Why (Not) Empirical Likelihood?

Marian Grendár

BioMed of Jessenius Faculty of Medicine, Comenius Univesity, Slovakia

Workshop on Recent Developments in Empirical Likelihood Methodology, IMS NUS, June 13-17, 2016

NUS National University of Singapore | IMS Institute for Mathematical Sciences

Plan of talk

- Empirical Likelihood vs. Fisher Likelihood
  - Multinomial Likelihood under convex constraints
  - Implications for EL
  - Continuous case and FL
- Empirical Likelihood vs. Generalized Minimum Contrast
  - Bayesian nonparametric consistency
  - Large Deviations and Bayesian Law of Large Numbers

EL vs. Fisher Likelihood

EL vs. FL

Based on:

1. MG and V. Špitalský, Multinomial and empirical likelihood under convex constraints: directions of recession, Fenchel duality, perturbations. arXiv:1408.5621, 2014[1].

2. MG and G. Judge, Empty set problem of maximum empirical likelihood methods. Electron. J. Statist., 3:1542-1555, 2009.

[1] Research supported by Slovanet a.s.

EL vs. FL

- EL is 'a multinomial likelihood in the sample'.

EL vs. FL

- EL is 'a multinomial likelihood in the sample'.
- As such, EL intentionally ignores information about the support.

EL vs. FL

- EL is 'a multinomial likelihood in the sample'.
- As such, EL intentionally ignores information about the support.
- Let's consider the discrete case, first.
  And compare EL with the Maximum Multinomial Likelihood.

EL vs. FL

- EL is 'a multinomial likelihood in the sample'.
- As such, EL intentionally ignores information about the support.
- Let's consider the discrete case, first.
  And compare EL with the Maximum Multinomial Likelihood.
- The continuous case will be handled by Fisher's concept of likelihood, later on.

Multinomial likelihood under convex constraints

- Maximum multinomial Likelihood (MmL) may put positive weight to unobserved outcome(s)
- MmL may be different than EL
- Multinomial likelihood ratio may lead to different conclusion than ELR

MmL under convex constraints: setup

- *alphabet* $\mathcal{X}$ of m *letters*
- *probability simplex* $\Delta_{\mathcal{X}} \triangleq \{q \in \mathbb{R}^m : q \geq 0, \sum q = 1\}$
- $(n_i)_{i \in \mathcal{X}}$ *realization* of the *closed multinomial distribution*
$$\Pr((n_i)_{i \in \mathcal{X}}; n, q) = n! \prod q_i^{n_i}/n_i!$$
with parameters $n \in \mathbb{N}$ and $q \in \Delta_{\mathcal{X}}$
- *multinomial likelihood kernel* $L(q) = L_\nu(q) \triangleq e^{-n\,\ell(q)}$, where $\ell = \ell_\nu : \Delta_{\mathcal{X}} \to \bar{\mathbb{R}}$, *Kerridge's inaccuracy*, is
$$\ell(q) \triangleq -\langle \nu, \log q \rangle \,,$$
and $\nu \triangleq (n_i/n)_{i \in \mathcal{X}}$ is the *type*
- $\log 0 = -\infty$, $0 \cdot (-\infty) = 0$; $\bar{\mathbb{R}}$ extended real line $[-\infty, \infty]$; $\langle a, b \rangle$ *scalar product*

MmL under convex constraints: primal $\mathcal{P}$

Consider the primal problem $\mathcal{P}$ of *minimization* of ℓ, restricted to a *convex, closed set* $C \subseteq \Delta_{\mathcal{X}}$:

$$\hat{\ell}_{\mathcal{P}} \triangleq \inf_{q \in C} \ell(q), \qquad S_{\mathcal{P}} \triangleq \{\hat{q} \in C : \ \ell(\hat{q}) = \hat{\ell}_{\mathcal{P}}\} \tag{$\mathcal{P}$}$$

The goal is to find the *solution set* $S_{\mathcal{P}}$ as well as the value $\hat{\ell}_{\mathcal{P}}$ of the objective function ℓ at the infimum over C

