UCL Université catholique de Louvain

INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

ISBA

DISCUSSION PAPER

2017/21

Hybrid combinations of parametric and empirical likelihoods

Hjort, N., McKeague, I. and I. Van Keilegom

Hybrid combinations of parametric and empirical likelihoods

Nils Lid Hjort[1], Ian W. McKeague[2], and Ingrid Van Keilegom[3]

University of Oslo, Columbia University, and KU Leuven

Abstract: This paper develops a hybrid likelihood (HL) method based on a compromise between parametric and nonparametric likelihoods. Consider the setting of a parametric model for the distribution of an observation Y with parameter θ. Suppose there is also an estimating function $m(\cdot, \mu)$ identifying another parameter μ via $\mathrm{E}\, m(Y, \mu) = 0$, at the outset defined independently of the parametric model. To borrow strength from the parametric model while obtaining a degree of robustness from the empirical likelihood method, we formulate inference about θ in terms of the hybrid likelihood function $H_n(\theta) = L_n(\theta)^{1-a} R_n(\mu(\theta))^a$. Here $a \in [0, 1)$ represents the extent of the compromise, L_n is the ordinary parametric likelihood for θ, R_n is the empirical likelihood function, and μ is considered through the lens of the parametric model. We establish asymptotic normality of the corresponding HL estimator and a version of the Wilks theorem. We also examine extensions of these results under misspecification of the parametric model, and propose methods for selecting the balance parameter a.

Key words and phrases: Agnostic parametric inference, Focus parameter, Semiparametric estimation, Robust methods

1 Some personal reflections on Peter

We are all grateful to Peter for his deeply influential contributions to the field of statistics, in particular to the areas of nonparametric smoothing, bootstrap, empirical likelihood (what this paper is about), functional data, high-dimensional data, measurement errors, etc., many of which were major breakthroughs in the area. His services to the profession were also exemplary and exceptional. It seems that he could simply not say 'no' to the many requests for recommendation letters, thesis reports, editorial duties, departmental reviews, and various other requests for help, and as many of us have experienced, he handled all this with an amazing speed, thoroughness and efficiency. We will also remember Peter as a very warm, gentle and humble person, who was particularly supportive to young people.

Nils Lid Hjort: I have many and uniformly warm remembrances of Peter. We had met and talked a few times at conferences, and then Peter invited me for a two-month stay in Canberra in 2000. This was both enjoyable, friendly and fruitful. I remember fondly not only technical discussions and the free-flowing of ideas on blackboards (and since Peter could think twice as fast as anyone else, that somehow improved my own arguing and thinking speed, or so I'd like to think), but also the positive, widely international, upbeat, but

[1]N.L. Hjort is supported via the Norwegian Research Council funded project FocuStat.
[2]I.W. McKeague is partially supported by NIH Grant R01GM095722.
[3]I. Van Keilegom is financially supported by the European Research Council (2016-2021, Horizon 2020, grant agreement No. 694409), and by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy).

unstressed working atmosphere. Among the pluses for my Down Under adventures was not merely meeting kangaroos in the wild while jogging and singing Schnittke, but teaming up with fellow visitors for several good projects, in particular with Gerda Claeskens; another sign of Peter's deep role in building partnerships and teams around him, by his sheer presence.

Then Peter and Jeannie visited us in Oslo for a six-week period in the autumn of 2003. For their first day there, at least Jeannie was delighted that I had put on my Peter Hall t-shirt and that I gave him a *Hall of Fame* wristwatch. For these Oslo weeks he was therefore elaboratedly introduced at seminars and meetings as *Peter Hall of Fame*; everyone understood that all other Peter Halls were considerably less famous. A couple of project ideas we developed together, in the middle of Peter's dozens and dozens of other ongoing real-time projects, are still in my drawers and files, patiently awaiting completion. Very few people can be as quietly and undramatically supremely efficient and productive as Peter. Luckily most of us others don't really have to, as long as we are doing decently well a decent proportion of the time. But once in a while, in my working life, when deadlines are approaching and I've lagged far behind, I put on my Peter Hall t-shirt, and think of him. It tends to work.

(a) (b)

Figure 1: (a) Peter at the occasion of his honorary doctorate at the Institute of Statistics in Louvain-la-Neuve in 1997; (b) Peter and Ingrid Van Keilegom in Tidbinbilla Nature Reserve near Canberra in 2002 (picture taking by Jeannie Hall).

Ingrid Van Keilegom: I first met Peter in 1995 during one of Peter's many visits to Louvain-la-Neuve (LLN). At that time I was still a graduate student at Hasselt University. Two years later, in 1997, Peter obtained an honorary doctorate from the Institute of Statistics in LLN (at the occasion of the fifth anniversary of the Institute), during which I discovered that Peter was not only a giant in his field but also a very human, modest and kind person. Figure 1(a) shows Peter at his acceptance speech. Later, in 2002, soon after I

started working as a young faculty member in LLN, Peter invited me to Canberra for six weeks, a visit of which I have extremely positive memories. I am very grateful to Peter for having given me the opportunity to work with him there. During this visit Peter and I started working on two papers, and although Peter was very busy with many other things, it was difficult to stay on top of all new ideas and material that he was suggesting and adding to the papers, day after day. At some point during this visit Peter left Canberra for a 10-day visit to London, and I (naively) thought I could spend more time on broadening my knowledge on the two topics Peter had introduced to me. However, the next morning I received a fax of 20 pages of hand-written notes, containing a difficult proof that Peter had found during the flight to London. It took me the full next 10 days to unraffle all the details of the proof! Although Peter was very focused and busy with his work, he often took his visitors on a trip during the weekends. I enjoyed very much the trip to the Tidbinbilla Nature Reserve near Canberra, together with him and his wife Jeannie. A picture taken in this park by Jeannie is seen in Figure 1(b).

After the visit to Canberra, Peter and I continued working on other projects, and in around 2004 Peter visited LLN for several weeks. I picked him up in the morning from the airport in Brussels. He came straight from Canberra and had been more or less 30 hours underway. I supposed without asking that he would like to go to the hotel to take a rest. But when we were approaching the hotel, Peter insisted that I would drive immediately to the Institute in order to start working straight away. He spent the whole day at the Institute discussing with people and working in his office, before going finally to his hotel! I always wondered where he found the energy, the motivation and the strength to do this. He will be remembered by many of us as an extremely hard working person, and as an example to all of us.

2 Introduction

For modelling data there are usually many options, ranging from purely parametric, semiparametric, to fully nonparametric. There are also numerous ways in which to combine parametrics with nonparametrics, say estimating a model density by a combination of a parametric fit with a nonparametric estimator, or by taking a weighted average of parametric and nonparametric quantile estimators. This article concerns a proposal for a bridge between a given parametric model and a nonparametric likelihood-ratio method. We construct a hybrid likelihood function, based on (i) the usual likelihood function for the parametric model, say $L_n(\theta)$, with n referring to sample size, as usual; and (ii) the empirical likelihood function for a given set of control parameters, say $R_n(\mu)$, where the μ parameters in question are "pushed through" the parametric model, leading to $R_n(\mu(\theta))$, say. Our hybrid likelihood $H_n(\theta)$, defined in (2) below, will be used for estimating the parameter vector of the working model; we term the $\widehat{\theta}_{\rm hl}$ in question the maximum hybrid likelihood

estimator. This in turn leads to estimates of other quantities of interest. If ψ is such a focus parameter, expressed via the working model as $\psi = \psi(\theta)$, then it will be estimated using $\widehat{\psi}_{\rm hl} = \psi(\widehat{\theta}_{\rm hl})$.

If the working parametric model is correct, these hybrid estimators will lose a certain amount in terms of efficiency, when compared to the usual maximum likelihood estimator. We shall demonstrate, however, that the efficiency loss under ideal model conditions is typically a small one, and that the hybrid estimators often will outperform the maximum likelihood when the working model is not correct. Thus the hybrid likelihood will be seen to offer parametric robustness, or protection against model misspecification, by borrowing strength from the empirical likelihood, via the selected control parameters.

