The Hybrid Likelihood: Combining Parametric and Empirical Likelihoods

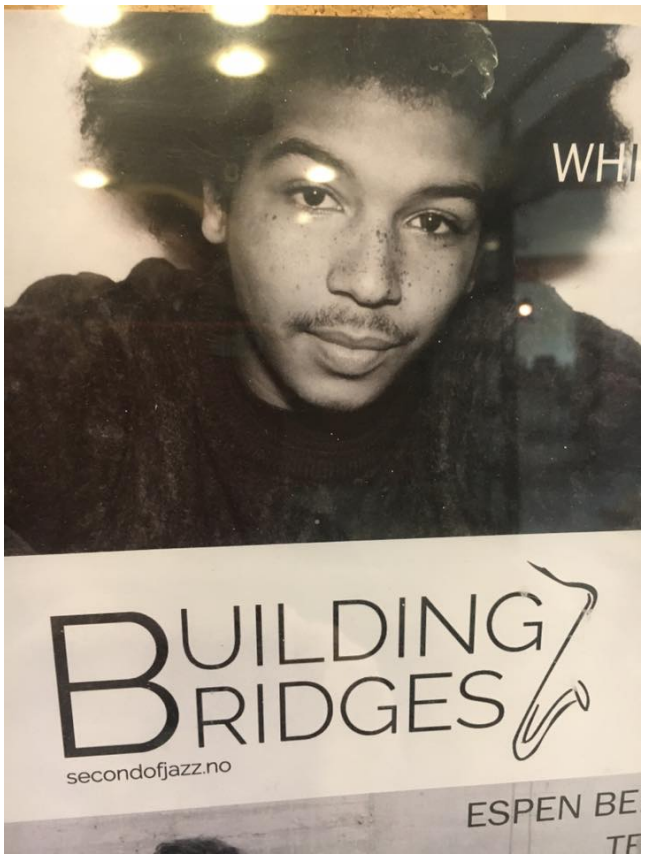
Nils Lid Hjort
(with Ingrid Van Keilegom and Ian McKeague)

Department of Mathematics, University of Oslo

Building Bridges (at Bislett)
May 2017, Teknologihuset

Bridging parametrics and nonparametrics

Building Bridges, connecting parametrics and nonparametrics: Here I discuss *one* such bridge operation (nonparametric extension of a parametric likelihood, or parametric focus for a nonparametric method). I use FIC tools (Focused Information Criteria) for fine-tuning and for selecting ingredients for the bridge.



Theme: Combining parametrics with nonparametrics

Suppose $y_1, \ldots, y_n$ i.i.d. f, with inference needed for focus parameter $\psi = \psi(f)$.

Parametric likelihood approach (perfect if model is perfect):

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

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

delta method giving

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

with $\kappa^2 = c^{\rm t} J(\theta)^{-1} c$ and $c = \partial\psi(\theta)/\partial\theta$. Also: Wilks theorem, χ^2_1.

Nonparametric likelihood approach (no conditions needed):

Identify ψ via ${\rm E}_f\, m(Y, \psi) = 0$. The empirical likelihood $R_n(\psi)$ is the maximum of $\prod_{i=1}^n (n w_i)$ under $\sum_{i=1}^n w_i = 1$, $\sum_{i=1}^n w_i m(y_i, \psi) = 0$, each $w_i > 0$. Then

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

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 $\mathrm{E}\, m_j(Y, \mu) = 0$ for $j = 1, \ldots, q$;
- put the parametric model through the EL, giving $R_n(\mu(\theta))$;