MmL under convex constraints: active/passive letters

For a type ν (or, more generally, for any $\nu \in \Delta_{\mathcal{X}}$),
the *active* and *passive alphabets* are

- $\mathcal{X}^a \triangleq \{i \in \mathcal{X} : \nu_i > 0\}$
- $\mathcal{X}^p \triangleq \{i \in \mathcal{X} : \nu_i = 0\}$

The elements of $\mathcal{X}^a$, $(\mathcal{X}^p)$ are called *active, (passive) letters*

MmL under convex constraints: notation

- π^a, π^p the *natural projections* onto active, passive letters
- $x = (x^a, x^p)$ for $x \in \mathbb{R}^m$, where $x^a = \pi^a(x)$, $x^p = \pi^p(x)$
- for $M \subseteq \mathbb{R}^m$ and $x \in M$ put

$$M^a \triangleq \pi^a(M), \qquad \textit{active projection}$$

$$M^a(x^p) \triangleq \{x^a \in \mathbb{R}^{m_a} : (x^a, x^p) \in M\} \quad x^p\textit{-slice}$$

- analogously define M^p and $M^p(x^a)$

MmL under convex constraints: H-set and Z-set

If a non-empty convex, closed set $C \subseteq \Delta_{\mathcal{X}}$ and a type ν are such that $C^a(0^p) = \emptyset$, then we say that C is an H-set with respect to ν

The set C is called a Z-set with respect to ν if $C^a(0^p)$ is non-empty but its support is strictly smaller than $\mathcal{X}^a$

MmL under convex constraints: putting positive weight to unobserved outcomes

Let

- $\mathcal{X} = \{-1, 0, 1\}$,
- $u = (-1, 0, 1)$,
- $C = \{q \in \Delta_{\mathcal{X}} : \langle q, u \rangle = 0\}$,
- $\nu = (1, 0, 0)$.

Thus $\mathcal{X}^a = \{-1\}$, $\mathcal{X}^p = \{0, 1\}$

Then, $S_{\mathcal{P}} = \{(1, 0, 1)/2\}$ and $\hat{\ell}_{\mathcal{P}} = \log 2$

MmL under convex constraints: putting positive weight to unobserved outcomes

Let

- $\mathcal{X} = \{-1, 0, 1\}$,
- $u = (-1, 0, 1)$,
- $C = \{q \in \Delta_{\mathcal{X}} : \langle q, u \rangle = 0\}$,
- $\nu = (1, 0, 0)$.

Thus $\mathcal{X}^a = \{-1\}$, $\mathcal{X}^p = \{0, 1\}$

Then, $S_{\mathcal{P}} = \{(1, 0, 1)/2\}$ and $\hat{\ell}_{\mathcal{P}} = \log 2$

Note: As C is the H-set wrt ν, MEL does not exist, here

MmL under convex constraints: putting positive weight to unobserved outcomes

Let

- $\mathcal{X} = \{-1, 0, 1\}$,
- $u = (-1, 0, 1)$,
- $C = \{q \in \Delta_{\mathcal{X}} : \langle q, u \rangle = 0\}$,
- $\nu = (1, 1, 0)/2$.

Thus $\mathcal{X}^a = \{-1, 0\}$, $\mathcal{X}^p = \{1\}$

Then, $S_{\mathcal{P}} = \{(1, 2, 1)/4\}$ and $\hat{\ell}_{\mathcal{P}} = 0.5 \log 8$

MmL under convex constraints: putting positive weight to unobserved outcomes

Let

- $\mathcal{X} = \{-1, 0, 1\}$,
- $u = (-1, 0, 1)$,
- $C = \{q \in \Delta_{\mathcal{X}} : \langle q, u \rangle = 0\}$,
- $\nu = (1, 1, 0)/2$.