Though our construction and methods can be lifted to e.g. regression models, see Section S.5 in the supplementary material, it is practical to start with the simpler i.i.d. framework, both for conveying the basic ideas and for developing theory. Thus let $Y_1, ..., Y_n$ be i.i.d. observations, stemming from some unknown density f. We wish to fit the data to some parametric family, say $f_\theta(y) = f(y, \theta)$, with $\theta = (\theta_1, \ldots, \theta_p)^{\rm t} \in \Theta$, where Θ is an open subset of $\mathbb{R}^p$. The purpose of fitting the data to the model is typically to make inference about certain quantities $\psi = \psi(f)$, termed *focus parameters*, but without necessarily trusting the model fully. Our machinery for constructing robust estimators for these focus parameters involves certain extra parameters, which we term *control parameters*, say $\mu_j = \mu_j(f)$ for $j = 1, \ldots, q$. These are context-driven parameters, selected to safeguard against certain types of model misspecification, and may or may not include the focus parameters. Suppose in general terms that $\mu = (\mu_1, \ldots, \mu_q)$ is identified via estimating equations, $\mathrm{E}_f\, m_j(Y, \mu) = 0$ for $j = 1, \ldots, q$. Now consider

$$R_n(\mu) = \max\Big\{\prod_{i=1}^n (n w_i) \colon \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i m(Y_i, \mu) = 0, \text{each } w_i > 0\Big\}. \tag{1}$$

This is the empirical likelihood function for μ, see Owen (2001), with further discussions in e.g. Hjort et al. (2009), Schweder and Hjort (2016, Ch. 11). One might e.g. choose $m(Y, \mu) = g(Y) - \mu$ for suitable $g = (g_1, \ldots, q_q)$, in which case the empirical likelihood machinery gives confidence regions for the parameters $\mu_j = \mathrm{E}_f\, g_j(Y)$. We can now introduce the *hybrid likelihood (HL) function*

$$H_n(\theta) = L_n(\theta)^{1-a} R_n(\mu(\theta))^a, \tag{2}$$

where $L_n(\theta) = \prod_{i=1}^n f(Y_i, \theta)$ is the ordinary likelihood, $R_n(\mu)$ is the empirical likelihood for μ, but here computed at the value $\mu(\theta)$, which is μ evaluated at f_θ, and with a being a balance parameter in $[0, 1)$. The associated maximum HL estimator is $\widehat{\theta}_{\rm hl}$, the maximiser of $H_n(\theta)$. If $\psi = \psi(f)$ is a parameter of interest, it is estimated as $\widehat{\psi}_{\rm hl} = \psi(f(\cdot, \widehat{\theta}_{\rm hl}))$. This means first expressing ψ in terms of the model parameters, say

$\psi = \psi(f(\cdot,\theta)) = \psi(\theta)$, and then plugging in the maximum HL estimator. Note that the general approach (2) works for multidimensional vectors Y_i, so the g_j functions could e.g. be set up to reflect covariances. For one-dimensional cases, options include $m_j(Y,\mu_j) = I\{Y \le \mu_j\} - j/q$ $(j = 1,\ldots,q-1)$ for quantile inference.

The hybrid method (2) provides a bridge from the purely parametric to the nonparametric empirical likelihood. The a parameter dictates the degree of balance. One may view (2) as a way for the empirical likelihood to borrow strength from a parametric family, and, alternatively, as a means of robustifying ordinary parametric model fitting by incorporating precision control for one or more μ_j parameters. There might be practical circumstances to assist one in selecting good μ_j parameters, or good estimating equations, or these may be singled out at the outset of the study as being quantities of primary interest.

A typical setting where the hybrid estimation method is attractive is where there is a single focus parameter of interest, say $\psi = \psi(f)$, identified via an estimating equation, $\mathrm{E}_f\, m(Y,\psi) = 0$. Here one may use the HL strategy above, with ψ taken to also play the role of the control parameter. In such a case, the HL estimator, say $\widehat{\theta}_{\mathrm{hl},a}$, can be read off as a full curve in a, starting at the ML estimator for $a = 0$ and then for $a \to 1$ tending to the EL based estimator.

Example 1. Let f_θ be the normal density with parameters (ξ,σ^2), and use $m_j(y,\mu_j) = I\{y \le \mu_j\} - j/4$ for $j = 1,2,3$. Then (1.1), with the ensuing $\mu_j(\xi,\sigma) = \xi + \sigma\,\Phi^{-1}(j/4)$ for $j = 1,2,3$, may be seen as estimating the normal parameters in a way which modifies the parametric ML method by taking into account the wish to have good model fit for the three quartiles. Alternatively, it may be viewed as a way of making inference for the three quartiles, borrowing strength from the normal family in order to hopefully do better than simply using the three empirical quartiles.

Example 2. Let f_θ be the Beta family with parameters (b,c), where ML estimates match moments for $\log Y$ and $\log(1-Y)$. Add to these the functions $m_j(y,\mu_j) = y^j - \mu_j$ for $j = 1,2$. Again, this is Beta fitting with modification for getting the mean and variance about right, or moment estimation borrowing strength from the Beta family.

Example 3. Consider the model $f(y,\theta) = \theta y^{\theta-1}$ for observations on the unit interval. The log-likelihood is $\ell_n(\theta) = n\{\log\theta - (\theta-1)Z_n\}$, with $Z_n = n^{-1}\sum_{i=1}^n \log(1/Y_i)$, and with $\widehat{\theta}_{\mathrm{ml}} = 1/Z_n$ the parametric ML estimator. The empirical log-likelihood for the mean μ, put through the model, for which $\mu = \theta/(\theta+1)$, yields the empirical log-likelihood $\log R_n(\mu(\theta))$. These may then be combined to the hybrid log-likelihood $\log H_n(\theta) = (1-a)\ell_n(\theta) + a\log R_n(\mu(\theta))$. Figure 2 displays these three log-likelihood functions for a set of $n = 50$ simulated data, with ML peaking at 0.629, EL at 0.534, and HL at 0.590, in this case using $\frac{1}{2}$-$\frac{1}{2}$ balance for $(1-a,a)$. Further variations may be worked through with empirical and hybrid log-likelihoods displayed for other functionals than the mean. We could compute these for the case of the median

$\mu' = (\frac{1}{2})^{1/\theta}$, and for the mean and the median (μ, μ') combined; in particular, the dimension q of estimating equations in (1), (2) is allowed to exceed the dimension p of the model parameter.

Example 4. Take your favourite parametric family $f(y,\theta)$, and for an appropriate data set specify an interval or region A that actually matters. Then use $m(y,p) = I\{y \in A\} - p$ as 'control equation' above, with $p = P\{Y \in A\} = \int_A f(y,\theta)\,dy$. The effect will be to push the parametric ML estimates, softly or not softly depending on the size of a, so as to make sure that the empirical binomial estimate $\widehat{p} = n^{-1}\sum_{i=1}^n I\{Y_i \in A\}$ is not far off from the estimated $p(A,\widehat{\theta}) = \int_A f(y,\widehat{\theta})\,\mathrm{d}y$. This can also be extended to using a partition of the sample space, say $A_1, \ldots, A_k$, with control equations $m_j(y,p) = I\{y \in A_j\} - p_j$ for $j = 1, \ldots, k-1$ (there is redundancy if trying to include also m_k). It will be seen via our theory, developed below, that the hybrid likelihood estimation strategy in this case is large-sample equivalent to maximising

$$(1-a)\ell_n(\theta) - \tfrac{1}{2}an\, r(Q_n(\theta)) = (1-a)\sum_{i=1}^n \log f(Y_i,\theta) - \tfrac{1}{2}an\frac{Q_n(\theta)}{1+Q_n(\theta)},$$

where $r(w) = w/(1+w)$ and $Q_n(\theta) = \sum_{j=1}^k \{\widehat{p}_j - p_j(\theta)\}^2/\widehat{p}_j$. Here $\widehat{p}_j$ is the direct empirical binomial estimate of $P\{Y \in A_j\}$ and $p_j(\widehat{\theta})$ the model-based estimate.

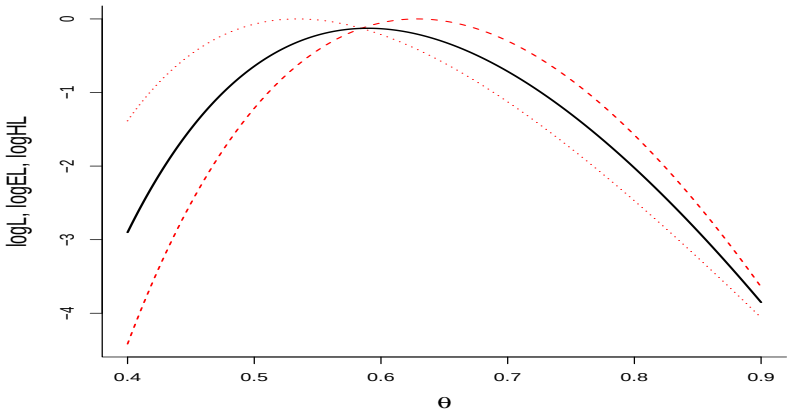

Figure 2: A dataset with $n = 50$ simulated observations on the unit interval is fitted to the model $f(y,\theta) = \theta y^{\theta-1}$, with mean $\mu(\theta) = \theta/(\theta+1)$; see Example 3. This gives rise to the log-likelihood $\ell_n(\theta)$ (dashed line), to the empirical likelihood $\log R_n(\mu(\theta))$ (dotted line), and the hybrid log-likelihood $\log H_n(\theta)$ (full curve).

In Section 3 we explore the basic properties of HL estimators and the ensuing $\widehat{\psi}_{\rm hl} = \psi(\widehat{\theta}_{\rm hl})$, under model conditions. Results here entail that the efficiency loss is typically small, and of size $O(a^2)$ in terms of the balance parameter a. In Section 4 we study the behaviour of HL in $O(1/\sqrt{n})$ neighbourhoods of the parametric model. It turns out that the HL estimator enjoys certain robustness properties, as compared to the ML estimator. Section 5 examines aspects related to fine-tuning the balance parameter a of (2), and we provide a recipe for its selection. An illustration of our HL methodology is given in Section 6, involving fitting a Gamma model to data of Roman era Egyptian life-lengths, a century BC.

Finally, coming back to the work of Peter Hall, a nice overview of all papers of Peter on EL can be found in Chang et al. (2017). We like to mention in particular the paper by DiCiccio et al. (1989), in which the features and behaviour of parametric and empirical likelihood functions are compared. Let us also mention that Peter also made very influential contributions to the somewhat related area of likelihood tilting, see e.g. Choi et al. (2000).