and form

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

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

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

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

I'll invent 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

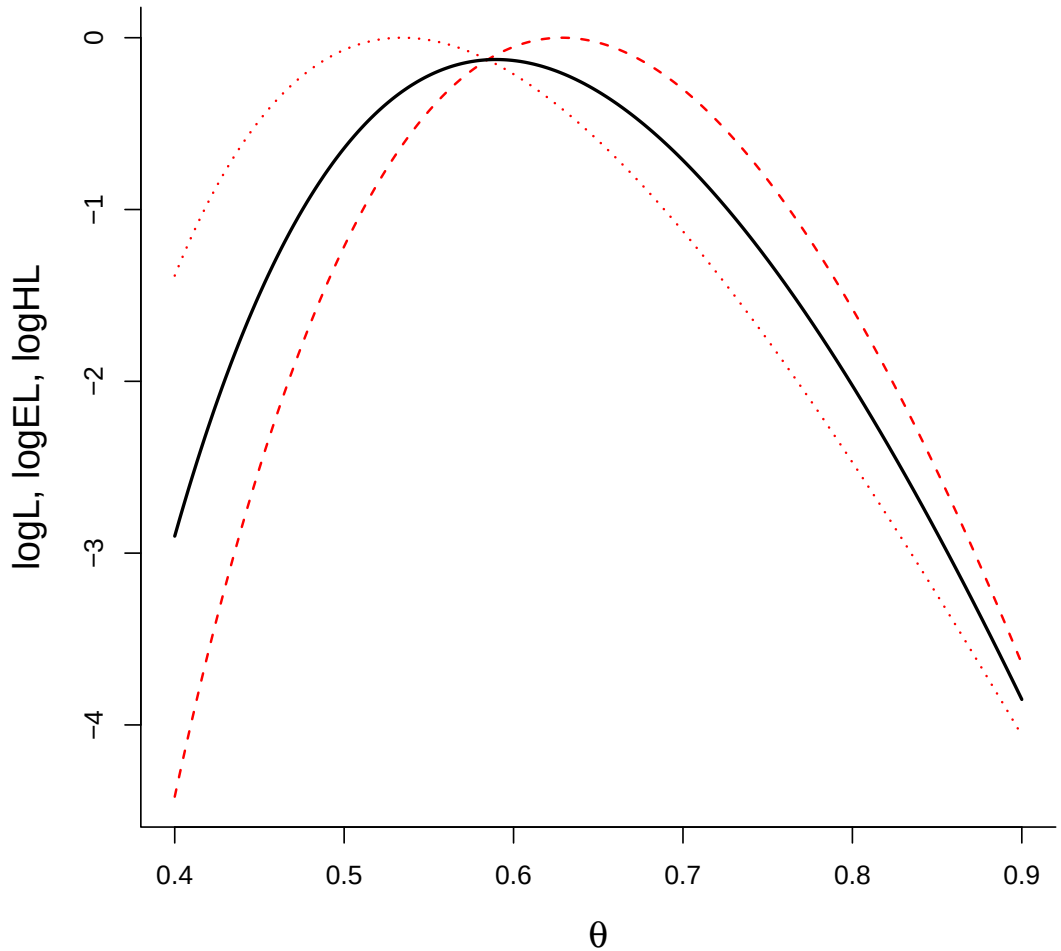
$$m_j(y, \mu_j) = I\{y \le \mu_j\} - j/4 \quad \text{for } j = 1, 2, 3.$$

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

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

Example 3. Consider $f(y,\theta) = \theta y^{\theta-1}$ on $(0,1)$. 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. I use Newcomb's 1889 speed of light data, with $n = 66$ and two grand outliers at -44 and -2. True value is 33.02.

ML ($a = 0$): bad estimates $(26.21, 10.75)$.