Thus $\mathcal{X}^a = \{-1, 0\}$, $\mathcal{X}^p = \{1\}$

Then, $S_{\mathcal{P}} = \{(1, 2, 1)/4\}$ and $\hat{\ell}_{\mathcal{P}} = 0.5 \log 8$

Note: As C is the Z-set wrt ν, MEL does not exist, here

MmL under convex constraints: and suboptimal EL

Let

- $\mathcal{X} = \{-1, 0, 10\}$,
- $u = (-1, 0, 10)$,
- $C = \{q \in \Delta_{\mathcal{X}} : \langle q, u \rangle = 0\}$,
- $\nu = (3, 2, 0)/5$.

Thus $\mathcal{X}^a = \{-1, 1\}$, $\mathcal{X}^p = \{10\}$

Then, $S_{\mathcal{P}} = \{(54, 44, 1)/99\}$ and $\hat{\ell}_{\mathcal{P}} = 0.6881$

MmL under convex constraints: and suboptimal EL

Let

- $\mathcal{X} = \{-1, 0, 10\}$,
- $u = (-1, 0, 10)$,
- $C = \{q \in \Delta_{\mathcal{X}} : \langle q, u \rangle = 0\}$,
- $\nu = (3, 2, 0)/5$.

Thus $\mathcal{X}^a = \{-1, 1\}$, $\mathcal{X}^p = \{10\}$

Then, $S_{\mathcal{P}} = \{(54, 44, 1)/99\}$ and $\hat{\ell}_{\mathcal{P}} = 0.6881$

Note: MEL is $\hat{q}_{\mathcal{E}} = (1, 1, 0)/2$, $\hat{\ell}_{\mathcal{E}} = 0.6931$ and $\hat{\ell}_{\mathcal{P}} < \hat{\ell}_{\mathcal{E}}$

MmL under convex constraints: decision making

Let $\mathcal{X} = \{-2, -1, 0, 1, 2\}$ and $C(\theta_j) = \{q \in \Delta_{\mathcal{X}} : \mathrm{E}_q(X^2) = \theta_j\}$, where $\theta_1 = 1.01$, $\theta_2 = 1.05$
Let $\nu = (6, 3, 0, 0, 1)/10$

The solution of $\mathcal{P}$ is

- $\hat{q}_{\mathcal{P}}(\theta_1) = (0.1515, 0.3030, 0.52025, 0, 0.02525)$, for θ_1
- $\hat{q}_{\mathcal{P}}(\theta_2) = (0.1575, 0.3150, 0.50125, 0, 0.02625)$, for θ_2
- $\mathrm{LR}_{21} = \exp\left(n[\ell(\hat{q}_{\mathcal{P}}(\theta_1)) - \ell(\hat{q}_{\mathcal{P}}(\theta_2))]\right) = 1.48$

which indicates inconclusive evidence

For both θ's the solution of EL primal exists and it is

- $\hat{q}_{\mathcal{E}}(\theta_1) = (0.00286, 0.99\bar{6}, 0.00048)$, for θ_1
- $\hat{q}_{\mathcal{E}}(\theta_2) = (0.01429, 0.98\bar{3}, 0.00238)$, for θ_2
- $\mathrm{ELR}_{21} = \exp\left(n[\ell(\hat{q}_{\mathcal{E}}(\theta_1)) - \ell(\hat{q}_{\mathcal{E}}(\theta_2))]\right) = 75031.31$

which indicates decisive evidence for θ_2

Implications for EL

- MmL under convex constraints always exists
- MmL may put positive weight to passive letters
- MEL does not exist if C is the H-set or Z-set, wrt ν
- Note that also the EL outer optimization problem may have no solution; cf. ESP, ❷
- If MEL exists, the value of EL at MEL may be smaller than the value of the multinomial likelihood at MmL
- If ELR exists, it may lead to different inferential conclusion than the Multinomial Likelihood Ratio

Continuous case and Fisher likelihood

Due to the finite precision of any measurement 'all actual sample spaces are discrete, and all observable random variables have discrete distributions', Pitman

Already Fisher's original notion of the likelihood reflects the finiteness of the sample space

