Combining Parametric and Empirical Likelihoods

Ian McKeague

June 22, 2016

(ongoing work, with Nils Hjort and Ingrid Van Keilegom)

Columbia University
in the City of New York

Theme: Combining parametrics with nonparametrics

Observe $y_1, \ldots, y_n \sim$ i.i.d. f, and suppose inference is needed for a focus parameter $\psi = \psi(f)$.

Parametric likelihood approach (perfect if model is perfect):

Fit f to $\{f_\theta \colon \theta \in \Theta\}$ via maximum likelihood, $\widehat\theta_{\rm ML}$ maximising log-likelihood $\ell_n(\theta) = \log L_n(\theta)$. Then

$$\sqrt{n}(\widehat\theta_{\rm ML} - \theta) \to_d {\rm N}_p(0, J^{-1}).$$

Delta method gives

$$\sqrt{n}(\widehat\psi_{\rm ML} - \psi) \to_d {\rm N}(0, \kappa^2),$$

with $\kappa^2 = c^{\rm T} J^{-1} c$ and $c = \partial\psi(\theta)/\partial\theta$. Wilks theorem.

Nonparametric likelihood approach (no conditions needed):

Identify ψ via $\mathbb{E}_f\, m(Y, \psi) = 0$. EL function $R_n(\psi)$ is the max of $\prod_{i=1}^n n w_i$ under $\sum_{i=1}^n w_i = 1$, $\sum_{i=1}^n w_i m(y_i, \psi) = 0$, $w_i > 0$.

$$-2 \log R_n(\psi) \to_d \chi^2_1.$$

How to combine parametric and empirical likelihood?

Main idea (with details and variations and applications to come):

- Decide on control parameters $\mu = (\mu_1, \ldots, \mu_q)$, identified via $\mathbb{E}\, m_j(Y, \mu) = 0$ for $j = 1, \ldots, q$;
- put the parametric model through the EL, giving $R_n(\mu(\theta))$;

and form

$$H_n(\theta) = L_n(\theta)^{1-a} R_n(\mu(\theta))^a.$$

We will show that the hybrid likelihood estimator $\widehat{\theta}_{\rm HL}$ maximising

$$h_n(\theta) = (1-a)\ell_n(\theta) + a \log R_n(\mu(\theta)),$$

along with focus parameter estimator $\widehat{\psi}_{\rm HL} = \psi(f(\cdot, \widehat{\theta}_{\rm HL}))$, have good properties.

FIC type schemes to assist in selecting balance parameter a in $[0, 1]$ and the control parameters $\mu_1, \ldots, \mu_q$.

Plan

General setup (so far for i.i.d., extensions later): With working model $f(y, \theta)$, leading to log-likelihood $\ell_n(\theta)$, and control parameters μ:

$$h_n(\theta) = (1 - a)\ell_n(\theta) + a \log R_n(\mu(\theta)).$$

A Examples

B Basics for the EL

C Theory: under the model

D Theory: outside the model

E Fine-tuning the balance parameter a

F Choosing the control parameters $\mu_1, \ldots, \mu_q$

G Concluding remarks (and questions)

A: Examples

Example 1. Let f_θ be the normal (ξ, σ^2), and use

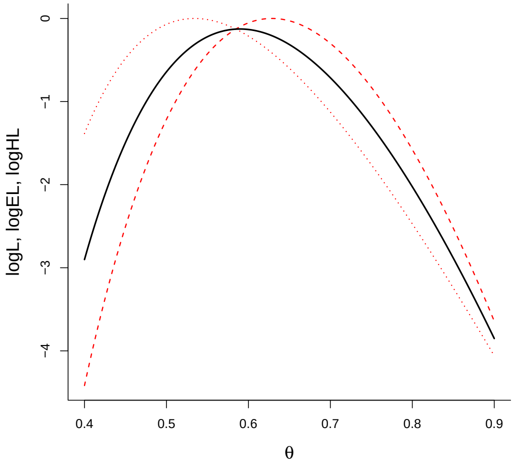
$$m_j(y, \mu_j) = I\{y \leq \mu_j\} - j/4 \quad \text{for } j = 1, 2, 3.$$

Then HL means estimating (ξ, σ) factoring in that the three quartiles ought to be estimated well too.

