Can we estimate a constant?

Christian P. Robert

Université Paris-Dauphine, Paris & University of Warwick, Coventry

December 11, 2015

CRiSM workshop, Warwick, 20-22 April, 2016

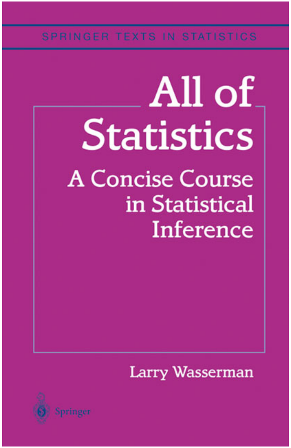
Estimating Constants

- wide range of computational methods for approximating normalising constants
- wide range of communities
- novel challenges associated with large data and highly complex models
- 13 talks, plus two poster sessions

[England in April, where else...?!]

Larry’s constant

1. Larry’s constant
2. Charlie’s logistic regression
3. Xiao-Li’s MLE
4. Larry’s and Jamie’s paradox



SPRINGER TEXTS IN STATISTICS

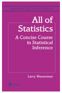
All of Statistics

A Concise Course in Statistical Inference

Larry Wasserman

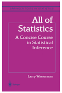
Springer

Example 11.10



"Suppose that f is a probability density function and that $f(x) = cg(x)$, where g is a known function and $c > 0$ is unknown. In principle, we can compute c since $\int f(x)dx = 1$ implies that $c = 1/\int g(x)dx$. But in many cases we can't do the integral $\int g(x)dx$ since g might be a complicated function and x could be high-dimensional. Despite the fact that c is unknown, it is often possible to draw $X_1, \ldots, X_n$ from f; see Chapter 24. Can we use the sample to estimate the normalizing constant c? Here is a frequentist solution: Let $\widehat{f}$ be a consistent estimate of the density f. Choose any point x and note that $c = f(x)/g(x)$. Hence $\hat{c} = \widehat{f}(x)/g(x)$ is a consistent estimate of c."

Example 11.10: “the” argument


All of Statistics
A Concise Course in Statistical Inference
Larry Wasserman

“Now let us try to solve the problem from a Bayesian approach. Let $\pi(c)$ *be a prior such that* $\pi(c) > 0$ *for all* $c > 0$*. The likelihood function is*

$$\mathcal{L}_n(c) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} c g(X_i) = c^n \prod_{i=1}^{n} g(X_i) \propto c^n \,.$$

Hence the posterior is proportional to $c^n \pi(c)$*. The posterior does not depend on* $X_1, \ldots, X_n$*, so we come to the startling conclusion that from the Bayesian point of view, there is no information in the data about* c*.”*

["Example 11.10 is due to Ed George (personal communication)"]

“Bayesians are slaves to the likelihood function”

- likelihood function $\mathcal{L}_n(c)$? which likelihood function?! [moving c does not modify the sample]
- “there is no information in the data about c”: right! Absolutely none whatsoever
- this is not a statistical problem, rather a numerical problem with many Monte Carlo solutions
- Monte Carlo methods are frequentist (LLN) and asymptotical (as in large numbers) [not an issue]

Is there any meaning in bringing a Bayesian flavour into the Monte Carlo dishes?

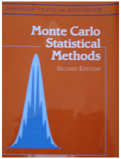
a broader picture

- Larry's problem somehow relates to the infamous harmonic mean estimator issue [see later?]
- highlight paradoxical differences between statistics and Monte Carlo methods:
    - statistics constrained by sample and its distribution
    - Monte Carlo free to generate samples
    - no best unbiased estimator or optimal solution in Monte Carlo integration
- paradox of the fascinating "Bernoulli factory" problem, which requires infinite sequence of Bernoullis [see later]

  [Flegal & Herbei, 2012; Jacob & Thiery, 2015]
- highly limited range of parameters allowing for unbiased estimation versus universal debiasing of converging sequences