For an *iid* sample $X_1^n \triangleq X_1, X_2, \dots, X_n$ and a finite partition $\mathcal{A} = \{A_l\}_1^m$ of a sample space $\mathcal{X}$ the *Fisher likelihood* $L_{\mathcal{A}}(q; X_1^n)$ which the data X_1^n provide to a pmf $q \in \Delta_{\mathcal{X}}$ is

$$L_{\mathcal{A}}(q; X_1^n) \triangleq \prod_{A_l \in \mathcal{A}} e^{n(A_l) \log q(A_l)},$$

where $n(A_l)$ is the number of observations in X_1^n that belong to A_l

This view thus carries the discordances between the multinomial and empirical likelihoods also to the continuous *iid* setting

Fisher likelihood with estimating equations: example

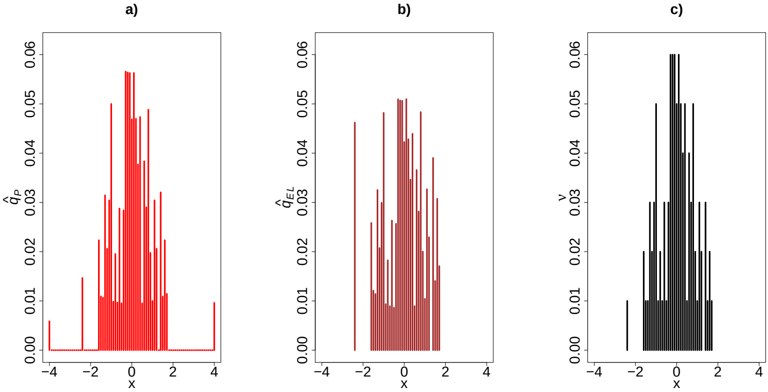


Figure : a) MmL $\hat{q}_{\mathcal{P}}$; b) MEL $\hat{q}_{\mathcal{E}}$; c) the observed type ν. Qin & Lawless, Ex. 1, with finite partition

EL vs. Generalized Minimum Contrast

EL vs. GMC

Based on:

1. MG and G. Judge, Asymptotic Equivalence of Empirical Likelihood and Bayesian MAP. Ann. Statist., 37(5A):2445-2457, 2009.

2. MG and G. Judge, Large deviations theory and econometric information recovery. In Handbook of empirical economics and finance, A.Ullah and D. E. A. Giles (eds.), pp. 155-182, Chapman & Hall/CRC, 2011.

3. MG and G. Judge, Not all empirical divergence minimizing statistical methods are created equal? In ICNPAA 2012, S. Sivasundaram (ed.), AIP (Melville), pp. 432-435, 2012.

EL vs. GMC

- EL is just one member of the GMC class of estimators and tests

EL vs. GMC

- EL is just one member of the GMC class of estimators and tests
- Are all GMC estimators created equal?

EL vs. GMC

- EL is just one member of the GMC class of estimators and tests
- Are all GMC estimators created equal?
- Bayesian Law of Large Numbers implies that EL is the only member of GMC that is consistent under misspecification

Estimating Equations

Setup:

Chance: r.v. $X \in \mathcal{X} \subseteq \mathrm{R}^d$, with cdf $Q_r \in \mathcal{Q}(\mathcal{X})$, where $\mathcal{Q}(\mathcal{X})$ is the set of all cdf's on $\mathcal{X}$.

Data: $X_1^n = X_1, \ldots, X_n$, iid from Q_r.

Model:
Estimating functions: $u(X;\theta) : \mathcal{X} \times \Theta \to \mathrm{R}^J$, where $\theta \in \Theta \subseteq \mathrm{R}^K$; K can be, in general, different than J.
Estimating equations (EE):

$$\Phi(\theta) = \{Q \in \mathcal{Q}(\mathcal{X}) : \mathrm{E}_Q u(X;\theta) = 0\}.$$

Model: $\Phi(\Theta) \; = \bigcup_{\theta \in \Theta} \Phi(\theta)$.