3 Behaviour of HL under the parametric model

The aim of this section is to explore asymptotic properties of the HL estimator under the parametric model $f(\cdot) = f(\cdot, \theta_0)$ for an appropriate true θ_0. We establish the local asymptotic normality of HL, the asymptotic normality of the estimator $\widehat{\theta}_{\rm hl}$, and a version of Wilks theorem. The HL estimator $\widehat{\theta}_{\rm hl}$ maximizes

$$h_n(\theta) = \log H_n(\theta) = (1-a)\ell_n(\theta) + a \log R_n(\mu(\theta)) \tag{3}$$

over θ, where $\ell_n(\theta) = \log L_n(\theta)$. We need to analyse the local behaviour of the two parts of $h_n(\cdot)$.

Consider the localised empirical likelihood $R_n(\mu(\theta_0 + s/\sqrt{n}))$, where s belongs to some compact $S \subset \mathbb{R}^p$. For simplicity of notation we write $m_{i,n}(s) = m(Y_i, \mu(\theta_0 + s/\sqrt{n}))$. Also, consider the functions $G_n(\lambda, s) = \sum_{i=1}^n 2\log\{1 + \lambda^{\rm t} m_{i,n}(s)/\sqrt{n}\}$ and $G_n^*(\lambda, s) = 2\lambda^{\rm t} V_n(s) - \lambda^{\rm t} W_n(s)\lambda$ of the q-dimensional λ, where

$$V_n(s) = n^{-1/2}\sum_{i=1}^n m_{i,n}(s) \quad \text{and} \quad W_n(s) = n^{-1}\sum_{i=1}^n m_{i,n}(s) m_{i,n}(s)^{\rm t}.$$

Note that G_n^* is the two-term Taylor expansion of G_n.

We now re-express the EL statistic in terms of Lagrange multipliers $\widehat{\lambda}_n$, which is pure analysis, not yet having anything to do with random variables, per se: $-2\log R_n(\mu(\theta_0 + s/\sqrt{n})) = \max_\lambda G_n(\lambda, s) = G_n(\widehat{\lambda}_n(s), s)$, with $\widehat{\lambda}_n(s)$ the solution to $\sum_{i=1}^n m_{i,n}(s)/\{1 + \lambda^{\rm t} m_{i,n}(s)/\sqrt{n}\} = 0$ for all s. This basic translation from the EL definition via Lagrange multipliers is contained in Owen (2001, Ch. 11); for a detailed proof, along with further discussion, see Hjort et al. (2009, Remark 2.7). The next lemma is crucial for understanding the basic properties of HL. The proof is in Section S.1 in the supplementary material. For any matrix $A = (a_{j,k})$, $\|A\| = (\sum_{j,k} a_{j,k}^2)^{1/2}$ denotes the Euclidean norm.

Lemma 1. *Suppose that (i)* $\sup_{s\in S}\|V_n(s)\| = O_{\rm pr}(1)$*; (ii)* $\sup_{s\in S}\|W_n(s) - W\| \to_{\rm pr} 0$*, where* $W = \operatorname{Var} m(Y, \mu(\theta_0))$ *is of full rank; (iii)* $n^{-1/2}\sup_{s\in S}\max_{i\le n}\|m_{i,n}(s)\| \to_{\rm pr} 0$*. Then, the maximisers* $\widehat{\lambda}_n(s) = \operatorname{argmax}_\lambda G_n(\lambda, s)$ *and* $\lambda_n^*(s) = \operatorname{argmax}_\lambda G_n^*(\lambda, s) = W_n^{-1}(s)V_n(s)$ *are both* $O_{\rm pr}(1)$ *uniformly in* $s \in S$*, and* $\sup_{s\in S}|\max_\lambda G_n(\lambda, s) - \max_\lambda G_n^*(\lambda, s)| = \sup_{s\in S}|G_n(\widehat{\lambda}_n(s), s) - G_n^*(\lambda_n^*(s), s)| \to_{\rm pr} 0$.

Note that we have an explicit expression for the maximiser of $G_n^*(\cdot, s)$, hence also its maximum, $\max_\lambda G_n^*(\lambda, s) = V_n(s)^{\rm t} W_n^{-1}(s) V_n(s)$. It follows that in situations covered by Lemma 1, $-2\log R_n(\mu(\theta_0 + s/\sqrt{n})) = V_n(s)^{\rm t} W_n^{-1}(s) V_n(s) + o_{\rm pr}(1)$, uniformly in $s \in S$. Also note that, by the law of large numbers, condition (ii) of Lemma 1 is valid if $\sup_s \|W_n(s) - W_n(0)\| \to_{\rm pr} 0$. If m and μ are smooth, then the latter holds using the mean value theorem. For the quantile example (see Example 1) we can use results on the oscillation behaviour of empirical distributions (see Stute (1982)).

For our Theorem 1 below we need assumptions on the $m(y, \mu)$ function involved in (1), and also on how $\mu = \mu(f_\theta) = \mu(\theta)$ is behaved close to θ_0. In addition to $\mathrm{E}\, m(Y, \mu(\theta_0)) = 0$, we assume that

$$\sup_{s \in S} \|V_n(s) - V_n(0) - \xi_n s\| = o_{\rm pr}(1), \tag{4}$$

with ξ_n of dimension $q \times p$ tending in probability to ξ_0. Suppose for illustration that $m(y, \mu(\theta))$ has a derivative at θ_0, and write $m(y, \mu(\theta_0 + \varepsilon)) = m(y, \mu(\theta_0)) + \xi(y)\varepsilon + r(y, \varepsilon)$, for the appropriate $\xi(y) = \partial m(y, \mu(\theta_0))/\partial\theta$, a $q \times p$ matrix, and with a remainder term $r(y, \varepsilon)$. This fits with (4), with $\xi_n = n^{-1} \sum_{i=1}^n \xi(Y_i) \to_{\rm pr} \xi_0 = \mathrm{E}\, \xi(Y)$, as long as $n^{-1/2} \sum_{i=1}^n r(Y_i, s/\sqrt{n}) \to_{\rm pr} 0$ uniformly in s. In smooth cases we would typically have $r(y, \varepsilon) = O(\|\varepsilon\|^2)$, making the mentioned term of size $O_{\rm pr}(1/\sqrt{n})$. On the other hand, when $m(y, \mu(\theta)) = I\{y \le \mu(\theta)\} - \alpha$, we have $V_n(s) - V_n(0) = f(\mu(\theta_0), \theta_0)s + O_{\rm pr}(n^{-1/4})$ uniformly in s (see Stute (1982)).

We rewrite the log-HL in terms of a local $1/\sqrt{n}$-scale perturbation around θ_0:

$$\begin{aligned} A_n(s) &= h_n(\theta_0 + s/\sqrt{n}) - h_n(\theta_0) \\ &= (1-a)\{\ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0)\} + a\{\log R_n(\mu(\theta_0 + s/\sqrt{n})) - \log R_n(\mu(\theta_0))\}. \end{aligned} \tag{5}$$

Below we show that $A_n(s)$ converges weakly to a quadratic limit $A(s)$ uniformly in s, which will then lead to our most important insights concerning HL based estimation and inference. By the multivariate central limit theorem,

$$\begin{pmatrix} U_{n,0} \\ V_{n,0} \end{pmatrix} = \begin{pmatrix} n^{-1/2} \sum_{i=1}^n u(Y_i, \theta_0) \\ n^{-1/2} \sum_{i=1}^n m(Y_i, \mu(\theta_0)) \end{pmatrix} \to_d \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \sim \mathrm{N}_{p+q}(0, \Sigma), \quad \text{where } \Sigma = \begin{pmatrix} J & C \\ C^{\rm t} & W \end{pmatrix}. \tag{6}$$

Here, $u(y, \theta) = \partial \log f(y, \theta)/\partial\theta$ is the score function, $J = \mathrm{Var}\, u(Y, \theta_0)$ is the Fisher information matrix of dimension $p \times p$, $C = \mathrm{E}\, u(Y, \theta_0) m(Y, \mu(\theta_0))^{\rm t}$ is of dimension $p \times q$, and $W = \mathrm{Var}\, m(Y, \mu(\theta_0))$ as before. The $(p+q) \times (p+q)$ variance matrix Σ is assumed to be positive definite. This ensures that the parametric and empirical likelihoods do not "tread on one another's toes", i.e. that the $m_j(y, \mu(\theta))$ functions are not in the span of the score functions, and vice versa.

Theorem 1. *Suppose that* $\sup_{\theta\in\mathcal{N}(\theta_0)} \max_{i,j,k\le q} |\mathrm{E}\,\partial^3 \log f(Y,\theta)/\partial\theta_i\partial\theta_j\partial\theta_k|$ *is finite, for some neighbourhood* $\mathcal{N}(\theta_0)$ *of* θ_0*, conditions (ii) and (iii) of Lemma 1 hold, condition (4) is valid with the appropriate* ξ_0*, and* Σ *has full rank. Then,* $A_n(s)$ *converges weakly to* $A(s) = s^{\mathrm{t}}U^* - \frac{1}{2}s^{\mathrm{t}}J^*s$ *as a process in* $s \in S$*, where*

$$U^* = (1-a)U_0 - a\xi_0^{\mathrm{t}}W^{-1}V_0 \quad \text{and} \quad J^* = (1-a)J + a\xi_0^{\mathrm{t}}W^{-1}\xi_0.$$

Here $U^* \sim \mathrm{N}_p(0, K^*)$*, with variance matrix* $K^* = (1-a)^2J + a^2\xi_0^{\mathrm{t}}W^{-1}\xi_0 - a(1-a)(CW^{-1}\xi_0 + \xi_0^{\mathrm{t}}W^{-1}C^{\mathrm{t}})$*. The convergence takes place in each function space* $\ell^\infty(S)$*, with* S *compact, and with the uniform topology; cf. van der Vaart and Wellner (1996, Ch. 1).*

The theorem, proved in Section S.2 of the supplementary material, is valid for each fixed balance parameter a in (2), with J^* and K^* also depending on a. We discuss ways of fine-tuning a in Section 5.