  [McLeish, 2011; Rhee & Glynn, 2012, 2013]

Exercice 3.33

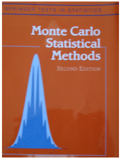


Accept–reject raw outcome: i.i.d. sequences $Y_1, Y_2, \ldots, Y_t \sim g$ and $U_1, U_2, \ldots, U_t \sim \mathcal{U}(0,1)$

Random number of accepted Y_i's

$$\mathbb{P}(N = n) = \binom{n-1}{t-1} \left(1/M\right)^t \left(1 - 1/M\right)^{n-t}$$

Exercice 3.33



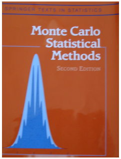
Accept–reject raw outcome: i.i.d. sequences $Y_1, Y_2, \ldots, Y_t \sim g$ and $U_1, U_2, \ldots, U_t \sim \mathcal{U}(0,1)$

Joint density of $(N, \mathbf{Y}, \mathbf{U})$

$$\begin{aligned}
&\mathbb{P}(N = n, Y_1 \le y_1, \ldots, Y_n \le y_n, U_1 \le u_1, \ldots, U_n \le u_n) \\
&\quad = \int_{-\infty}^{y_n} g(t_n)(u_n \wedge w_n) dt_n \int_{-\infty}^{y_1} \ldots \int_{-\infty}^{y_{n-1}} g(t_1) \ldots g(t_{n-1}) \\
&\quad \times \sum_{(i_1,\cdots,i_{t-1})} \prod_{j=1}^{t-1} (w_{i_j} \wedge u_{i_j}) \prod_{j=t}^{n-1} (u_{i_j} - w_{i_j})^+ dt_1 \cdots dt_{n-1},
\end{aligned}$$

where $w_i = f(y_i)/Mg(y_i)$ and sum over all subsets of $\{1, \ldots, n-1\}$ of size $t-1$

Exercice 3.33



Accept–reject raw outcome: i.i.d. sequences $Y_1, Y_2, \dots, Y_t \sim g$ and $U_1, U_2, \dots, U_t \sim \mathcal{U}(0,1)$

Marginal joint density of $(Y_i, U_i)|N = n,\ i < n$

$$\begin{aligned}
&\mathbb{P}(N = n, Y_1 \leq y, U_1 \leq u_1) \\
&\quad = \binom{n-1}{t-1} \left(\frac{1}{M}\right)^{t-1} \left(1 - \frac{1}{M}\right)^{n-t-1} \\
&\quad \times \left[\frac{t-1}{n-1}(w_1 \wedge u_1)\left(1 - \frac{1}{M}\right) + \frac{n-t}{n-1}(u_1 - w_1)^+ \left(\frac{1}{M}\right)\right] \int_{-\infty}^{y} g(t_1)dt_1
\end{aligned}$$

much ado about noise

Accept-reject sample $(X_1, \ldots, X_m)$ associated with $(U_1, \ldots, U_N)$ and $(Y_1, \ldots, Y_N)$

N is stopping time for acceptance of m variables among Y_j's

Rewrite estimator of $\mathbb{E}[h]$ as

$$\frac{1}{m} \sum_{i=1}^{m} h(X_i) = \frac{1}{m} \sum_{j=1}^{N} h(Y_j)\, \mathbb{I}_{U_j \leq w_j}\,,$$

with $w_j = {}^{f(Y_j)}\!/\!{}_{Mg(Y_j)}$

[Casella & Robert, 1996]