Estimating Equations: examples

Examples:

Ex. 1: $\mathcal{X} = \mathrm{R}$, $\Theta = [0, \infty)$, $u(X; \theta) = X - \theta$.

Ex. 2: (Brown & Chen) $\mathcal{X} = \mathrm{R}$, $\Theta = \mathrm{R}$,
$u(X; \theta) = \{X - \theta, \mathrm{sgn}(X - \theta)\}$.

Ex. 3: (Qin & Lawless) $\mathcal{X} = \mathrm{R}$, $\Theta = \mathrm{R}$,
$u(X; \theta) = \{X - \theta, X^2 - (2\theta^2 + 1)\}$.

Objective: selection

Given a random sample X_1^n from Q_r, the objective is to select a $\hat{Q}$ from $\Phi(\Theta)$, and in this way provide a point estimate $\hat{\theta}$ of the 'true' value θ_r.

If the model is correctly specified (i.e., $Q_r \in \Phi(\Theta)$), θ_r solves $\mathrm{E}_{Q_r}\, u(X;\theta) = 0$.

If the model is misspecified (i.e., $Q_r \notin \Phi(\Theta)$), then $\theta_r =$???.

Empirical Estimating Equations

To connect the model $\Phi(\Theta)$ with the data X_1^n, replace the model $\Phi(\Theta)$ by its empirical, data-based analogue $\Phi_n(\Theta) = \bigcup_{\theta\in\Theta} \Phi_n(\theta)$, where

$$\Phi_n(\theta) = \{Q_n \in \mathcal{Q}(X_1^n) : \mathrm{E}_{Q_n} u(X;\theta) = 0\}$$

are the empirical estimating equations.

Empirical Estimating Equations (E^3) approach to estimation and inference replaces the set $\Phi(\Theta)$ of cdf's supported on $\mathcal{X}$ by the set $\Phi_n(\Theta)$ of cdf's that are supported on the data X_1^n.

An estimate $\hat{\theta}$ of θ_r is obtained by means of a rule that selects $\hat{Q}_n(x;\hat{\theta})$ from the empirical set $\Phi_n(\Theta)$.

Generalized Minimum Contrast rule

GMC selects $\hat{Q}_n(x;\hat{\theta})$ from $\Phi_n(\Theta)$:

$$\hat{Q}_n(x;\hat{\theta}) = \arg \inf_{Q_n(x;\theta)\in\Phi_n(\Theta)} D_\phi(Q_n \,||\, \hat{Q}_r) \tag{1}$$

where

- $\phi(\cdot)$ is a convex function with minimum at 1,
- $\hat{Q}_r(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n}$ is the empirical cdf

GMC rule selects the member of $\Phi_n(\Theta)$ which is closest to the empirical cdf $\hat{Q}_r$, in the sense of the divergence $D_\phi(\cdot \,||\, \cdot)$.

Typical GMC rules

Typical choices of $\phi(\cdot)$ are:

- $-\log x$; leads to *Maximum Empirical Likelihood*, assoc. with the L-divergence,
- $x \log x$; leads to *Exponential Empirical Likelihood*, assoc. with the I-divergence,
- $2/(\alpha(\alpha+1))(x^{-\alpha}-1)$; leads to the Cressie-Read family-based *Generalized Empirical Likelihood*, assoc. with the CR-divergences

GMC estimator

The θ part of the GMC optimization problem (1):

$$\hat{\theta} = \arg \inf_{\theta \in \Theta} \inf_{Q_n(x) \in \Phi_n(\theta)} \mathrm{E}_{\hat{Q}_r} \phi \left(\frac{dQ}{d\hat{Q}_r} \right), \tag{2}$$

The convex dual form of (2):

$$\hat{\theta} = \arg \inf_{\theta \in \Theta} \sup_{\mu \in \mathrm{R}, \lambda \in \mathrm{R}^J} \left[\mu - \mathrm{E}_{\hat{Q}_r} \phi^*(\mu + \lambda' u(x; \theta)) \right],$$

where $\phi^*(y) = \sup_x xy - \phi(x)$ is the Legendre Fenchel transformation of $\phi(x)$.