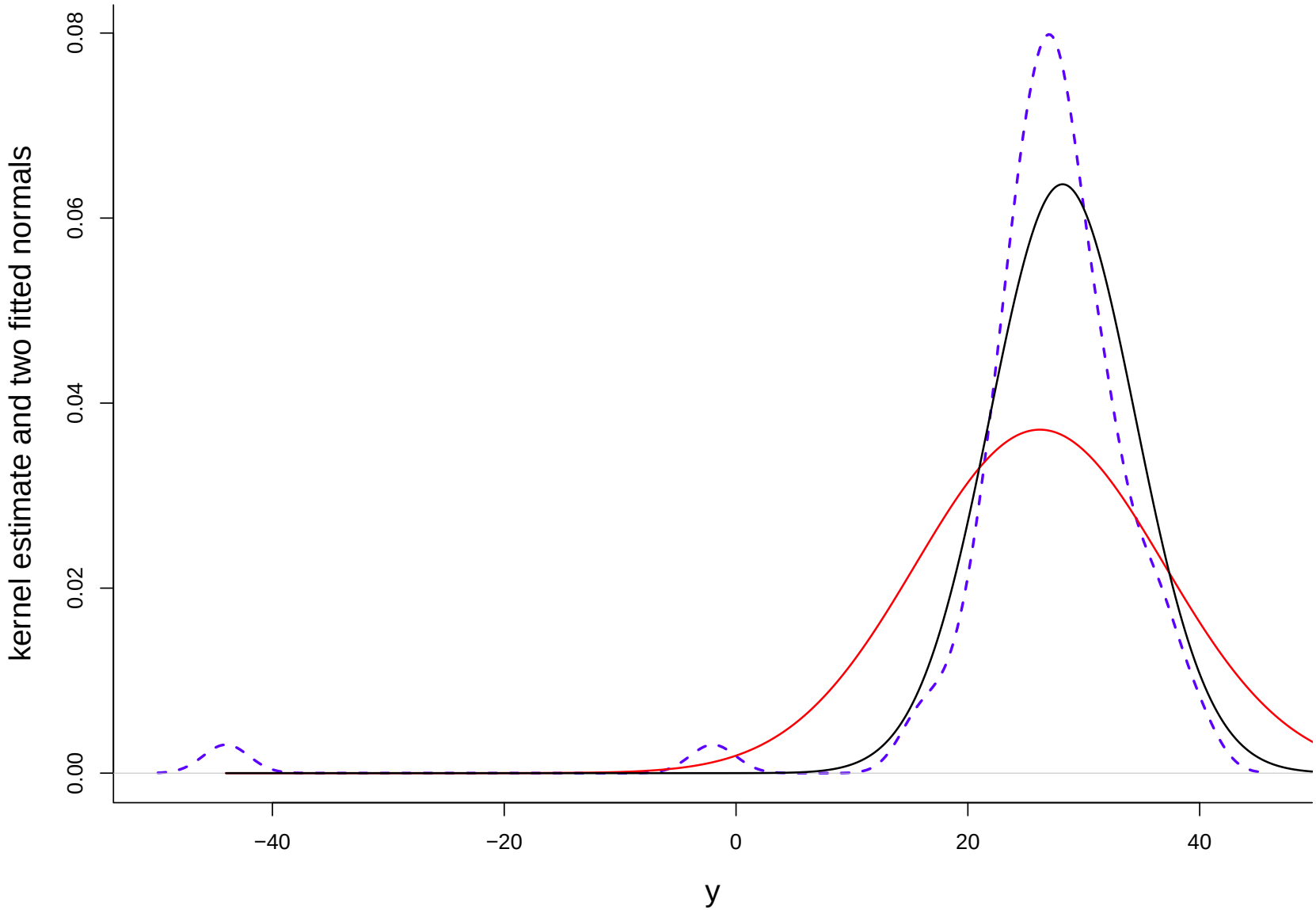
Removing outliers: $(27.75, 5.08)$.

I 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 I 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), I 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): I 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 I use the HL with $p+1$ parameters, the working model plus my focus ψ. – For the normal, I 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 (1-page course)

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

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

Let

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

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

First: $-2\log\Lambda_n = \max G_n = G_n(\widehat{\lambda})$.

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

$$-2\log\Lambda_n = V_n^{\mathrm{t}} W_n^{-1} V_n + o_{\mathrm{pr}}(1).$$

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

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

Understanding behaviour of $A_n \Longrightarrow$ understanding behaviour of $\widehat{\theta}_{\mathrm{HL}}$ (et al.). Under mild conditions, with $u(y,\theta)$ the score function:

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

Note that $J = J_{\mathrm{fish}}$ is the information matrix of the working model.

Lemma: there is a well-defined and well-behaved limiting quadratic process:

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

where

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

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

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

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

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

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

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

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

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

$$\sqrt{n}(\widehat{\psi}_{\mathrm{HL}} - \psi_0) \to_d c^{\mathrm{t}} \Lambda \sim \mathrm{N}(0, \kappa^2),$$

with $\kappa^2 = c^{\mathrm{t}} (J^*)^{-1} K^* (J^*)^{-1} c$, and $c = \partial \phi(\theta_0) / \partial \theta$.