Rao-Blackwellisation: smaller variance produced by integrating out the U_i's,

$$\frac{1}{m}\sum_{j=1}^{N} \mathbb{E}[\mathbb{I}_{U_j \le w_j} | N, Y_1, \ldots, Y_N]\ h(Y_j) = \frac{1}{m}\sum_{i=1}^{N} \rho_i h(Y_i),$$

where $(i < n)$

$$\begin{aligned}\rho_i &= \mathbb{P}(U_i \le w_i | N = n, Y_1, \ldots, Y_n) \\ &= w_i \frac{\sum_{(i_1,\ldots,i_{m-2})} \prod_{j=1}^{m-2} w_{i_j} \prod_{j=m-1}^{n-2} (1 - w_{i_j})}{\sum_{(i_1,\ldots,i_{m-1})} \prod_{j=1}^{m-1} w_{i_j} \prod_{j=m}^{n-1} (1 - w_{i_j})},\end{aligned}$$

and $\rho_n = 1$.
Numerator sum over all subsets of $\{1, \ldots, i-1, i+1, \ldots, n-1\}$ of size $m-2$, and denominator sum over all subsets of size $m-1$

[Casella & Robert, 1996]

Accepted Metropolis–Hastings proposals

Yet another representation of Metropolis–Hastings estimator δ as

$$\delta = \frac{1}{n}\sum_{t=1}^{n} h(x^{(t)}) = \frac{1}{n}\sum_{i=1}^{M_n} \mathfrak{n}_i h(\mathfrak{z}_i)\,,$$

where

- $(x_t)_t$ original MCMC chain
- $\mathfrak{z}_i$'s are the accepted y_j's
- M_n is the number of accepted y_j's till time n
- $\mathfrak{n}_i$ is the number of times $\mathfrak{z}_i$ appears in the sequence $(x^{(t)})_t$

[Douc & Robert, 2011]

Accepted Metropolis–Hastings proposals

Yet another representation of Metropolis–Hastings estimator δ as

$$\delta = \frac{1}{n}\sum_{t=1}^{n} h(x^{(t)}) = \frac{1}{n}\sum_{i=1}^{M_n} \mathfrak{n}_i h(\mathfrak{z}_i)\,,$$

where

1. $(\mathfrak{z}_i, \mathfrak{n}_i)_i$ is a Markov chain;
2. $\mathfrak{z}_{i+1}$ and $\mathfrak{n}_i$ are independent given $\mathfrak{z}_i$;
3. $\mathfrak{n}_i$ is distributed as a geometric random variable with probability parameter

$$p(\mathfrak{z}_i) := \int \alpha(\mathfrak{z}_i, y)\, q(y|\mathfrak{z}_i)\, dy\,;$$

4. $(\mathfrak{z}_i)_i$ is a Markov chain with transition kernel $\tilde{Q}(\mathfrak{z}, dy) = \tilde{q}(y|\mathfrak{z})dy$ and stationary distribution $\tilde{\pi}$ such that

$$\tilde{q}(\cdot|\mathfrak{z}) \propto \alpha(\mathfrak{z}, \cdot)\, q(\cdot|\mathfrak{z}) \quad \text{and} \quad \tilde{\pi}(\cdot) \propto \pi(\cdot)p(\cdot)\,.$$

[Douc & Robert, 2011]

The Bernoulli factory

Estimate of $1/p(\mathfrak{z}_i)$,

$$\mathfrak{n}_i = 1 + \overbrace{\sum_{j=1}^{\infty} \prod_{\ell \leq j}}^{\text{virtual infinite sum}} \mathbb{I}\{u_\ell \geq \alpha(\mathfrak{z}_i, y_\ell)\} \,,$$

improved by integrating u_ℓ's

$$\hat{\xi}_i = 1 + \sum_{j=1}^{\infty} \prod_{\ell \leq r} \{1 - \alpha(\mathfrak{z}_i, y_\ell)\}$$

- unbiased estimator of $1/p(\mathfrak{z}_i)$
- lower [conditional on $\mathfrak{n}_i$] variance than geometric $\{1 - p(\mathfrak{z}_i)\}/p^2(\mathfrak{z}_i)$
- mileage may vary....