MEL as GMC

Recall that $\phi(x) = -\log x$ leads to Maximum Empirical Likelihood (MEL).

$$\hat{\theta}_{\mathrm{MEL}} = \arg \inf_{\theta \in \Theta} \sup_{\lambda \in \mathrm{R}^J} \mathrm{E}_{\hat{Q}_r} \log(1 + \lambda' u(x; \theta)).$$

MEL selects among the data-supported cdf's from the empirical model $\Phi_n(\Theta)$ the one with the highest value of the likelihood.

Question

Are all the GMC methods created equal?

Answer: Bayesian infinite dimensional consistency under misspecification.

Bayesian infinite dimensional consistency

A prior Π is put on $\Phi(\Theta)$; (it induces a prior $\Pi(\theta)$ over Θ). The prior Π combines with the data X_1^n to define the posterior:

$$\Pi_n(A \mid X_1^n) = \frac{\int_A e^{-l_n(Q)} \Pi(dQ)}{\int_\Phi e^{-l_n(Q)} \Pi(dQ)},$$

where $l_n(Q) = -\mathrm{E}_{\hat{Q}_r} \log \frac{dQ}{d\hat{Q}_r}$, and $A \subseteq \Phi$.

Bayesian infinite-dimensional consistency: the objective – to determine the distribution(s) on which the posterior Π_n concentrates as n gets large.

Bayesian consistency under misspecification

If the model is misspecified, i.e., $Q_r \notin \Phi(\Theta)$,
then the true value θ_r can be defined as the value $\hat{\theta}_{\mathrm{L}}$ corresponding to the distribution $\hat{Q}_{\mathrm{L}}$ on which the posterior concentrates.

Bayesian consistency under misspecification

If the model is misspecified, i.e., $Q_r \notin \Phi(\Theta)$,
then the true value θ_r can be defined as the value $\hat{\theta}_{\mathrm{L}}$ corresponding to the distribution $\hat{Q}_{\mathrm{L}}$ on which the posterior concentrates.

An estimator $\hat{\theta}$ of θ is consistent under misspecification if $\hat{\theta} \xrightarrow{p} \hat{\theta}_{\mathrm{L}}$.

Bayesian Law of Large Numbers

BLLN. (G&J, 09) Under some regularity conditions the posterior concentrates on the union of weak ϵ-balls that are centered at the L-projections $\hat{Q}_{\mathrm{L}}$ of Q_r on Φ.

The L-divergence and L-projection

The L-projection $\hat{Q}_{\mathrm{L}}$ of Q_r on Φ

$$\hat{Q}_{\mathrm{L}} = \arg \inf_{Q \in \Phi} \mathrm{L}(Q \,||\, Q_r),$$

where $\mathrm{L}(Q \,||\, Q_r)$ is the L-divergence (aka the reverse I-divergence) of Q wrt Q_r

$$\mathrm{L}(Q \,||\, Q_r) = -\,\mathrm{E}_{Q_r} \log \frac{dQ}{dQ_r}.$$

The BLLN is an extension of Schwartz' consistency theorem to the case of misspecified model.

Answer

Recall that GMC selects

$$\hat{\theta} = \arg \inf_{\theta \in \Theta} \inf_{Q_n(x) \in \Phi_n(\theta)} \mathrm{E}_{\hat{Q}_r} \phi \left(\frac{dQ}{d\hat{Q}_r} \right).$$

BLLN implies that

- MEL (i.e., $\phi(x) = -\log x$) is consistent under misspecification,
- other GMC methods are not.

MC study: inconsistency-under-misspecification of EEL

Recall that the Exponential Empirical Likelihood (EEL) is associated with $\phi(x) = x \log x$, i.e., the empirical I-projection of $\hat{Q}_r$ on $\Phi_n(\Theta)$.