The $p\times q$-dimensional block component C of the variance matrix Σ of (6) can be worked with and represented in different ways. Suppose that μ is differentiable at $\theta=\theta_0$, and denote the vector of partial derivatives by $\frac{\partial\mu}{\partial\theta}$, with derivatives at θ_0, and with this matrix arranged as a $p\times q$ matrix, with columns $\partial\mu_j(\theta_0)/\partial\theta$ for $j=1,\ldots,q$. Note that from $\int m(y,\mu(\theta))f(y,\theta)\,\mathrm{d}y = 0$ for all θ follows the $q\times p$-dimensional equation $\int m^*(y,\mu(\theta_0))f(y,\theta_0)\,\mathrm{d}y\,(\frac{\partial\mu}{\partial\theta})^{\mathrm{t}} + \int m(y,\mu(\theta_0))f(y,\theta_0)u(y,\theta_0)^{\mathrm{t}}\,\mathrm{d}y = 0$, where $m^*(y,\mu) = \partial m(y,\mu)/\partial\mu$, in case m is differentiable with respect to μ. This means $C = -\frac{\partial\mu}{\partial\theta}\,\mathrm{E}_\theta\, m^*(Y,\mu(\theta_0))$. If $m(y,\mu) = g(y)-\mu$, for example, corresponding to parameters $\mu = \mathrm{E}\,g(Y)$, we have $C = \frac{\partial\mu}{\partial\theta}$. Also, using (4) we have $\xi_0 = -(\frac{\partial\mu}{\partial\theta})^{\mathrm{t}}$, of dimension $q\times p$. Applying Theorem 1 yields $U^* = (1-a)U_0 + a\frac{\partial\mu}{\partial\theta}W^{-1}V_0$, along with

$$J^* = (1-a)J + a\frac{\partial\mu}{\partial\theta}W^{-1}(\frac{\partial\mu}{\partial\theta})^{\mathrm{t}} \quad \text{and} \quad K^* = (1-a)^2J + \{1-(1-a)^2\}\frac{\partial\mu}{\partial\theta}W^{-1}(\frac{\partial\mu}{\partial\theta})^{\mathrm{t}}. \tag{7}$$

For the next corollary of Theorem 1, we need to introduce the random function $\Gamma_n(\theta) = n^{-1}\{h_n(\theta) - h_n(\theta_0)\}$ along with its population version

$$\Gamma(\theta) = -(1-a)\,\mathrm{KL}(f_{\theta_0}, f_\theta) - a\,\mathrm{E}\,\log\big(1+\xi(\theta)^{\mathrm{t}}m(Y,\mu(\theta))\big).$$

Here $\mathrm{KL}(f,f_\theta) = \int f\log(f/f_\theta)\,\mathrm{d}y$ the Kullback–Leibler divergence, in this case from f_{θ_0} to f_θ, and with $\xi(\theta)$ the solution of

$$\mathrm{E}\,\frac{m(Y,\mu(\theta))}{1+\xi^{\mathrm{t}}m(Y,\mu(\theta))} = 0$$

(that this solution exists and is unique is shown in the proof of Corollary 1 below). Then, $\widehat\theta_{\mathrm{hl}} = \operatorname{argmax}\Gamma_n(\cdot)$, and $\theta_0 = \operatorname{argmax}\Gamma(\cdot)$, as is again shown in the course of the proof.

Corollary 1. *Under the conditions of Theorem 1 and under conditions (A1)-(A3) given in Section S.3 of the supplementary material, there is consistency of* $\widehat{\theta}_{\rm hl}$ *towards* θ_0*, and*

$$\sqrt{n}(\widehat{\theta}_{\rm hl}-\theta_0) \to_d (J^*)^{-1}U^* \sim {\rm N}_p\big(0,(J^*)^{-1}K^*(J^*)^{-1}\big), \tag{8}$$

$$2\{h_n(\widehat{\theta}_{\rm hl})-h_n(\theta_0)\} \to_d (U^*)^{\rm t}(J^*)^{-1}U^*. \tag{9}$$

These results allow us to construct confidence regions for θ_0 and confidence intervals for its components. Of course we are not merely interested in the individual parameters of a model, but in certain functions of these, namely focus parameters. Assume $\psi=\psi(\theta)=\psi(\theta_1,\ldots,\theta_p)$ is such a parameter, with ψ differentiable at θ_0 and denote $c=\partial\psi(\theta_0)/\partial\theta$. The HL estimator for this ψ is the plug-in $\widehat{\psi}_{\rm hl}=\psi(\widehat{\theta}_{\rm hl})$. With $\psi_0=\psi(\theta_0)$ as the true parameter value, we then have via the delta method that

$$\sqrt{n}(\widehat{\psi}_{\rm hl}-\psi_0) \to_d c^{\rm t}(J^*)^{-1}U^* \sim {\rm N}(0,\kappa^2), \quad \text{where } \kappa^2=c^{\rm t}(J^*)^{-1}K^*(J^*)^{-1}c. \tag{10}$$

The focus parameter ψ could, for example, be one of the components of $\mu=\mu(\theta)$ used in the EL part of the HL, say μ_j, for which $\sqrt{n}(\widehat{\mu}_{j,{\rm hl}}-\mu_{0,j})$ has a normal limit with variance $(\frac{\partial\mu_j}{\partial\theta})^{\rm t}(J^*)^{-1}K^*(J^*)^{-1}\frac{\partial\mu_j}{\partial\theta}$, in terms of $\frac{\partial\mu_j}{\partial\theta}=\partial\mu_j(\theta_0)/\partial\theta$. Armed with (8), (9) and (10), we can set up Wald and likelihood-ratio type confidence regions and tests for θ, and confidence intervals for ψ. Consistent estimators $\widehat{J}^*$ and $\widehat{K}^*$ of J^* and K^* would then be required, but these are readily obtained via plug-in. Also, an estimate of J^* is typically obtained via the Hessian of the optimisation algorithm used to find the HL estimator in the first place.

In order to investigate how much is lost in efficiency when using the HL estimator under model conditions, consider the case of small a. We have $J^*=J+aA_1$ and $K^*=J+aA_2+O(a^2)$, with $A_1=\xi_0^{\rm t}W^{-1}\xi_0-J$ and $A_2=-2J-CW^{-1}\xi_0-\xi_0^{\rm t}W^{-1}C^{\rm t}$. For the class of functions of the form $m(y,\mu)=g(y)-T(\mu)$, corresponding to $\mu=T^{-1}({\rm E}\,g(Y))$, we have $A_2=2A_1$. It is assumed that $T(\cdot)$ has a continuous inverse at $\mu(\theta)$ for θ in a neighbourhood of θ_0. Hence

$$\begin{aligned}(J^*)^{-1}K^*(J^*)^{-1} &= (J^{-1}-aJ^{-1}A_1J^{-1})(J+aA_2)(J^{-1}-aJ^{-1}A_1J^{-1})+O(a^2)\\ &= J^{-1}+J^{-1}(A_2-2A_1)J^{-1}+O(a^2)=J^{-1}+O(a^2),\end{aligned}$$

which in particular means that the efficiency loss is a very small one when a is small.

4 Hybrid likelihood outside model conditions

In Section 3 we investigated the hybrid likelihood estimation strategy under the conditions of the parametric model. Under suitable conditions, the HL will be consistent and asymptotically normal, with a certain mild loss of efficiency under model conditions, compared to the parametric ML method, i.e. to the special case $a = 0$. In the present section we investigate the behaviour of the HL outside the conditions of the parametric model, which is now viewed as a working model. It will turn out that HL often outperforms ML by reducing model bias, which in mean squared error terms might more than compensate for a slight increase in variability. This in turn calls for methods for fine-tuning the balance parameter a in our basic hybrid construction (2), and we shall deal with this problem too, in Section 5.