[Douc & Robert, 2011]

reverse logistic regression

1. Larry's constant
2. Charlie's logistic regression
3. Xiao-Li's MLE
4. Larry's and Jamie's paradox

regression estimator

Given

$$X_{ij} \sim f_i(x) = c_i h_i(x)$$

with h_i known and c_i unknown $(i = 1, \ldots, k,\ j = 1, \ldots, n_i)$, constants c_i estimated by a "reverse logistic regression" based on the quasi-likelihood

$$\mathcal{L}(\eta) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log p_i(x_{ij}, \eta)$$

with

$$p_i(x, \eta) = \exp\{\eta_i\} h_i(x) \Big/ \sum_{i=1}^{k} \exp\{\eta_i\} h_i(x)$$

[Anderson, 1972; Geyer, 1992]

Approximation

$$\log \hat{c}_i = \log n_i / n - \hat{\eta}_i$$

regression estimator

Given

$$X_{ij} \sim f_i(x) = c_i h_i(x)$$

with h_i known and c_i unknown $(i = 1, \ldots, k,\ j = 1, \ldots, n_i)$, constants c_i estimated by a “reverse logistic regression” based on the quasi-likelihood

$$\mathcal{L}(\eta) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log p_i(x_{ij}, \eta)$$

with

$$p_i(x, \eta) = \exp\{\eta_i\} h_i(x) \Big/ \sum_{i=1}^{k} \exp\{\eta_i\} h_i(x)$$

[Anderson, 1972; Geyer, 1992]

Approximation

$$\log \hat{c}_i = \log n_i / n - \hat{\eta}_i$$

statistical framework?

Existence of a central limit theorem:

$$\sqrt{n}\,(\hat{\eta}_n - \eta) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, B^+AB)$$

[Geyer, 1992; Doss & Tan, 2015]

- strong convergence properties
- asymptotic approximation of the precision
- connection with bridge sampling and auxiliary model [mixture]
- ...but nothing statistical there [no estimation]
- which optimality? [weights unidentifiable]

[Kong et al., 2003; Chopin & Robert, 2011]

statistical framework?

Existence of a central limit theorem:

$$\sqrt{n}\,(\hat{\eta}_n - \eta) \xrightarrow{\mathcal{L}} \mathcal{N}_k(0, B^{+}AB)$$

[Geyer, 1992; Doss & Tan, 2015]

- strong convergence properties
- asymptotic approximation of the precision
- connection with bridge sampling and auxiliary model [mixture]
- ...but nothing statistical there [no estimation]
- which optimality? [weights unidentifiable]

[Kong et al., 2003; Chopin & Robert, 2011]

evidence from a posterior sample

Use of the identity

$$\begin{aligned}\mathbb{E}\left[\frac{\varphi(\theta)}{\pi(\theta)L(\theta)}\right] &= \int \frac{\varphi(\theta)}{\pi(\theta)L(\theta)} \frac{\pi(\theta)L(\theta)}{\mathfrak{Z}} \,\mathrm{d}\theta \\ &= \frac{1}{\mathfrak{Z}}\end{aligned}$$

no matter what the proposal $\varphi(\theta)$ is.

[Gelfand & Dey, 1994; Bartolucci et al., 2006]

harmonic mean estimator

Constraint opposed to usual importance sampling constraints: $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi(\theta)L(\theta)$ for the approximation

$$\widehat{\mathfrak{Z}_1} = 1 \Bigg/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta^{(t)})}{\pi(\theta^{(t)})L(\theta^{(t)})}$$

to have a finite variance

[Robert & Wraith, 2012]

mixtures as proposals

Design specific mixture for simulation purposes, with density

$$\tilde{\varphi}(\theta) \propto \omega_1 \pi(\theta) L(\theta) + \varphi(\theta)\,,$$

where $\varphi(\theta)$ is arbitrary (but normalised)
Note: ω_1 is not a probability weight