Example 2. Let f_θ be the Beta with parameters (b, c). ML means moment matching for $\log y_i$ and $\log(1 - y_i)$. Add to these functions $m_1(y, \mu_1) = y - \mu_1$ and $m_2(y, \mu_2) = y^2 - \mu_2$. Then HL is Beta fitting with getting mean and variance not far from

$$\mathbb{E}_{\mathrm{Beta}}\, Y = \frac{b}{b+c} \quad \text{and} \quad \mathrm{Var}_{\mathrm{Beta}}\, Y = \frac{1}{b+c+1}\frac{b}{b+c}\frac{c}{b+c}.$$

Example 3. $f(y, \theta) = \theta y^{\theta - 1}$, $y \in (0, 1)$, $\theta > 0$. The log-likelihood is $n\{\log \theta - (\theta - 1)Z_n\}$, with $Z_n = (1/n)\sum_{i=1}^{n} \log(1/y_i)$, and $\widehat{\theta}_{\mathrm{ML}} = 1/Z_n$. Then put the EL for the mean μ through the model, yielding $R_n(\mu(\theta))$ with $\mu(\theta) = \theta/(\theta + 1)$. This is HL with $a = \frac{1}{2}$:



Example 4. Newcomb's 1889 speed of light data

$n = 66$ and two grand outliers at -44 and -2. True value is 33.02.

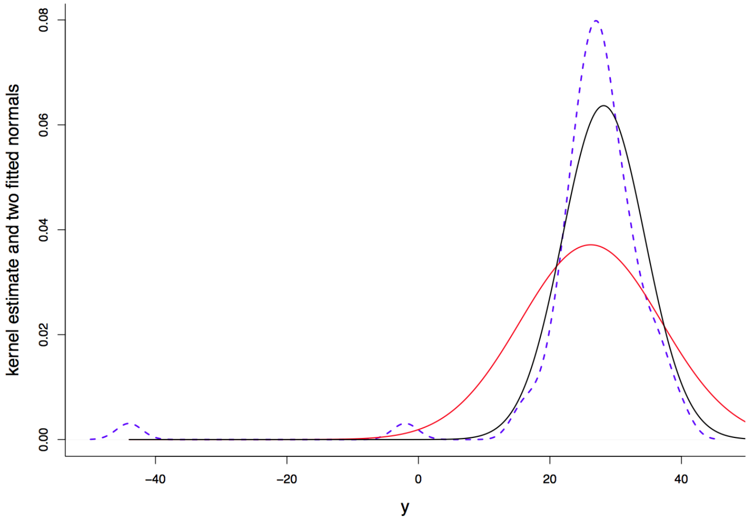
Normal model: estimates of mean and variance are $(26.21, 10.75)$.
and after removing outliers $(27.75, 5.08)$.

Now use HL with histogram associated control parameters, with $k = 6$ cells

$(-\infty, 10.5], (10.5, 20.5], (20.5, 25.5], (25.5, 30.5], (30.5, 35.5], (35.5, \infty)$.

The HL, with $a = 0.50$: $(28.23, 6.37)$.

$a = 1$: Close to minimum chi-squared.


kernel estimate and two fitted normals
0.00
0.02
0.04
0.06
0.08
−40
−20
0
20
40
y

Two (related) viewpoints

Which $\mu_1, \ldots, \mu_q$ should we use in $(1-a)\ell_n(\theta) + a \log R_n(\mu(\theta))$? Robustify a parametric model, and/or helping to focus the nonparametric method?

Viewpoint One (focused robustness): Using control parameters to help the parametric fit do well for these too. – For the normal (ξ, σ^2), we might want not only mean and standard deviation to be ok, but also $\widehat{\xi} - 0.675\,\widehat{\sigma}, \widehat{\xi} + 0.675\,\widehat{\sigma}$ to reasonably match quartiles $F_n^{-1}(\frac{1}{4}), F_n^{-1}(\frac{3}{4})$.