May then compute $(J^*)^{-1} K^* (J^*)^{-1}$ and compare to J_{fish}^{-1}.

Result: The HL loses rather little compared to the ML, under model conditions, if say $a \leq 0.15$:

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

D: Theory: outside the model

Results so far: behaviour of $\widehat{\theta}_{\mathrm{HL}}$ and consequent $\widehat{\psi}_{\mathrm{HL}}$ well understood under parametric model conditions, where they 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 neighbourhood model framework:

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

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

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

with

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

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

$$\mathrm{mse}(\delta) = (b^{\mathrm{t}}\delta)^2 + \tau^2.$$

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

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

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

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

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

This implies

$$\sqrt{n}(\widehat{\psi}_{\text{ML}} - \psi_{\text{true}}) \to_d \text{N}(\omega^{\text{t}}\delta, \tau_0^2)$$

with

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

Hence we know

$$\text{mse}(\delta) = (\omega^{\text{t}}\delta)^2 + \tau_0^2$$

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

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

of dimension $q \times r$, along with

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

Then (via various efforts):

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

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

$$\sqrt{n}(\widehat{\psi}_{\mathrm{HL}} - \psi_{\mathrm{true}}) \to_d \mathrm{N}(\omega_{\mathrm{HL}}^{\mathrm{t}}\delta, \tau_{0,\mathrm{HL}}^2)$$

with

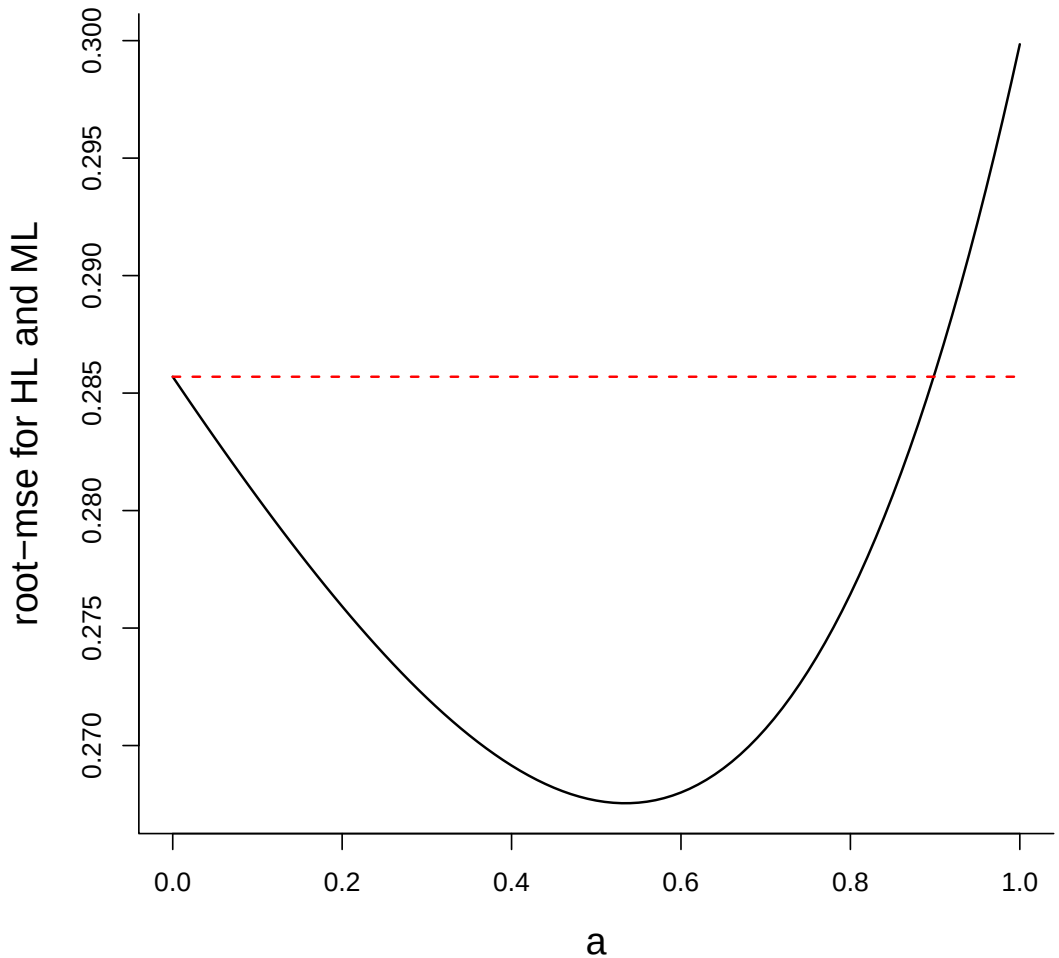
$$\begin{aligned}
\omega_{\mathrm{HL}} &= \omega_{\mathrm{HL},a} = L_{10}(J^*)^{-1}\frac{\partial\psi}{\partial\theta} - \frac{\partial\psi}{\partial\gamma}, \\
\tau_{0,\mathrm{HL}}^2 &= \tau_{0,\mathrm{HL},a}^2 = (\tfrac{\partial\psi}{\partial\theta})^{\mathrm{t}}(J^*)^{-1}K^*(J^*)^{-1}\frac{\partial\psi}{\partial\theta}.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