The posterior odds $\mathrm{PO}_{\mathrm{IL}}$ of the empirical I-projection $\hat{Q}_{\mathrm{I},n}(x; \hat{\theta}_{\mathrm{EEL}})$ to the empirical L-projection $\hat{Q}_{\mathrm{L},n}(x; \hat{\theta}_{\mathrm{EL}})$ is proportional to

$$\Delta_n = \frac{1}{n} \sum_{i=1}^{n} \log \frac{d\, \hat{Q}_{\mathrm{I},n}(x_i; \hat{\theta}_{\mathrm{EEL}})}{d\, \hat{Q}_{\mathrm{L},n}(x_i; \hat{\theta}_{\mathrm{EL}})},$$

which converges almost sure to

$$\Delta = \mathrm{L}(\hat{Q}_{\mathrm{L}} \,||\, Q_r) - \mathrm{L}(\hat{Q}_{\mathrm{I}} \,||\, Q_r);$$

there $\hat{Q}_{\mathrm{I}}$ is the I-projection of Q_r on Φ.

MC study (cont'd)

Setting: Ex. 3, the gaussian $n(1.1, \sigma = 2.75)$ source. Model is misspecified.

Table : MC estimates of the Mean Squared Error (MSE) of EL, EEL and Δ_n estimators.

n	$\mathrm{MSE}(\hat{\theta}_{\mathrm{EL}})$	$\mathrm{MSE}(\hat{\theta}_{\mathrm{EEL}})$	$\mathrm{MSE}(\Delta_n)$
100	0.0375	0.0359	0.000225
500	0.0082	0.0088	0.000042
1000	0.0041	0.0046	0.000024
5000	0.0019	0.0012	0.000008

MC study (cont'd)

The large deviations convergence of the posterior odds $\mathrm{PO}_{\mathrm{IL}}$ can be informally stated as

$$\mathrm{PO}_{\mathrm{IL}} \approx e^{n\,\Delta}.$$

Since $\Delta = -0.0147$, this implies the inconsistency under misspecification of the parametric component $\hat{\theta}_{\mathrm{EEL}}$ of $\hat{Q}_{\mathrm{I},n}$, which is based on selection of the I-projection.

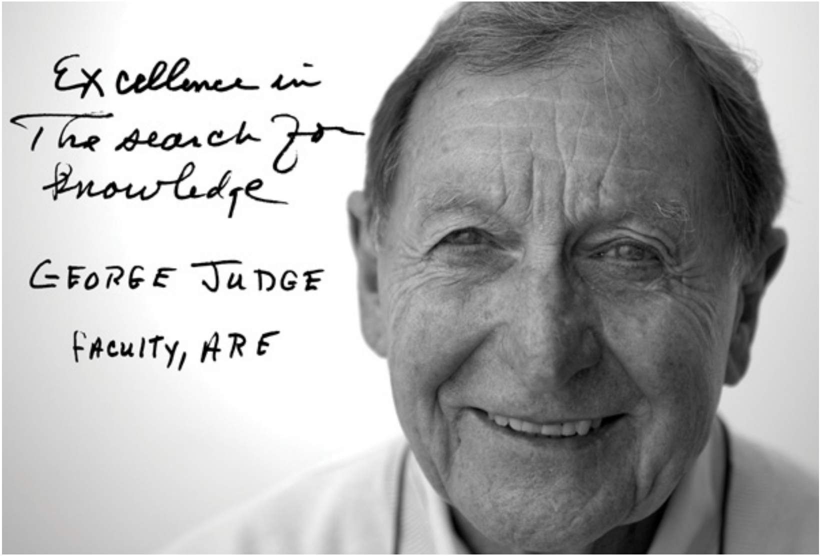
Reverse I-divergence and I-divergence; iid case

In the problem of selecting a sampling distribution, BLLN disqualifies the GMC methods other than the L-divergence (reverse I-divergence) based MEL.

Recall that in the problem of selecting an empirical distribution, the Conditional Law of Large Numbers disqualifies the maximum entropy methods other than the I-divergence based MaxEnt.

Dedication

To George Judge



Thank you for your attention!