Viewpoint Two (with focus parameter): We wish the fitted model to give a particularly good estimate of $\psi = \psi(f)$ via $\widehat{\psi}_{\rm HL} = \psi(f(\cdot, \widehat{\theta}_{\rm HL}))$. Then we use the HL with $p+1$ parameters, the working model plus the focus ψ. – For the normal, we may put in $m(y, \mu) = I\{y \le \mu\} - 3/4$, and use $\widehat{\xi}_{\rm HL} + 0.675\,\widehat{\sigma}_{\rm HL}$ to estimate $F^{-1}(\frac{3}{4})$.

B: Empirical likelihood

For q-vectors $m_1, \ldots, m_n$, consider

$$R_n = \max\Big\{\prod_{i=1}^n n w_i \colon \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i m_i = 0, \text{each } w_i > 0\Big\}.$$

Let

$$G_n(\lambda) = \sum_{i=1}^n 2\log(1 + \lambda^{\mathsf{T}} m_i/\sqrt{n}) \text{ and } G_n^*(\lambda) = 2\lambda^{\mathsf{T}} V_n - \lambda^{\mathsf{T}} W_n \lambda,$$

where $V_n = n^{-1/2}\sum_{i=1}^n m_i$ and $W_n = n^{-1}\sum_{i=1}^n m_i m_i^{\mathsf{T}}$.

Dual optimization: $-2\log R_n = \max_\lambda G_n(\lambda) = G_n(\widehat{\lambda})$.

With the m_i random; eigenvalues of W_n away from zero and infinity; $n^{-1/2}\max_{i\le n}\|m_i\| \to_{\rm pr} 0$; V_n bounded in probability: then $G_n \approx G_n^*$ where it matters, and

$$-2\log R_n = V_n^{\mathsf{T}} W_n^{-1} V_n + o_{\rm pr}(1).$$

This machinery is then used with $m_i = m(Y_i, \mu(\theta))$.

C: Theory: under the model

First aim: working out how the HL behaves under model conditions (it will lose some to ML there, but how much?). With

$$h_n(\theta) = (1-a)\ell_n(\theta) + a \log R_n(\mu(\theta)),$$

and θ_0 the true value, define

$$A_n(s) \quad = \quad h_n(\theta_0 + s/\sqrt{n}) - h_n(\theta_0).$$

Understanding behavior of $A_n \Longrightarrow$ understanding behaviour of $\widehat{\theta}_{\rm HL}$ (et al.). With $u(\cdot,\theta) = \dot{\ell}_\theta$ as the score function,

$$\begin{pmatrix} U_{n,0} \\ V_{n,0} \end{pmatrix} \quad = \quad \begin{pmatrix} n^{-1/2}\sum_{i=1}^n u(Y_i,\theta_0) \\ n^{-1/2}\sum_{i=1}^n m(Y_i,\mu(\theta_0)) \end{pmatrix}$$
$$\to_d \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \sim {\rm N}_{p+q}(0, \begin{pmatrix} J & C \\ C^{\sf T} & W \end{pmatrix})$$

where $J = J_{\rm fish}$ is the Fisher information matrix.

Local asymptotic normality (LAN)

Theorem: There is a limiting quadratic process:

$$A_n(s) = h_n(\theta_0 + s/\sqrt{n}) - h_n(\theta_0) \to_d A(s) = s^{\mathsf{T}} U^* - \tfrac{1}{2} s^{\mathsf{T}} J^* s$$

over compacta, where

$$\begin{aligned} U^* &= (1-a)U_0 - a\xi_0^{\mathsf{T}} W^{-1} V_0, \\ J^* &= (1-a)J + a\xi_0^{\mathsf{T}} W^{-1} \xi_0. \end{aligned}$$

Here $\xi_0 = \mathrm{E}\,\partial m(Y, \mu(\theta_0))/\partial\theta$. Also, $U^* \sim \mathrm{N}_p(0, K^*)$ with

$$K^* = (1-a)^2 J + a^2 \xi_0^{\mathsf{T}} W^{-1} \xi_0 - a(1-a)(C W^{-1} \xi_0 + \xi_0^{\mathsf{T}} W^{-1} C^{\mathsf{T}}).$$

The most important aspects of how $\widehat{\theta}_{\mathrm{HL}}$ behaves can now be read off from $A_n(s) \to_d A(s)$.