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

May then compare

$$\mathrm{mse}_{\mathrm{ML}}(\delta) = (\omega^{\mathrm{t}}\delta)^2 + \tau_0^2,$$
$$\mathrm{mse}_{\mathrm{HL},a}(\delta) = (\omega_{\mathrm{HL},a}^{\mathrm{t}}\delta)^2 + \tau_{0,\mathrm{HL},a}^2,$$

in different special setups.



E: Fine-tuning the balance parameter

The precision of $\widehat{\psi}_{\mathrm{HL}}$ for estimating ψ_{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

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

Here

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

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

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

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

Since $D_n = \sqrt{n}(\widehat{\gamma}_{\mathrm{ML}} - \gamma_0) \approx_d \mathrm{N}_r(\delta, Q)$, $D_n D_n^{\mathrm{t}}$ overestimates $\delta\delta^{\mathrm{t}}$, and

$$\mathrm{E}\,(c^{\mathrm{t}} D_n)^2 \doteq (c^{\mathrm{t}}\delta)^2 + c^{\mathrm{t}} Q c.$$

Hence we estimate the squared bias

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

in the 'FIC way', using

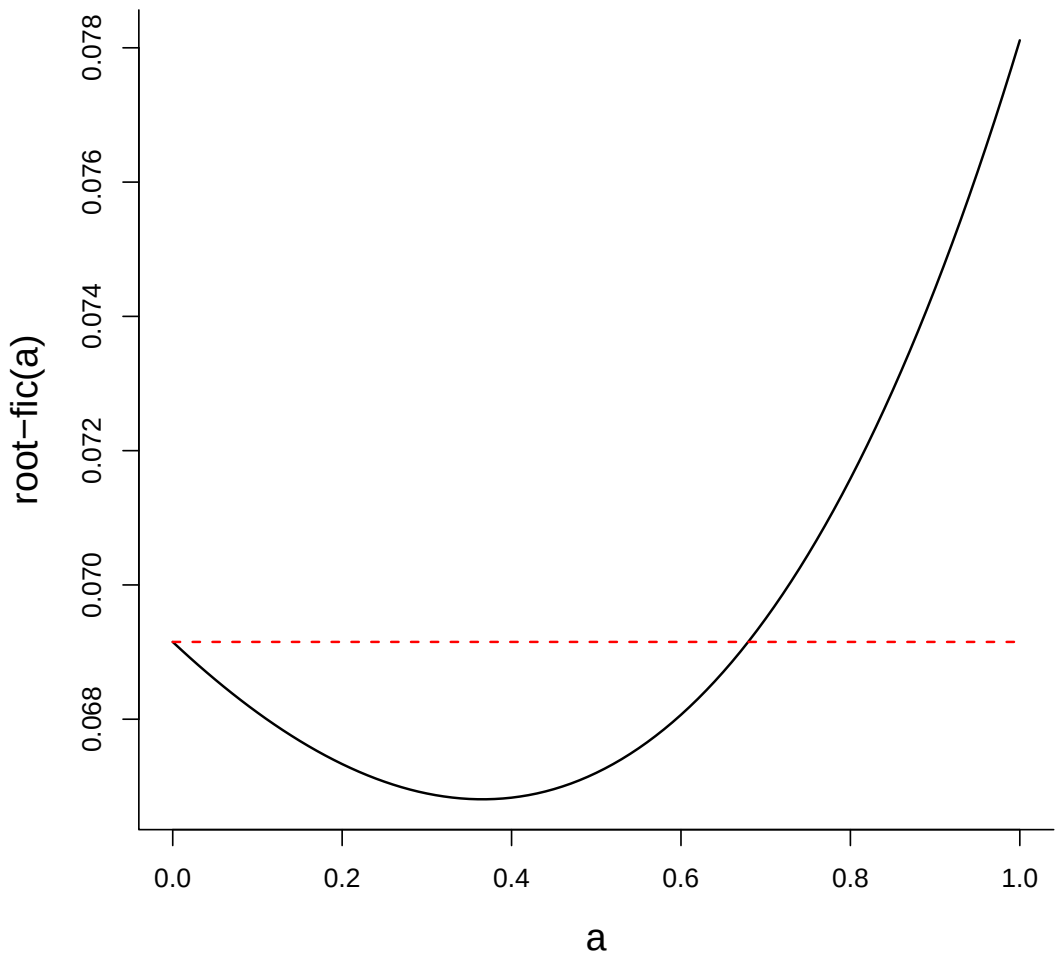
$$\begin{aligned}\widehat{\mathrm{sqb}} &= \max\{(\widehat{\omega}_{\mathrm{HL},a}^{\mathrm{t}} D_n)^2 - \widehat{\omega}_{\mathrm{HL},a}^{\mathrm{t}}\widehat{Q}\widehat{\omega}_{\mathrm{HL},a}, 0\} \\ &= \begin{cases} n\{\widehat{\omega}_{\mathrm{HL},a}^{\mathrm{t}}(\widehat{\gamma}_{\mathrm{ML}} - \gamma_0)\}^2 - \widehat{\omega}_{\mathrm{HL},a}^{\mathrm{t}}\widehat{Q}\widehat{\omega}_{\mathrm{HL},a} & \text{if nonnegative,} \\ 0 & \text{if else.} \end{cases}\end{aligned}$$

