatever be γ [and thus $\rho(x;\gamma)$] and $\pi(\gamma)$ ..

Thus a Bayesian interested in $\psi(\theta^*)$ with independent priors would decline the offer

if offered the true propenisty score $\rho(x;\gamma^*)$, especially if he had to pay

·The same for a pure likelihoodist

· I can force a Bayesian to have independent priors in this example

·That isit is way to insure a Bayesian has independent priors in thsi example.

THEN What Would Bayesian With Independent Priors Do

·Regression of Y on (D, X); X -25,000 dimensional

Allow marginal of X to be unrestricted (not the point not essential)

Posterior (and its mean and median) will

(i) generally concentrate around a value different from $\psi(\theta^*)$

or

(ii) not concentrate and be too wide to be useful for a decision licensing

· Thousands of Deaths will occur as many bad drugs licensed and good drugs rejected.

·The same for a pure-likelihoodist

·**Nonparametric Bayes TOO WIDE as likelihood flat**: Size of space

$\{f\,(y|D=d,X=x;\theta_1)\}$ too large

·**Parametric Bayes**: Misspecification unless model

$E[Y|D,X;\theta_1]=\theta_{11}^T m\,(X)+\theta_{10}^T Dm\,(X)$ correct (essentially never).

Will not concentrate around $\psi\,(\theta^*)$

·

· Smooth or sparse priors will be wrong and not concentrate around $\psi\,(\theta^*)$

What Would Frequentist Do: Horvitz-Thompson

$$
\begin{aligned}
V &= DY/\rho(X;\gamma^{*}) - \{(1-D)\,Y\}\,/\,\{1-\rho(X;\gamma^{*})\} \\
\widehat{\psi}_{HT} &= P_n\,[V] \equiv n^{-1}\sum_{i=1}^{n} V_i
\end{aligned}
$$

· $\widehat{\psi}_{HT}$ is uniformly $\sqrt{n}$-consistent and provides for valid asymptotic and finite sample CI

·$\widehat{\psi}_{HT}$ inefficient because does not use data on X. Will improve it later.

Math: Theorem (Robins and Ritov, 1997): When γ^* is unknown or not used in an estimator, no uniformly consistent estimator of $\psi(\theta^*)$ exists. That is, at no finite sample size, can we have an estimator that is close to $\psi(\theta)$ over all laws (θ, γ) in $\Theta \times \Gamma$. In fact true in any neighborhood a a given (θ^*, γ^*).

·**Corollary**: ·Standard likelihood-based and Bayesian estimators methods (with ind priors) will fail to be uniformly consistent.

·Uniformity is important because it links asymptotic behavior to finite sample behavior.

·**Finite Samples** The deficiencies in Bayesian methods are not only in large samples..

Any interval estimator [eg a highest 1-α posterior interval] that is not a function of γ^* will not be "valid"

By valid we mean that under all $\theta \times \gamma \in \Theta \times \Gamma$

(i) the frequentist coverage is at least $(1 - \alpha)$ at each sample size n and

(ii) the expected length goes to zero with increasing sample size.

· HT is a valid interval estimators for ψ^* whose length shrinks at rate but these depend on γ^* and hence are not SFB and they violate LP.

·Note if γ^* is known (as in our randomized trials) and used, there do exist uniformly valid 1-α CI

·An example of such interval estimators are $\widehat{\psi}_{HT} \pm c_{HT} n^{-1/2}$ and $\widehat{\psi}_{loc,eff} \pm c_{loc,eff} n^{-1/2}$where the $c's$ can be explicitly calculated using Chebychev's inequality.

∘This interval is not only valid, but its length shrinks at rate $n^{-1/2}$

Locally Efficient Semiparametric Estimators:

· Begin with a finite parametric for $E[Y|D,X]$ such as

$$\sum_{m=1}^{k} \eta_{0,m} W_m(X) + \eta_{1,m} D \eta_{1,m} W_m(X)$$

·Add to the model 2 magic covariates

$$\begin{aligned} & h(X,D;\eta,\varphi) \\ = \ & \sum_{m=1}^{k} \eta_{0,m} W_m(X) + \eta_{1,m} D \eta_{1,m} W_m(X) \\ & + \varphi_1 \frac{1-D}{1-\rho(X;\gamma^*)} + \frac{D}{\rho(X;\gamma^*)} \end{aligned}$$

·Fit by OLS.