Fact 1 [using $\operatorname{argmax}(A_n) \to_d \operatorname{argmax}(A)$]:

$$\sqrt{n}(\widehat{\theta}_{\rm HL} - \theta_0) \to_d (J^*)^{-1} U^* \sim {\rm N}_p(0, (J^*)^{-1} K^* (J^*)^{-1}).$$

Fact 2 [using $\max A_n \to_d \max A$]:

$$Z_n(\theta_0) = 2\{h_n(\widehat{\theta}_{\rm HL}) - h_n(\theta_0)\} \to_d Z = (U^*)^{\sf T} (J^*)^{-1} U^*.$$

Fact 3 [applying the delta method]: With $\widehat{\psi}_{\rm HL} = \psi(\widehat{\theta}_{\rm HL})$ and $\psi_0 = \psi(\theta_0)$ at true value,

$$\sqrt{n}(\widehat{\psi}_{\rm HL} - \psi_0) \to_d {\rm N}(0, \kappa^2),$$

with $\kappa^2 = c^{\sf T} (J^*)^{-1} K^* (J^*)^{-1} c$ and $c = \partial\psi(\theta_0)/\partial\theta$.

Result: HL loses rather little compared to the ML under model conditions:

$$(J^*)^{-1} K^* (J^*)^{-1} = J_{\rm fish}^{-1} + O(a^2).$$

LAN for the parametric likelihood

$$\ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = s^{\mathsf{T}} U_{n,0} - \tfrac{1}{2} s^{\mathsf{T}} J s + o_{\mathrm{pr}}(1)$$

See, for example, van der Vaart's *Asymptotic Statistics*:

7.2 Theorem. *Suppose that Θ is an open subset of $\mathbb{R}^k$ and that the model $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at θ. Then $P_\theta \dot{\ell}_\theta = 0$ and the Fisher information matrix $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ exists. Furthermore, for every converging sequence $h_n \to h$, as $n \to \infty$,*

$$\log \prod_{i=1}^{n} \frac{p_{\theta + h_n/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T \dot{\ell}_\theta(X_i) - \frac{1}{2} h^T I_\theta h + o_{P_\theta}(1).$$

LAN for the hybrid likelihood will then hold since

$$\begin{aligned} A_n(s) &= h_n(\theta_0 + s/\sqrt{n}) - h_n(\theta_0) \\ &= (1-a)\{\ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0)\} \\ &\qquad + a\{\log R_n(\mu(\theta_0 + s/\sqrt{n})) - \log R_n(\mu(\theta_0))\}, \end{aligned}$$

provided we also have LAN jointly for the empirical likelihood.

LAN for the empirical likelihood

By the quadratic approximation to $-2\log R_n$,

$$\log R_n(\mu(\theta_n)) = -\tfrac{1}{2} V_n^{\mathsf{T}} W_n^{-1} V_n + o_{\rm pr}(1)$$

where $\theta_n = \theta_0 + s/\sqrt{n}$,

$$V_n = n^{-1/2}\sum_{i=1}^{n} m(Y_i, \mu(\theta_n)) = V_{n,0} + \xi_n s + o_{\rm pr}(1)$$

[if, say, $m(y, \mu(\theta))$ has a first-order Taylor expansion in θ],

$$W_n = n^{-1}\sum_{i=1}^{n} m(Y_i, \mu(\theta_n)) m(Y_i, \mu(\theta_n))^{\mathsf{T}} = W_{n,0} + o_{\rm pr}(1).$$

$V_{n,0} \to_d V_0$, and $\xi_n = \mathbb{P}_n\xi \to \mathbb{E}\xi(Y) = \xi_0$, $W_{n,0} \to W$ (by LLN).