This leads to

$$\widehat{\mathrm{risk}}(a) = (\widehat{\frac{\partial\psi}{\partial\theta}})^{\mathrm{t}}(\widehat{J}_a^*)^{-1}\widehat{K}_a^*(\widehat{J}_a^*)^{-1}\widehat{\frac{\partial\psi}{\partial\theta}} + \widehat{\mathrm{sqb}}.$$

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) $\mu = \mathrm{E}\, Y^2$. FIC plot for selecting a in the HL estimation strategy:



Note: Sometimes $a = 0$ is best, i.e. choose ML.

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 $\mu = \mathrm{E}\, g(Y_i)$.

- 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.

D. Instead of $h_n(\theta) = (1-a)\ell_n(\theta) + a \log R_n(\mu(\theta))$, may work with a certain cousin operation,

$$\widetilde{h}_n(\theta) = (1-a)\ell_n(\theta) + a \log \widetilde{R}_n(\mu(\theta)),$$

with

$$\log \widetilde{R}_n(\mu(\theta)) = -\tfrac{1}{2} V_n(\mu(\theta))^{\mathrm{t}} W_n(\mu(\theta))^{-1} V_n(\mu(\theta)),$$

i.e. using the quadratic approximation to the EL rather than using the EL itself. This is (a) easier, numerically and mathematically; (b) equivalent, up to $O(1/\sqrt{n})$ neighbourhoods; (c) valid more generally.

This gives (with some efforts) a minimum divergence method based on

$$d(f, f_\theta) = (1-a)\mathrm{KL}(f, f_\theta) + a d_Q(f, f_\theta),$$

with

$$d_Q(f, f_\theta) = \tfrac{1}{2} \frac{(\mu_\theta - \mu_0)^{\mathrm{t}} \Omega_0^{-1} (\mu_\theta - \mu_0)}{1 + (\mu_\theta - \mu_0)^{\mathrm{t}} \Omega_0^{-1} (\mu_\theta - \mu_0)}.$$