[Chopin & Robert, 2011]

mixtures as proposals

Design specific mixture for simulation purposes, with density

$$\tilde{\varphi}(\theta) \propto \omega_1 \pi(\theta) L(\theta) + \varphi(\theta)\,,$$

where $\varphi(\theta)$ is arbitrary (but normalised)
Note: ω_1 is not a probability weight

[Chopin & Robert, 2011]

evidence approximation by mixtures

Rao-Blackwellised estimate

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) \Big/ \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + \varphi(\theta^{(t)}) \,,$$

converges to $\omega_1 \mathfrak{Z} / \{\omega_1 \mathfrak{Z} + 1\}$
Deduce $\hat{\mathfrak{Z}}$ from

$$\omega_1 \hat{\mathfrak{Z}} / \{\omega_1 \hat{\mathfrak{Z}} + 1\} = \hat{\xi}$$

[Chopin & Robert, 2011]

partition function and maximum likelihood

For parametric family

$$f(x;\theta) = p(x;\theta)/Z(\theta)$$

- normalising constant $Z(\theta)$ also called *partition function*
- ...if normalisation possible
- essential part of inference
- estimation by score matching [matching scores of model and data]
- ...and by noise-contrastive estimation [generalised Charlie's regression]

[Gutmann & Hyvärinen, 2012, 2015]

partition function and maximum likelihood

For parametric family

$$f(x;\theta) = p(x;\theta)/Z(\theta)$$

Generic representation with auxiliary data y from known distribution f_y and regression function

$$h(u;\theta) = \left\{1 + \frac{n_x}{n_y}\exp(-G(u;\theta))\right\}^{-1}$$

Objective function

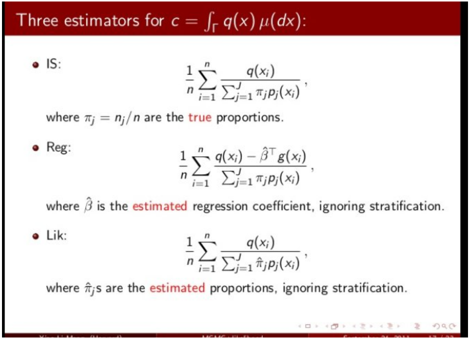
$$J(\theta) = \sum_{i=1}^{n_x} \log h(x_i;\theta) + \sum_{i=1}^{n_y} \log\{1 - h(y_i;\theta)\}$$

that can be maximised with no normalising constant

[Gutmann & Hyvärinen, 2012, 2015]

Xiao-Li's MLE

1. Larry's constant
2. Charlie's logistic regression
3. Xiao-Li's MLE
4. Larry's and Jamie's paradox



Three estimators for $c = \int_\Gamma q(x)\,\mu(dx)$:

- IS: $$\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i)}{\sum_{j=1}^{J}\pi_j p_j(x_i)},$$ where $\pi_j = n_j/n$ are the true proportions.
- Reg: $$\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i) - \hat{\beta}^\top g(x_i)}{\sum_{j=1}^{J}\pi_j p_j(x_i)},$$ where $\hat{\beta}$ is the estimated regression coefficient, ignoring stratification.
- Lik: $$\frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i)}{\sum_{j=1}^{J}\hat{\pi}_j p_j(x_i)},$$ where $\hat{\pi}_j$s are the estimated proportions, ignoring stratification.

[Meng, 2011, IRCEM]

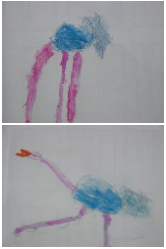
Xiao-Li's MLE

"The task of estimating an integral by Monte Carlo methods is formulated as a statistical model using simulated observations as data.
The difficulty in this exercise is that we ordinarily have at our disposal all of the information required to compute integrals exactly by calculus or numerical integration, but we choose to ignore some of the information for simplicity or computational feasibility."