Our framework for investigating such properties involves extending the working model $f(y, \theta)$ to a $f(y, \theta, \gamma)$ model, where $\gamma = (\gamma_1, \ldots, \gamma_r)$ is a vector of extra parameters. There is a null value $\gamma = \gamma_0$ which brings this extended model back to the working model. We shall examine behaviour of the ML and the HL schemes when γ is in the neighbourhood of γ_0. Suppose in fact that

$$f_{\text{true}}(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}), \tag{11}$$

with the $\delta = (\delta_1, \ldots, \delta_r)$ parameter dictating the relative distance from the null model. In this framework, suppose an estimator $\widehat{\theta}$ has the property that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}_p(B\delta, \Omega), \tag{12}$$

with a suitable $p \times r$ matrix B related to how much the model bias affects the estimator of θ, and limit variance matrix Ω. Then a parameter $\psi = \psi(f)$ of interest can in this wider framework be expressed as $\psi = \psi(\theta, \gamma)$, with true value $\psi_{\text{true}} = \psi(\theta_0, \gamma_0 + \delta/\sqrt{n})$. The spirit of these investigations is that the statistician uses the working model with only θ present, without knowing the extension model or the size of the δ discrepancy. The ensuing estimator for ψ is hence $\widehat{\psi} = \psi(\widehat{\theta}, \gamma_0)$. The delta method then leads to

$$\sqrt{n}(\widehat{\psi} - \psi_{\text{true}}) \to_d \mathrm{N}(b^{\mathrm{t}}\delta, \tau^2), \tag{13}$$

with $b = B^{\mathrm{t}} \frac{\partial \psi}{\partial \theta} - \frac{\partial \psi}{\partial \gamma}$ and $\tau^2 = (\frac{\partial \psi}{\partial \theta})^{\mathrm{t}} \Omega \frac{\partial \psi}{\partial \theta}$, and with partial derivatives evaluated at the working model, i.e. at (θ_0, γ_0). The limiting mean squared error, for such an estimator of μ, is

$$\mathrm{mse}(\delta) = (b^{\mathrm{t}}\delta)^2 + \tau^2. \tag{14}$$

Among the consequence of using the narrow working model when it is moderately wrong, at the level of $\gamma = \gamma_0 + \delta/\sqrt{n}$, is the bias $b^{\rm t}\delta$. Note that the size of this bias depends on the focus parameter, and that it may be zero for some foci, even when the model is incorrect.

We shall now examine both the ML and the HL methods in this framework, exhibiting the associated B and Ω matrices and hence the mean squared errors, via (13)–(14). Consider the parametric ML estimator $\widehat{\theta}_{\rm ml}$ first. To present the necessary results, consider the $(p+r)\times(p+r)$ Fisher information matrix

$$J_{\rm wide} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \tag{15}$$

for the $f(y,\theta,\gamma)$ model, computed at the null values (θ_0,γ_0). In particular, the $p\times p$ block J_{00}, corresponding to the model with only θ and without γ, is equal to the earlier J matrix of (6) and appearing in Theorem 1 etc. Here one may demonstrate, under appropriate mild regularity conditions, that $\sqrt{n}(\widehat{\theta}_{\rm narr}-\theta_0) \to_d {\rm N}_p(J_{00}^{-1}J_{01}\delta, J_{00}^{-1})$. Just as (13) followed from (12), one finds for $\widehat{\psi}_{\rm ml}=\psi(\widehat{\theta}_{\rm ml})$ that

$$\sqrt{n}(\widehat{\psi}_{\rm ml}-\psi_{\rm true}) \to_d {\rm N}(\omega^{\rm t}\delta, \tau_0^2), \tag{16}$$

featuring $\omega = J_{10}J_{00}^{-1}\frac{\partial\psi}{\partial\theta} - \frac{\partial\psi}{\partial\gamma}$ and $\tau_0^2 = (\frac{\partial\psi}{\partial\theta})^{\rm t}J_{00}^{-1}\frac{\partial\psi}{\partial\theta}$. See also Hjort and Claeskens (2003) and Claeskens and Hjort (2008, Chs. 6, 7) for further details, discussion, and precise regularity conditions.

In such a situation, with a clear interest parameter ψ, using the HL construction, we get $\widehat{\psi}_{\rm hl}=\psi(\widehat{\theta}_{\rm hl},\gamma_0)$. We shall work out what happens with $\widehat{\theta}_{\rm hl}$, in this framework (generalising what is found in the previous section). For the following, introduce $S(y)=\partial\log f(y,\theta_0,\gamma_0)/\partial\gamma$, the score function of the extended model in direction of these extension parameters, and let $K_{01}=\int f(y,\theta_0)m(y,\mu(\theta_0))S(y)\,{\rm d}y = {\rm E}\,m(Y,\mu(\theta_0))S(Y)$, of dimension $q\times r$, along with $L_{01}=(1-a)J_{01}-a(\frac{\partial\psi}{\partial\theta})^{\rm t}W^{-1}K_{01}$, of dimension $p\times r$, and with transpose $L_{10}=L_{01}^{\rm t}$. We then have the following result, proved in Section S.4 in the supplementary material.

Theorem 2. *Assume data stem from the extended $p+r$-dimensional model (11), and that the conditions listed in Corollary 1 are in force. For the HL method, with the focus parameter $\psi=\psi(f)$ built into the construction (2), results (12)–(14) hold, with $B=(J^*)^{-1}L_{01}$ and $\Omega=(J^*)^{-1}K^*(J^*)^{-1}$.*

The limiting distribution for $\widehat{\psi}_{\rm hl}=\psi(\widehat{\theta}_{\rm hl})$ can again be read off, just as (13) follows from (12):

$$\sqrt{n}(\widehat{\psi}_{\rm hl}-\psi_{\rm true}) \to_d {\rm N}(\omega_{\rm hl}^{\rm t}\delta, \tau_{0,{\rm hl}}^2), \tag{17}$$

with $\omega_{\rm hl}=L_{10}(J^*)^{-1}\frac{\partial\psi}{\partial\theta}-\frac{\partial\psi}{\partial\gamma}$ and $\tau_{0,{\rm hl}}^2=(\frac{\partial\psi}{\partial\theta})^{\rm t}(J^*)^{-1}K^*(J^*)^{-1}\frac{\partial\psi}{\partial\theta}$.

Note that the quantities involved in describing these large-sample properties for the HL estimator, as with J^* and K^* in (8)–(10) and the $\omega_{\rm hl}$ and $\tau_{0,\rm hl}$ involved in (17), depend on the balance parameter a employed in the basic HL construction (2). For $a = 0$ we are back to the ML, with (17) specialising to (16). When a moves away from zero, more emphasis is given to the EL part, in effect to push θ to get $n^{-1}\sum_{i=1}^n m(Y_i, \mu(\theta))$ closer to zero. The result is typically a lower bias $|\omega_{\rm hl}(a)^{\rm t}\delta|$, compared to $|\omega^{\rm t}\delta|$, and a slightly larger standard deviation $\tau_{0,\rm hl}$, compared to τ_0. Thus selecting a good value of a is a bias-variance balancing game, which we discuss in the following section.

5 Fine-tuning the balance parameter

The basic HL construction of (2) first entails selecting context relevant control parameters $\mu_1, \ldots, \mu_q$, and then a focus parameter ψ. A special case is that of using the focus ψ as the single control parameter. In each case, there is also the balance parameter a to decide upon. Ways of fine-tuning the balance are discussed in this section.

5.1 Balancing robustness and efficiency

By allowing the empirical likelihood to be combined with the likelihood from a given parametric model, one may buy robustness, via the control parameters μ in the HL construction, at the expense of a certain mild loss of efficiency. One way to fine-tune the balance, after having decided on the control parameters, is to select a so that the loss of efficiency under the conditions of the parametric working model is limited by a fixed, small amount, say 5%. This may be achieved by using the corollaries of Section 3 by comparing the inverse Fisher information matrix J^{-1}, associated with the ML estimator, to the sandwich matrix $(J_a^*)^{-1}K_a^*(J_a^*)^{-1}$, for the HL estimator. Here we refer to the corollaries of Section 3, see e.g. (7), and have added the subscript a, for emphasis. If there is special interest in some focus parameter ψ, one may select a so that

$$\kappa_a = \{c^{\rm t}(J_a^*)^{-1}K_a^*(J_a^*)^{-1}c\}^{1/2} \le (1+\varepsilon)\kappa_0 = (1+\varepsilon)(c^{\rm t}J^{-1}c)^{1/2}, \tag{18}$$

with ε the required threshold. With $\varepsilon = 0.05$, for example, one ensures that confidence intervals are only 5% wider than those based on the ML, but with the additional security of having controlled well for the μ parameters in the process, e.g. for robustness reasons. Pedantically speaking, in (10) there is really a $c_a = \partial\psi(\theta_{0,a})/\partial\theta$ also depending on the a, associated with the limit in probability $\theta_{0,a}$ of the HL estimator, but in the present case, discussing efficiencies at the parametric model, the $\theta_{0,a}$ is the same as the true θ_0, so c_a is the same as $c = \partial\psi(\theta_0)/\partial\theta$. A concrete illustration of this approach is in the following section.

5.2 Features of the mse(a)

The methods above, as with (18), rely on the theory developed in Section 3, under the conditions of the parametric working model. In what follows we need the theory given in Section 4, examining the behaviour of the HL estimator in a neighbourhood around the working model. Results there can first be used to examine the limiting mse properties for the ML and the HL estimators where it will be seen that the HL often can behave better; a slightly larger variance is being compensated for with a smaller modelling bias. Secondly, the mean squared error curve, as a function of the balance parameter a, can be estimated from data. This leads to the idea of selecting a to be the minimiser of this estimated risk curve, pursued in Section 5.3 below.

For given focus parameter ψ, the limit mse when using the HL with parameter a is found from (17):

$$\text{mse}(a) = \{\omega_{\text{hl}}(a)^{\text{t}}\delta\}^2 + \tau_{0,\text{hl}}(a)^2. \tag{19}$$

The first exercise is to evaluate this curve, as a function of the balance parameter a, in situations with given model extension parameter δ. The mse(a) at $a = 0$ corresponds to the mse for the ML estimator. If mse(a) is smaller than mse(0), for some a, then the HL is doing a better job than the ML.