HL can be as good as ML

Example 5. Let $f_\theta = N(\theta, 1)$ and use the median as the control parameter, so $\mu(\theta) = \theta$ and we take

$$m(y, \mu) = I\{y \leq \mu\} - 1/2.$$

Note: $m(y, \mu(\theta))$ has no Taylor expansion in θ. Donsker gives

$$V_n - V_{n,0} = n^{-1/2} \sum_{i=1}^{n} 1\{\theta_0 < Y_i \leq \theta_0 + s/\sqrt{n}\} \to_{\mathrm{pr}} 0$$

so we still have LAN for the HL, and find that $\xi_0 = 0$.

This implies that $\widehat{\theta}_{\mathrm{HL}}$ and $\widehat{\theta}_{\mathrm{ML}}$ have the same asymp variance:

$$(J^*)^{-1} K^* (J^*)^{-1} = J_{\mathrm{fish}}^{-1} \quad \text{for all choices of a.}$$

D: Theory: outside the model

Results so far: behaviour of $\widehat{\theta}_{\rm HL}$ and consequent $\widehat{\psi}_{\rm HL}$ well understood under parametric model conditions, where they may lose a little, but not much compared to ML.

Will now show (though a bigger machinery and more efforts are required) that HL is (often) better than ML just outside the parametric model.

Framework: extend $f(y,\theta)$ model (with $\dim(\theta)=p$) to a bigger $f(y,\theta,\gamma)$ model (with $\dim(\gamma)=r$), and such that $\gamma=\gamma_0$ corresponds to the start model; $f(y,\theta,\gamma_0)=f(y,\theta)$.

Local neighborhood model framework:

$$f_{\rm true}(y)=f(y,\theta_0,\gamma_0+\delta/\sqrt{n}).$$

Thus $\psi_{\rm true}=\psi(\theta_0,\gamma_0+\delta/\sqrt{n})$, etc.

Under $f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$, suppose an estimation strategy $\widehat{\theta}$ has the property

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}_p(B\delta, \Omega),$$

for appropriate B ($p \times r$ matrix, related to how the model bias affects the estimator) and Ω.

For $\psi = \psi(f) = \psi(\theta, \gamma)$, may use $\widehat{\psi} = \psi(\widehat{\theta}, \gamma_0)$. Then analysis leads to

$$\sqrt{n}(\widehat{\psi} - \psi_{\mathrm{true}}) \to_d \mathrm{N}(b^{\mathsf{T}}\delta, \tau^2),$$

with

$$b = B^{\mathsf{T}}\frac{\partial\psi}{\partial\theta} - \frac{\partial\psi}{\partial\gamma} \quad \text{and} \quad \tau^2 = (\tfrac{\partial\psi}{\partial\theta})^{\mathsf{T}}\Omega\frac{\partial\psi}{\partial\theta}$$

with derivatives at narrow model (θ_0, γ_0). Hence limit mean squared error is

$$\mathrm{mse}_{\widehat{\psi}}(\delta) = (b^{\mathsf{T}}\delta)^2 + \tau^2.$$

Next: Examining estimation strategies ML and HL, to find B and Ω, and hence the $\mathrm{mse}_{\widehat{\psi}}(\delta)$. For ML: as in Hjort and Claeskens (2003); for HL: new.

The story for the ML: Essentially from Hjort and Claeskens (2003, 2008). Need the $(p+r)\times(p+r)$ Fisher information matrix

$$J_{\rm wide} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$$

at the narrow model. From this (via various efforts):

$$\sqrt{n}(\widehat\theta_{\rm ML} - \theta_0) \to_d {\rm N}_p(J_{00}^{-1}J_{01}\delta, J_{00}^{-1}).$$

This implies

$$\sqrt{n}(\widehat\psi_{\rm ML} - \psi_{\rm true}) \to_d {\rm N}(\omega^{\rm T}\delta, \tau_0^2)$$

with

$$\omega = J_{10}J_{00}^{-1}\frac{\partial\psi}{\partial\theta} - \frac{\partial\psi}{\partial\gamma} \quad \text{and} \quad \tau_0^2 = (\tfrac{\partial\psi}{\partial\theta})^{\rm T} J_{00}^{-1}\frac{\partial\psi}{\partial\theta}.$$

Hence we know

$$\mathrm{mse}_{ML}(\delta) = (\omega^{\rm T}\delta)^2 + \tau_0^2$$

and should compare this with what we may find for the HL.