[Kong, McCullagh, Meng, Nicolae & Tan, 2003]

Xiao-Li's MLE

"Our proposal is to use a semiparametric statistical model that makes explicit what information is ignored and what information is retained. The parameter space in this model is a set of measures on the sample space, which is ordinarily an infinite dimensional object. None-the-less, from simulated data the base-line measure can be estimated by maximum likelihood, and the required integrals computed by a simple formula previously derived by Geyer and by Meng and Wong using entirely different arguments."

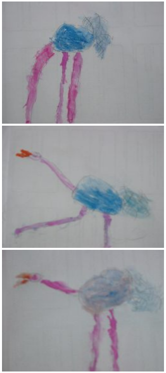


[Kong, McCullagh, Meng, Nicolae & Tan, 2003]

Xiao-Li's MLE

> *"By contrast with Geyer's retrospective likelihood, a correct estimate of simulation error is available directly from the Fisher information. The principal advantage of the semiparametric model is that variance reduction techniques are associated with submodels in which the maximum likelihood estimator in the submodel may have substantially smaller variance than the traditional estimator."*

[Kong, McCullagh, Meng, Nicolae & Tan, 2003]

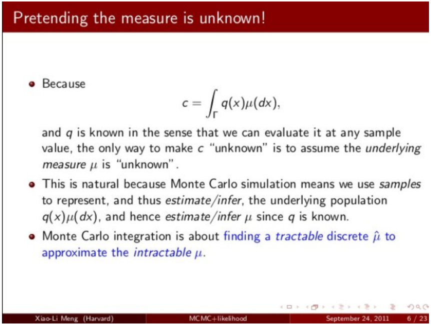


(c.) Rachel^{2002}

Xiao-Li's MLE

"At first glance, the problem appears to be an exercise in calculus or numerical analysis, and not amenable to statistical formulation"

- use of Fisher information
- non-parametric MLE based on simulations
- comparison of sampling schemes through variances
- Rao–Blackwellised improvements by invariance constraints



Pretending the measure is unknown!

- Because
$$c = \int_\Gamma q(x)\mu(dx),$$
and q is known in the sense that we can evaluate it at any sample value, the only way to make c "unknown" is to assume the *underlying measure* μ is "unknown".
- This is natural because Monte Carlo simulation means we use *samples* to represent, and thus *estimate/infer*, the underlying population $q(x)\mu(dx)$, and hence *estimate/infer* μ since q is known.
- Monte Carlo integration is about finding a *tractable* discrete $\hat{\mu}$ to approximate the *intractable* μ.

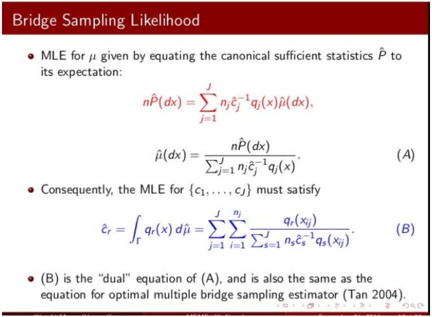
Xiao-Li Meng (Harvard) MCMC+likelihood September 24, 2011 6 / 23

[Meng, 2011, IRCEM]

Xiao-Li's MLE

"At first glance, the problem appears to be an exercise in calculus or numerical analysis, and not amenable to statistical formulation"

- use of Fisher information
- non-parametric MLE based on simulations
- comparison of sampling schemes through variances
- Rao–Blackwellised improvements by invariance constraints



Bridge Sampling Likelihood

- MLE for μ given by equating the canonical sufficient statistics $\hat{P}$ to its expectation:

$$n\hat{P}(dx) = \sum_{j=1}^{J} n_j \hat{c}_j^{-1} q_j(x) \hat{\mu}(dx),$$

$$\hat{\mu}(dx) = \frac{n\hat{P}(dx)}{\sum_{j=1}^{J} n_j \hat{c}_j^{-1} q_j(x)}. \qquad (A)$$

- Consequently, the MLE for $\{c_1, \ldots, c_J\}$ must satisfy

$$\hat{c}_r = \int_{\Gamma} q_r(x)\, d\hat{\mu} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \frac{q_r(x_{ij})}{\sum_{s=1}^{J} n_s \hat{c}_s^{-1} q_s(x_{ij})}. \qquad (B)$$

- (B) is the "dual" equation of (A), and is also the same as the equation for optimal multiple bridge sampling estimator (Tan 2004).