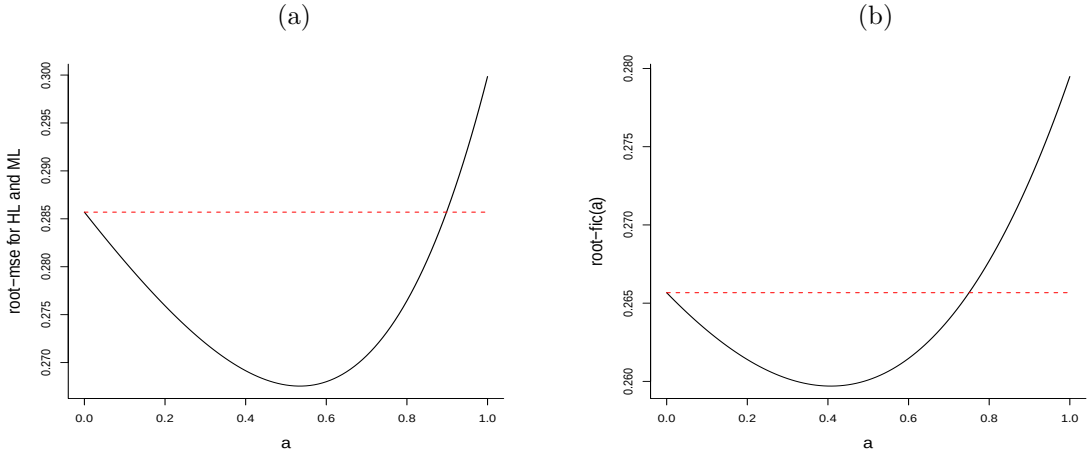

Figure 3: (a) The dotted horizontal line indicates the root-mse for the ML estimator, and the full curve the root-mse for the HL estimator, as a function of the balance parameter a in the HL construction. (b) The root-fic(a), as a function of the balance parameter a, constructed on the basis of $n = 100$ simulated observations, from a case where $\gamma = 1 + \delta/\sqrt{n}$, with δ described in the text.

Figure 3(a) displays the root-mse(a) curve in a simple setup, where the parametric start model is the Beta$(\theta, 1)$, i.e. with density $\theta y^{\theta-1}$, and the focus parameter used for the HL construction is $\psi = \text{E}\, Y^2$, which is $\theta/(\theta + 2)$ under model conditions. The extended model, under which we examine the mse properties of the ML and the HL, is the Beta(θ, γ), with $\gamma = 1 + \delta/\sqrt{n}$ in (11). The δ for this illustration is chosen to be $Q^{1/2} = (J^{11})^{1/2}$, from (20) below, which may be interpreted as one standard deviation away from

the null model. The root-mse(a) curve, computed via numerical integration, shows that the HL estimator $\widehat{\theta}_{\rm hl}/(\widehat{\theta}_{\rm hl}+2)$ does better than the parametric ML estimator $\widehat{\theta}_{\rm ml}/(\widehat{\theta}_{\rm ml}+2)$, unless a is close to 1. Similar curves are seen for other δ, for other focus parameters, and for more complex models. Occasionally, mse(a) is increasing in a, indicating in such cases that ML is better than HL, but this typically happens only when the model discrepancy parameter δ is small, i.e. when the working model is nearly correct.

It is of interest to note that $\omega_{\rm hl}(a)$ in (17) starts out for $a=0$ at $\omega = J_{10}J_{00}^{-1}\frac{\partial\psi}{\partial\theta} - \frac{\partial\psi}{\partial\gamma}$ in (16), associated with the ML method, but then it decreases in size towards zero, as a grows from zero to one. Hence, when HL employs only a small part of the ordinary log-likelihood in its construction, the consequent $\widehat{\psi}_{{\rm hl},a}$ has small bias, but potentially a bigger variance than ML. The HL may thus be seen as a debiasing operation, for the control and focus parameters, in cases where the parametric model $f(\cdot,\theta)$ cannot be fully trusted.

5.3 Estimation of mse(a)

Concrete evaluation of the mse(a) curves of (19) shows that the HL scheme typically is worthwhile, in that the mse is lower than that of the ML, for a range of a values. To find a good value of a from data, a natural idea is to estimate the mse(a) and then pick its minimiser. For mse(a), the ingredients $\omega_{\rm hl}(a)$ and $\tau_{0,{\rm hl}}(a)$ involved in (17) may be estimated consistently via plug-in of the relevant quantities. The difficulty lies with the δ part, and more specifically with $\delta\delta^{\rm t}$ in $\omega_{\rm hl}(a)\delta\delta^{\rm t}\omega_{\rm hl}(a)$. For this parameter, defined on the $O(1/\sqrt{n})$ scale via $\gamma=\gamma_0+\delta/\sqrt{n}$, the essential information lies in $D_n=\sqrt{n}(\widehat{\gamma}_{\rm ml}-\gamma_0)$, via parametric ML estimation in the extended $f(y,\theta,\gamma)$ model. As demonstrated and discussed in Claeskens and Hjort (2008, Chs. 6–7), in connection with construction of their Focused Information Criterion (FIC), we have

$$D_n \to_d D \sim {\rm N}_r(\delta, Q), \quad \text{with } Q = J^{11} = (J_{11}-J_{10}J_{00}^{-1}J_{01})^{-1}. \tag{20}$$

The factor $\delta/\sqrt{n}$ in the $O(1/\sqrt{n})$ construction cannot be estimated consistently. Since $DD^{\rm t}$ has mean $\delta\delta^{\rm t}+Q$ in the limit, we estimate squared bias parameters of the type $(b^{\rm t}\delta)^2 = b\delta\delta^{\rm t}b$ using $\{b^{\rm t}(D_nD_n^{\rm t}-\widehat{Q})b\}_+$, in which $\widehat{Q}$ estimates $Q=J^{11}$, and $x_+=\max(x,0)$. We construct the $r\times r$ matrix $\widehat{Q}$ from estimating and then inverting the full $(p+r)\times(p+r)$ Fisher information matrix $J_{\rm wide}$ of (15). This leads to estimating mse(a) using

$$\text{fic}(a) = \{\widehat{\omega}_{\rm hl}(a)^{\rm t}(D_nD_n^{\rm t}-\widehat{Q})\widehat{\omega}_{\rm hl}(a)\}_+ + \widehat{\tau}_{0,{\rm hl}}(a)^2 = \big[n\widehat{\omega}_{\rm hl}(a)^{\rm t}\{(\widehat{\gamma}-\gamma_0)(\widehat{\gamma}-\gamma_0)^{\rm t}-\widehat{Q}\}\widehat{\omega}_{\rm hl}(a)\big]_+ + \widehat{\tau}_{0,{\rm hl}}(a)^2.$$

Figure 3(b) displays such a root-fic curve, the estimated root-mse(a) as a function of the balance parameter a. Whereas the root-mse(a) curve shown in Figure 3(a) is coming from considerations and numerical

investigation of the extended $f(y,\theta,\gamma)$ model alone, pre-data, the root-fic(a) curve is constructed for a given dataset. The start model and its extension are as with Figure 3(a), a Beta$(\theta,1)$ inside a Beta(θ,γ), with $n=100$ simulated data points using $\gamma=1+\delta/\sqrt{n}$ with δ chosen as for Figure 3(a). Again, the HL method was applied, using the second moment $\psi=\mathrm{E}\,Y^2$ as both control and focus. The estimated risk is smallest for $a=0.41$, and we would use this as the balance parameter.

6 An illustration: Roman era Egyptian life-lengths

A fascinating dataset on $n=141$ life-lengths from Roman era Egypt, a century BC, is examined in Pearson (1902), where he compares life-length distributions from two societies, two thousand years apart. The data are also discussed, modelled and analysed in (Claeskens and Hjort, 2008, Ch. 2).

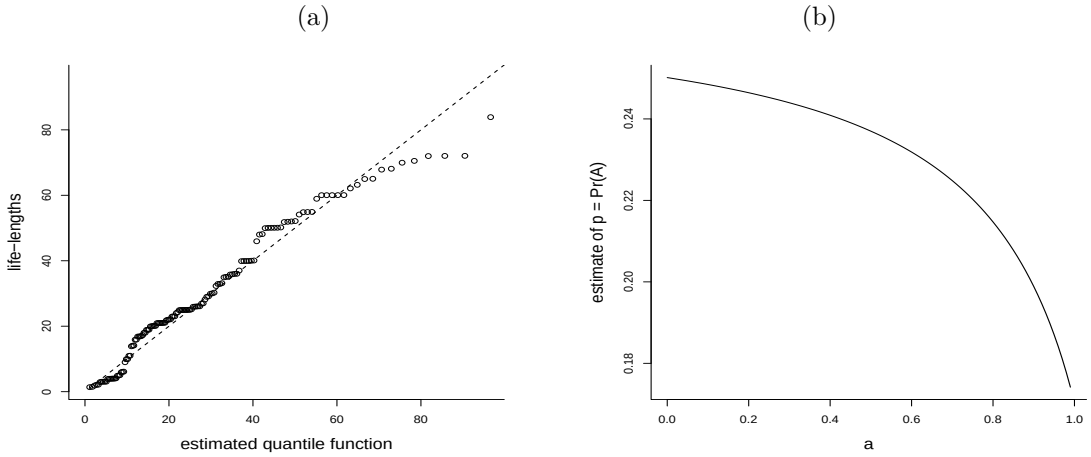

Figure 4: (a) The q-q plot shows the ordered life-lengths $y_{(i)}$ plotted against the ML-estimated gamma quantile function $F^{-1}(i/(n+1),\widehat{b},\widehat{c})$. (b) The curve $\widehat{p}_a$, with the probability $p=P\{Y\in[9.5,20.5]\}$ estimated via the HL estimator, is displayed, as a function of the balance parameter a. At balance position $a=0.61$, the efficiency loss is 10% compared to the ML precision under ideal gamma model conditions.