The story for the HL: For $S(y) = \partial \log f(y, \theta_0, \gamma_0)/\partial\gamma$, let

$$K_{01} = \mathbb{E}\, m(Y, \mu(\theta_0)) S(Y)$$

of dimension $q \times r$, along with

$$L_{01} = (1-a) J_{01} - a(\tfrac{\partial\psi}{\partial\theta})^{\mathsf{T}} W^{-1} K_{01}.$$

Then (via various efforts):

$$\sqrt{n}(\widehat\theta_{\rm HL} - \theta_0) \to_d {\rm N}_p(B\delta, \Omega)$$

with $B = (J^*)^{-1} L_{01}$ and $\Omega = (J^*)^{-1} K^* (J^*)^{-1}$. This yields

$$\sqrt{n}(\widehat\psi_{\rm HL} - \psi_{\rm true}) \to_d {\rm N}(\omega_{\rm HL}^{\mathsf{T}}\delta, \tau_{0,{\rm HL}}^2)$$

with

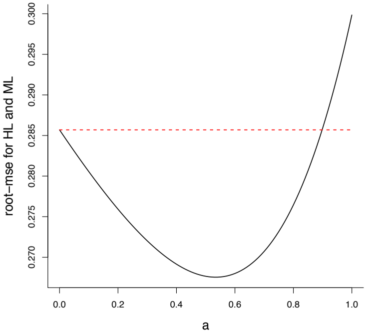
$$\begin{aligned} \omega_{\rm HL} &= \omega_{{\rm HL},a} = L_{10}(J^*)^{-1}\tfrac{\partial\psi}{\partial\theta} - \tfrac{\partial\psi}{\partial\gamma}, \\ \tau_{0,{\rm HL}}^2 &= \tau_{0,{\rm HL},a}^2 = (\tfrac{\partial\psi}{\partial\theta})^{\mathsf{T}}(J^*)^{-1}K^*(J^*)^{-1}\tfrac{\partial\psi}{\partial\theta}. \end{aligned}$$

Here J^*, K^*, L_{10} depend on the balance parameter a.

May then compare

$$\begin{aligned} \mathrm{mse}_{\mathrm{ML}}(\delta) &= (\omega^{\mathsf{T}}\delta)^2 + \tau_0^2, \\ \mathrm{mse}_{\mathrm{HL},a}(\delta) &= (\omega_{\mathrm{HL},a}^{\mathsf{T}}\delta)^2 + \tau_{0,\mathrm{HL},a}^2, \end{aligned}$$

in different special setups.



E: Fine-tuning the balance parameter

The precision of $\widehat{\psi}_{\rm HL}$ for estimating $\psi_{\rm true}$ depends on the underlying truth and on the balance parameter a.

In the $f(y,\theta_0,\gamma_0+\delta/\sqrt{n})$ framework, the best balance a is the minimiser of

$$\text{risk}(a)=\text{mse}_{\text{HL},a}(\delta)=(\omega_{\text{HL},a}^{\mathsf T}\delta)^2+\tau_{0,\text{HL},a}^2.$$

Here

$$\begin{aligned}\omega_{\text{HL},a} &= L_{10,a}(J_a^*)^{-1}\frac{\partial\psi}{\partial\theta}-\frac{\partial\psi}{\partial\gamma},\\ \tau_{0,\text{HL},a}^2 &= \Big(\frac{\partial\psi}{\partial\theta}\Big)^{\mathsf T}(J_a^*)^{-1}K_a^*(J_a^*)^{-1}\frac{\partial\psi}{\partial\theta}.\end{aligned}$$

may be estimated consistently from data, with δ less visible:

$$D_n=\sqrt{n}(\widehat{\gamma}_{\text{ML}}-\gamma_0)\to_d \text{N}_r(\delta,Q),$$

with $Q=J^{11}$ from J_{wide}^{-1}.