· Compute

$$\begin{aligned} & \widehat{\psi}_{loc,eff} \\ = \; & \widehat{E}[Y_1] - \widehat{E}[Y_0] \\ = \; & P_n[h(X, D = 1; \widehat{\eta}, \widehat{\varphi})] - P_n[h(X, D = 0; \widehat{\eta}, \widehat{\varphi}) \end{aligned}$$

·Like $\widehat{\psi}_{HT}$, $\widehat{\psi}_{loc,eff}$ $n^{1/2}-$consistent even if model totally misspecified.

· It is semiparmatric efficient if the model $h(X, D; \eta, \varphi)$ correct which it is not

· Closer to correct more efficient.

· In obs study use $\rho(X; \widehat{\gamma})$ used. Estimator is doubly robust.

·**Bayes Locally Efficient Semiparametric Estimators:**

·Place noninformative prior on η, φ and use known γ^*

·Assume normal errors with non-informmative prior on the variance

·The posterior mean has same large sample distribution as $\widehat{\psi}_{loc,eff}$

· Conclusion: By using carefully tuned dependent priors on (γ, θ)

have Bayes estimator mimicking a locally semiparametric efficient frequentist estimator.

·But is this a Pyrrhic victory?

If we need to engineer the dependent prior just to mimic a frequentist answer, is it really Bayesian inference?

What is going on:

· Likelihood in NP model too complex for data.

· Need to throw away data but true Bayes cannot

· Christian Robert response: the *curse of marginalization*: "the classical Bayesian approach is an holistic system that cannot remove information to process a subset of the original problem." ie a functiona

· But more: Parameter of ancillary statistic known needs to be incorporated

Not Anti-Bayes

·Bayes inference that maps prior beliefs to posterior beliefs via the likelihood function, without regard for the frequentist properties can be informative.

·Second, when modelling complex phenomena (particularly in small and moderate samples), there are be Bayesian approaches well motivated and easy to implement even when there is no good frequentist alternative,

·the Bayes estimator is the best, or perhaps the only, frequentist game in town.

·Third, to improve decision making under uncertainty, one can adopt a Bayes-frequentist compromise (

·combines honest subjective Bayesian inference with good frequentist behavior even when, the model is so large and the likelihood function so complex that standard (uncompromised) Bayes procedures have poor frequentist performance.

·However such a compromise requires that our subjective Bayesian decision maker is only allowed to observe a specified vector function of X (depending on $\rho(X;\gamma^*)$) but not X itself. In this way one can circumvent the problem referred to by Christian Robert

·**What Could a Subjective Bayesian Do With Independent Priors:**

·Odysseus tied the mast **because he want to hear the enchanting sirens but does not want to his overwhelming desire to go to them result in his being torn apart.**

·Analogy:Subjective Bayesian is Odysseus

·Sirens the likelihood ass with the full Data:

all 25000 covariates an overwhelming call as for Bayesian more data better inference.

·Crew level headed ; do not hear the call of the full data likelihood.

·$Only$ offer Odysseus

$$\begin{aligned} U(\psi) &= \left\{\widehat{\psi}_{par,Bayes} - \psi\right\} / se_{bs}\left(\widehat{\psi}_{par,Bayes} - \psi\right) \\ & U(\psi) \text{ dis} N(\psi, 1) \end{aligned}$$

·Tied to the mast by the crew

How to Guarantee Independent Priors

Bayesian queries the randomizer who selected $\rho(X;\gamma^*)$ about his reasoned opinions concerning $f\,(Y|D,X;\theta_1)$(but not about $\rho(X;\gamma^*)$) until he gets no more info

· He updates his prior on θ_1

·The Bayesian will then have independent priors.

· The updated prior will not change if you learn now what $\rho(X;\gamma^*)$ he chose (if not gaming you) as no more info in randomizer about $f\,(Y|D,X;\theta_1)$