Here we have fitted the data to the Gamma(b,c) distribution, first using the ML, with parameter estimates $(1.6077,0.0524)$. The q-q plot of Figure 4(a) displays the points $(F^{-1}(i/(n+1),\widehat{b},\widehat{c}),y_{(i)})$, with $F^{-1}(\cdot,b,c)$ denoting the quantile function of the Gamma and $y_{(i)}$ the ordered life-lengths, from 1.5 to 96. We learn that the gamma distribution does a decent job for these data, but that the fit is not good for the longer lives. There is hence scope for the HL for estimating and assessing relevant quantities in a more robust and indeed controlled fashion than via the ML. Here we focus on $p=p(b,c)=P\{Y\in[L_1,L_2]\}=\int_{L_1}^{L_2}f(y,b,c)\,\mathrm{d}y$, for age groups $[L_1,L_2]$ of interest. The hybrid log-likelihood is hence $h_n(b,c)=(1-a)\ell_n(b,c)+a\log R_n(p(b,c))$, with $R_n(p)$ being the EL associated with $m(y,p)=I\{y\in[L_1,L_2]\}-p$. We may then, for each balance parameter a, maximise this function and read off both the HL estimates $(\widehat{b}_a,\widehat{c}_a)$ and the consequent $\widehat{p}_a=p(\widehat{b}_a,\widehat{c}_a)$.

Figure 4(b) displays this $\widehat{p}_a$, as a function of the balance parameter, for the age group $[9.5, 20.5]$. For $a = 0$ we have the ML based estimate 0.251, and with increasing a there is more weight to the EL, which has the point estimate 0.171.

To decide on a good balance, recipes of Section 5 may be appealed to. The relatively speaking simplest of these is that associated with (18), where we numerically compute $\kappa_a = \{c^{\rm t}(J^*)^{-1}K^*(J^*)^{-1}c\}^{1/2}$ for each a, at the ML position in the parameter space of (b, c), and with J^* and K^* from (7). The loss of efficiency κ_a/κ_0 is quite small for small a, and is at level 1.10 for $a = 0.61$. For this value of a, where confidence intervals are stretched 10% compared to the gamma-model-based ML solution, we find $\widehat{p}_a$ equal to 0.232, with estimated standard deviation $\widehat{\kappa}_a/\sqrt{n} = 0.188/\sqrt{n} = 0.016$. Similarly the HL machinery may be put to work for other age intervals, for each such using the $p = P\{Y \in [L_1, L_2]\}$ as both control and focus, and for models other than the gamma. We may employ the HL with a collection of control parameters, like age groups, before landing on inference for a focus parameter; see Example 4. The more elaborate recipe of selecting a, developed in Section 5.3 and using fic(a), can also be used here, but would involve the use of an extended parametric model.

7 Further developments and the Supplementary Material section

Various concluding remarks and extra developments are placed in the article's Supplementary Material section. In particular, proofs of Lemma 1, Theorems 1 and 2 and Corollary 1 are given there. Other material involves (i) the important extension of the basic HL construction from i.i.d. data to regression type data, in Section S.5; (ii) log-HL-profiling operations and a deviance fuction, leading to a full confidence curve for a focus parameter, in Section S.6; (iii) an implicit goodness-of-fit test for the parametric vehicle model, in Section S.7; and finally (iv) a related but different hybrid likelihood construction, in Section S.8.

References

Chang, J., Guo, J., and Tang, C. Y. (2017). Peter Hall's contribution to empirical likelihood. *Statistica Sinica*, **xx**, xx–xx.

Choi, E., Hall, P., and Presnell, B. (2000). Rendering parametric procedures more robust by empirically tilting the model. *Biometrika*, **87**, 453–465.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

DiCiccio, T., Hall, P., and Romano, J. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika*, **76**, 465–476.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association*, **98**, 879–899.

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics*, **37**, 1079–1111.

Molanes López, E., Van Keilegom, I., and Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion function. *Scandinavian Journal of Statistics*, **36**, 413–432.

Owen, A. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, London.

Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika*, **1**, 261–264.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, **61**, 123–137.

Stute, W. (1982). The oscillation behavior of empirical processes. *Annals of Probability*, **10**, 86–107.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Supplementary material

This additional section contains the following sections. Sections S.1, S.2, S.3, S.4 give the technical proofs of Lemma 1, Theorem 1, Cororally 1 and Theorem 2. Then Section S.5 crucially indicates how the HL methodology can be lifted from the i.i.d. case to regression type models, whereas a Wilks type theorem based on HL-profiling, useful for constructing confidence curves for focus parameters, is developed in Section S.6. An implicit goodness-of-fit test for the parametric working model is identified in Section S.7. Finally Section S.8 describes an alternative hybrid approach, related to, but different from the HL. This alternative method is first-order equivalent to the HL method inside $O(1/\sqrt{n})$ neighbourhoods of the parametric vehicle model, but not at farther distances.

S.1 Proof of Lemma 1

The proof is based on techniques and arguments related to those of Hjort et al. (2009), but with necessary extensions and modifications.

For the maximiser of $G_n(\cdot, s)$, write $\widehat{\lambda}_n(s) = \|\widehat{\lambda}_n(s)\|u(s)$ for a vector $u(s)$ of unit length. With arguments as in Owen (2001, p. 220),

$$\|\widehat{\lambda}_n(s)\|\{u(s)^{\mathrm{t}}W_n(s)u(s) - E_n(s)u(s)^{\mathrm{t}}V_n(s)\} \le u(s)^{\mathrm{t}}V_n(s),$$

with $E_n(s) = n^{-1/2}\max_{i\le n}\|m_{i,n}(s)\|$, which tends to zero in probability uniformly in s by assumption (iii). Also from assumption (i), $\sup_{s\in S}|u(s)^{\mathrm{t}}V_n(s)| = O_{\mathrm{pr}}(1)$. Moreover, $u(s)^{\mathrm{t}}W_n(s)u(s) \ge e_{n,\min}(s)$, the smallest eigenvalue of $W_n(s)$, which converges in probability to the smallest eigenvalue of W, and this is bounded away from zero by assumption (ii). It follows that $\|\widehat{\lambda}_n(s)\| = O_{\mathrm{pr}}(1)$ uniformly in s. Also, $\lambda_n^*(s) = W_n(s)^{-1}V_n(s)$ is bounded in probability uniformly in s. Via $\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3h(x)$, where $|h(x)| \le 2$ for $|x| \le \frac{1}{2}$, write

$$G_n(\lambda, s) = 2\lambda^{\mathrm{t}}V_n(s) - \tfrac{1}{2}\lambda^{\mathrm{t}}W_n(s)\lambda + r_n(\lambda, s) = G_n^*(\lambda, s) + r_n(\lambda, s).$$

For arbitrary $c > 0$, consider any λ with $\|\lambda\| \le c$. Then we find

$$|r_n(\lambda, s)| \le \frac{2}{3}\sum_{i=1}^n |\lambda^{\mathrm{t}}m_{i,n}(s)/\sqrt{n}|^3\,|h(\lambda^{\mathrm{t}}m_{i,n}(s)/\sqrt{n})| \le \frac{4}{3}E_n(s)\|\lambda\|\lambda^{\mathrm{t}}W_n(s)\lambda \le \frac{4}{3}E_n(s)c^3e_{n,\max}(s),$$

in terms of the largest eigenvalue of $W_n(s)$, as long as $cE_n(s) \le \frac{1}{2}$. Choose c big enough to have both $\widehat{\lambda}_n(s)$

and $\lambda_n^*(s)$ inside this ball for all s with probability exceeding $1-\varepsilon'$, for a preassigned small ε'. Then,

$$\begin{aligned}
&P\Big(\sup_{s\in S}|\max_\lambda G_n(\lambda,s)-\max_\lambda G_n^*(\lambda,s)|\ge\varepsilon\Big)\\
&\le P\Big(\sup_{s\in S}\sup_{\|\lambda\|\le c}|G_n(\lambda,s)-G_n^*(\lambda,s)|\ge\varepsilon\Big)\\
&\le P\Big((4/3)c^3\sup_{s\in S}(E_n(s)e_{n,\max}(s))\ge\varepsilon\Big)+P\Big(\sup_{s\in S}\|\widehat\lambda_n(s)\|>c\Big)+P\Big(\sup_{s\in S}\|\lambda_n^*(s)\|>c\Big)+P\Big(c\sup_{s\in S}E_n(s)>\tfrac12\Big).
\end{aligned}$$

The lim-sup of the probability sequence on the left hand side is hence bounded by $4\varepsilon'$. We have proven that $\sup_{s\in S}|\max_\lambda G_n(\lambda,s)-\max_\lambda G_n^*(\lambda,s)|\to_{\rm pr}0$. □

S.2 Proof of Theorem 1

We work with the two components of (5) separately. First,

$$\ell_n(\theta_0+s/\sqrt n)-\ell_n(\theta_0)\to_d s^{\rm t}U_0-\tfrac12 s^{\rm t}Js \tag{21}$$

by a third order Taylor expansion of the function $\log f(y,\theta)$. Secondly, we shall see that Lemma 1 may be applied, implying

$$\log R_n(\mu(\theta_0+s/\sqrt n))=-\tfrac12 V_n(s)^{\rm t}W_n(s)^{-1}V_n(s)+o_{\rm pr}(1), \tag{22}$$

uniformly in $s\in S$. For this to be valid it is in view of Lemma 1 sufficient to check condition (i) of that lemma (we assumed conditions (ii) and (iii)). Here (i) follows using (4), since

$$\sup_s\|V_n(s)\|=\sup_s\|V_n(0)+\xi_n s\|+o_{\rm pr}(1)\to_d\sup_s\|V_0+\xi_0 s\|.$$

Hence, $\sup_s\|V_n(s)\|=O_{\rm pr}(1)$.