Since $D_n = \sqrt{n}(\widehat\gamma_{\rm ML} - \gamma_0) \approx_d {\rm N}_r(\delta, Q)$, $D_n D_n^{\sf T}$ overestimates $\delta\delta^{\sf T}$, and

$$\mathbb{E}\,(c^{\sf T} D_n)^2 \doteq (c^{\sf T}\delta)^2 + c^{\sf T} Qc.$$

Hence we estimate the squared bias

$$\mathrm{sqb} = (\omega_{\mathrm{HL},a}^{\sf T}\delta)^2$$

in the 'FIC way', using

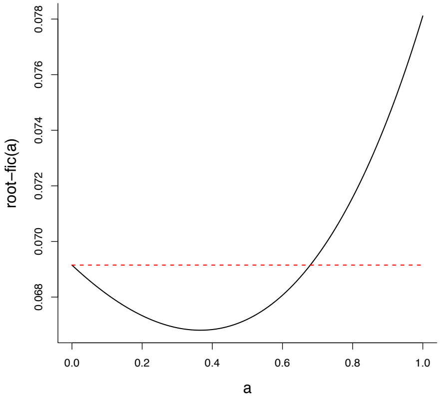
$$\begin{aligned} \widehat{\mathrm{sqb}} &= \max\{(\widehat\omega_{\mathrm{HL},a}^{\sf T} D_n)^2 - \widehat\omega_{\mathrm{HL},a}^{\sf T}\widehat Q\widehat\omega_{\mathrm{HL},a}, 0\} \\ &= \begin{cases} n\{\widehat\omega_{\mathrm{HL},a}^{\sf T}(\widehat\gamma_{\rm ML} - \gamma_0)\}^2 - \widehat\omega_{\mathrm{HL},a}^{\sf T}\widehat Q\widehat\omega_{\mathrm{HL},a} & \text{if nonnegative,} \\ 0 & \text{if else.} \end{cases} \end{aligned}$$

This leads to

$$\widehat{\mathrm{risk}}(a) = (\widehat{\frac{\partial\psi}{\partial\theta}})^{\sf T}(\widehat J_a^*)^{-1}\widehat K_a^*(\widehat J_a^*)^{-1}\widehat{\frac{\partial\psi}{\partial\theta}} + \widehat{\mathrm{sqb}}.$$

Via this FIC scheme we select balance parameter a as the minimiser of $\widehat{\mathrm{risk}}(a)$.

Example: $n = 100$ data points on $(0,1)$, fitted to $f(y,\theta) = \theta y^{\theta-1}$, with control parameter (now equal to the focus parameter) $\mu = \mathbb{E}\, Y^2$. FIC plot for selecting a in the HL estimation strategy:



F: Choosing the control parameters

The general hybrid likelihood estimation method is via constructing

$$h_n(\theta) = (1-a)\ell_n(\theta) + a \log R_n(\mu(\theta)),$$

which starts with choosing control parameters $\mu_1, \ldots, \mu_q$.

These aim at fitting models such that certain issues are well calibrated – outside those taken care of by the ML, which concentrates on the score functions $u_1(y,\theta), \ldots, u_p(y,\theta)$. Can choose $m(y,\mu) = g(y) - \mu$ to make sure that the HL incorporates aspects of $\mu = \mathbb{E}\, g(Y_i)$.

- Favourite case: For a given focus parameter $\psi = \psi(f)$, use this as the single control parameter.
- For a given focus parameter $\psi = \psi(f)$, may also select among candidate μ_j controls via FIC schemes.
- May 'stretch the idea', including a slowly increasing sequence of $\mu_1, \mu_2, \ldots$, with a FIC (or AFIC) stopping criterion.

G: Concluding remarks (and questions)

A. The methodology works for multidimensional data y_i, and can be extended to regression settings.

B. We fine-tune the balance parameter a by minimising the curve $\widehat{\mathrm{risk}}(a)$ over $[0,1]$. If the model gives a good fit, $\widehat{\mathrm{risk}}(a)$ is minimal at $a = 0$, and we use the ML, after all. This is also an implied goodness-of-fit test.

C. So far: large-sample approximation framework and methodology, with fixed

- p (dimension of θ),
- q (number of control parameters),
- r (number of extra γ_j model extension parameters).

It is of interest to let these grow with n – but more difficult mathematically.