[Meng, 2011, IRCEM]

NPMLE

Observing

$$Y_{ij} \sim F_i(t) = c_i^{-1} \int_{-\infty}^{t} \omega_i(x)\, dF(x)$$

with ω_i known and F unknown

NPMLE

Observing

$$Y_{ij} \sim F_i(t) = c_i^{-1} \int_{-\infty}^{t} \omega_i(x)\, dF(x)$$

with ω_i known and F unknown
“Maximum likelihood estimate” defined by weighted empirical cdf

$$\sum_{i,j} \omega_i(y_{ij}) p(y_{ij}) \delta_{y_{ij}}$$

maximising in p

$$\prod_{ij} c_i^{-1} \omega_i(y_{ij})\, p(y_{ij})$$

NPMLE

Observing

$$Y_{ij} \sim F_i(t) = c_i^{-1} \int_{-\infty}^{t} \omega_i(x)\, dF(x)$$

with ω_i known and F unknown
"Maximum likelihood estimate" defined by weighted empirical cdf

$$\sum_{i,j} \omega_i(y_{ij}) p(y_{ij}) \delta_{y_{ij}}$$

maximising in p

$$\prod_{ij} c_i^{-1} \omega_i(y_{ij})\, p(y_{ij})$$

Result such that

$$\sum_{ij} \frac{\hat{c}_r^{-1} \omega_r(y_{ij})}{\sum_s n_s \hat{c}_s^{-1} \omega_s(y_{ij})} = 1$$

[Vardi, 1985]

NPMLE

Observing

$$Y_{ij} \sim F_i(t) = c_i^{-1} \int_{-\infty}^{t} \omega_i(x)\, dF(x)$$

with ω_i known and F unknown
Result such that

$$\sum_{ij} \frac{\hat{c}_r^{-1} \omega_r(y_{ij})}{\sum_s n_s \hat{c}_s^{-1} \omega_s(y_{ij})} = 1$$

[Vardi, 1985]

Bridge sampling estimator

$$\sum_{ij} \frac{\hat{c}_r^{-1} \omega_r(y_{ij})}{\sum_s n_s \hat{c}_s^{-1} \omega_s(y_{ij})} = 1$$

[Gelman & Meng, 1998; Tan, 2004]

end of the 2002 discussion

"...essentially every Monte Carlo activity may be interpreted as parameter estimation by maximum likelihood in a statistical model. We do not claim that this point of view is necessary; nor do we seek to establish a working principle from it."

- restriction to discrete support measures [may be] suboptimal
  [Ritov & Bickel, 1990; Robins et al., 1997, 2000, 2003]
- group averaging versions in-between multiple mixture estimators and quasi-Monte Carlo version
  [Owen & Zhou, 2000; Cornuet et al., 2012; Owen, 2003]
- statistical analogy provides at best narrative thread

end of the 2002 discussion

"The hard part of the exercise is to construct a submodel such that the gain in precision is sufficient to justify the additional computational effort"

- garden of forking paths, with infinite possibilities
- no free lunch (variance, budget, time)
- Rao–Blackwellisation may be detrimental in Markov setups

end of the 2002 discussion

"The statistician can considerably improve the efficiency of the estimator by using the known values of different functionals such as moments and probabilities of different sets. The algorithm becomes increasingly efficient as the number of functionals becomes larger. The result, however, is an extremely complicated algorithm, which is not necessarily faster." Y. Ritov

"...the analyst must violate the likelihood principle and eschew semiparametric, nonparametric or fully parametric maximum likelihood estimation in favour of non-likelihood-based locally efficient semiparametric estimators." J. Robins

...and beginning of the 2015 discussion

Questions about probabilistic numerics:

- answer to the zero variance estimator
- significance of a probability statement about a mathematical constant other than epistemic
- posterior in functional spaces mostly reflect choice of prior rather than information...
- ...and idem for loss function
- big world versus small worlds debate

  [Robbins & Wasserman, 2000]
- questionable coherence of Bayesian inference in functional spaces
- unavoidable recourse to (and impact of) Bayesian prior modelling

Robins–Wasserman–Ritov paradox

1. Larry’s constant
2. Charlie’s logistic regression
3. Xiao-Li’s MLE
4. Larry’s and Jamie’s paradox

small worlds...

Observations $x_i \in [0,1]^d$, $r_i \in \{0,1\}$, and censored $y_i \in \{0,1\}$ with joint complete model

$$p(x)\pi(x)^r(1-\pi(x))^{1-r}\theta(x)^y\{1-\theta(x)\}^{1-y}$$

with $p(\cdot)$, $\pi(\cdot)$ known

Quantity of interest

$$\psi = \mathbb{P}(Y=1) = \int_{[0,1]^d} \theta(x)p(x)dx$$

Horwitz-Thompson estimator

$$\hat{\psi}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i r_i}{\pi(x_i)}$$

unbiased and consistent

[Robins & Wasserman, 2012]

Improved versions

[Rotnitzky et al, 2012]

small worlds...

Observations $x_i \in [0,1]^d$, $r_i \in \{0,1\}$, and censored $y_i \in \{0,1\}$ with joint complete model

$$p(x)\pi(x)^r(1-\pi(x))^{1-r}\theta(x)^y\{1-\theta(x)\}^{1-y}$$

with $p(\cdot)$, $\pi(\cdot)$ known
Quantity of interest

$$\psi = \mathbb{P}(Y=1) = \int_{[0,1]^d} \theta(x)p(x)dx$$

Horwitz-Thompson estimator

$$\hat{\psi}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i r_i}{\pi(x_i)}$$

unbiased and consistent

[Robins & Wasserman, 2012]

Improved versions

[Rotnitzky et al, 2012]

...big world

"Any estimator [of $\theta(\cdot)$] that is not a function of $\pi(\cdot)$ cannot be uniformly consistent."

[Robins & Ritov, 1997]

Introducing a prior on $\theta(\cdot)$ does not provide satisfactory answer:
"...the likelihood has no information (...) If the prior on $\theta(\cdot)$ is independent from $\pi(\cdot)$, then the posterior will not concentrate."

[Robins & Wasserman, 2012]

- biased sampling of the Y_i's with known weight $\pi(x_i)$
- above Monte Carlo based solutions available by estimating the reference measure à la Vardi
- not "Bayesian enough"?
- open challenge for today's audience?!

...big world

"Any estimator [of $\theta(\cdot)$] that is not a function of $\pi(\cdot)$ cannot be uniformly consistent."

[Robins & Ritov, 1997]

Introducing a prior on $\theta(\cdot)$ does not provide satisfactory answer:
"...the likelihood has no information (...) If the prior on $\theta(\cdot)$ is independent from $\pi(\cdot)$, then the posterior will not concentrate."

[Robins & Wasserman, 2012]

- biased sampling of the Y_i's with known weight $\pi(x_i)$
- above Monte Carlo based solutions available by estimating the reference measure à la Vardi
- not "Bayesian enough"?
- open challenge for today's audience?!

closure

- pardon my French!
- Ich bin ein Berliner (1993): perfectly coherent to be Bayesian outside statistical frameworks
- the more the merrier