From these efforts we find

$$\begin{aligned}
\log R_n(\mu(\theta_0+s_n/\sqrt n))-\log R_n(\mu(\theta_0))&\to_d-\tfrac12(V_0+\xi_0 s)^{\rm t}W^{-1}(V_0+\xi_0 s)+\tfrac12 V_0^{\rm t}W^{-1}V_0\\
&=-V_0^{\rm t}W^{-1}\xi_0 s-\tfrac12 s^{\rm t}\xi_0^{\rm t}W^{-1}\xi_0 s.
\end{aligned}$$

This convergence also takes place jointly with (21), in view of (6), and we arrive at the conclusion of the theorem. □

S.3 Proof of Corollary 1

Corollary 1 is valid under the following conditions :

(A1) For all $\epsilon > 0$, $\sup_{\|\theta-\theta_0\|>\epsilon} \Gamma(\theta) < \Gamma(\theta_0)$.

(A2) The class $\left\{y \to \frac{\partial}{\partial\theta} \log f(y,\theta) \colon \theta \in \Theta\right\}$ is Donsker (see e.g. van der Vaart and Wellner (1996, Ch. 2)).

(A3) Conditions (C0)–(C2) and (C4)–(C6) in Molanes López et al. (2009) are valid, with their function $g(X, \mu_0, \nu)$ replaced by our function $m(Y, \mu)$, except that instead of demanding boundedness of our function $m(Y, \mu)$ we assume merely that the class

$$y \to \frac{m(y,\mu)m(y,\mu)^{\mathrm{t}}}{\{1+\xi^t m(y,\mu)\}^2},$$

with μ and ξ in a neighbourhood of $\mu(\theta_0)$ and 0, is Donsker (this is a much milder condition than boundedness).

First note that $\Gamma_n(\theta)$ can be written as

$$\Gamma_n(\theta) = (1-a)\, n^{-1} \sum_{i=1}^{n} \{\log f(Y_i,\theta) - \log f(Y_i,\theta_0)\} - a\, n^{-1} \sum_{i=1}^{n} \log\left(1 + \widehat{\xi}(\theta)^{\mathrm{t}} m(Y_i, \mu(\theta))\right),$$

where $\widehat{\xi}(\theta)$ is the solution of

$$n^{-1} \sum_{i=1}^{n} \frac{m(Y_i,\mu(\theta))}{1+\xi^{\mathrm{t}} m(Y_i,\mu(\theta))} = 0.$$

Note that this corresponds with the formula of $\log R_n$ given below Lemma 1 but with $\lambda(\theta)/\sqrt{n}$ relabelled as $\xi(\theta)$. To prove the consistency part, we make use of Theorem 5.7 in van der Vaart (1998). It suffices by condition (A1) to show that $\sup_\theta |\Gamma_n(\theta) - \Gamma(\theta)| \to_{\mathrm{pr}} 0$, which we show separately for the ML and the EL part. For the parametric part we know that $n^{-1}\ell_n(\theta) - \mathrm{E} \log f(Y,\theta)$ is $o_{\mathrm{pr}}(1)$ uniformly in θ by condition (A2). For the EL part, the proof is similar to the proof of Lemma 4 in Molanes López et al. (2009) (except that no rate is required here and that the convergence is uniformly in θ), and hence details are omitted.

Next, to prove (8) we make use of Theorems 1 and 2 in Sherman (1993) about the asymptotics for the maximiser of a (not necessarily concave) optimisation function, and the results in Molanes López et al. (2009), who use the Sherman (1993) paper to establish asymptotic normality and a version of Wilks theorem in an EL context with nuisance parameters. For the verification of the conditions of Theorem 1 (which shows root-n consistency of $\widehat{\theta}_{\mathrm{hl}}$) and Theorem 2 (which shows asymptotic normality of $\widehat{\theta}_{\mathrm{hl}}$) in Sherman (1993), we consider separately the ML part and the EL part. For the EL part all the work is already done using our

Theorem 1 and Lemmas 1–6 in Molanes López et al. (2009), that are valid under condition (A3). Next, the conditions of Theorems 1 and 2 in Sherman (1993) for the ML part follow using standard arguments from parametric likelihood theory and condition (A2). It now follows that $\widehat{\theta}_{\rm hl}$ is asymptotically normal, and its asymptotic variance is equal to $(J^*)^{-1}K^*(J^*)^{-1}$ using Theorem 1.

Finally, (9) follows from a combination of Theorem 1 with $s=\sqrt{n}(\widehat{\theta}_{\rm hl}-\theta_0)$ and the asymptotic normality of $\sqrt{n}(\widehat{\theta}_{\rm hl}-\theta_0)$ to $(J^*)^{-1}U^*$. Indeed,

$$2\{h_n(\widehat{\theta}_{\rm hl})-h_n(\theta_0)\}\to_d 2\{(U^*)^{\rm t}(J^*)^{-1}U^*-\tfrac{1}{2}(U^*)^{\rm t}(J^*)^{-1}J^*(J^*)^{-1}U^*\}=(U^*)^{\rm t}(J^*)^{-1}U^*,$$

and this finishes the proof of the corollary. □

S.4 Proof of Theorem 2

To prove Theorem 2, we revisit several previous arguments for the $A_n(\cdot)\to_d A(\cdot)$ part of Theorem 1, but now needing to extend these to the case of the model departure parameter δ being present. First, we have

$$\ell_n(\theta_0+s/\sqrt{n})-\ell_n(\theta_0)=U_ns-\tfrac{1}{2}s^{\rm t}J_ns+o_{\rm pr}(1)\to_d (U+J_{01}\delta)^{\rm t}s-\tfrac{1}{2}s^{\rm t}J_{00}s.$$

This is essentially since $U_n=n^{-1/2}\sum_{i=1}^n u(Y_i,\theta_0)$ now is seen to have mean $J_{01}\delta$, but the same variance, up to the required order. We need a parallel result for $V_{n,0}=n^{-1/2}\sum_{i=1}^n m(Y_i,\mu(\theta_0))$ under $f_{\rm true}$. Here

$$\begin{aligned}\mathrm{E}_{\rm true}\, m(Y,\mu(\theta_0))&=\int m(y,\mu(\theta_0))f(y,\theta_0)\{1+S(y)^{\rm t}\delta/\sqrt{n}+o(1/\sqrt{n})\}\,{\rm d}y\\&=0+K_{01}\delta/\sqrt{n}+o(1/\sqrt{n}),\end{aligned}$$

yielding $V_{n,0}\to_d V_0+K_{01}\delta$. Along with some further details, this leads to the required extension of the $A_n\to_d A$ part of Theorem 1 and its proof, to the present local neighbourhood model state of affairs;

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

with J^* as defined earlier and with

$$U^*_{\rm plus}=(1-a)(U+J_{01}\delta)-a\xi_0^{\rm t}W^{-1}(V_0+K_{01}\delta)=U^*+L_{01}\delta.$$

Following and then modifying the technical details of the proof of Corollary 1, we arrive at

$$\sqrt{n}(\widehat{\theta}_{\rm hl}-\theta_0)\to_d (J^*)^{-1}(U^*+L_{01}\delta)\sim {\rm N}_p((J^*)^{-1}L_{01}\delta,(J^*)^{-1}K^*(J^*)^{-1}),$$

as required. □

S.5 The HL for regression models

Our HL machinery can be lifted from the iid framework to regression. The following example illustrates the general idea. Consider the normal linear regression model for data (x_i,y_i), with covariate vector x_i of dimension say d, and with y_i having mean $x_i^{\rm t}\beta$. The ML solution is associated with the estimation equation ${\rm E}\, m(Y,X,\beta)=0$, where $m(y,x,\beta)=(y-x^{\rm t}\beta)x$. The underlying regression parameter can be expressed as $\beta=({\rm E}\, XX^{\rm t})^{-1}\,{\rm E}\, XY$, involving also the covariate distribution. Consider now a subvector x_0, of dimension say $d_0<d$, and the associated estimating equation $m_0(y,x,\gamma)=(y-x_0^{\rm t}\gamma)x_0$. This invites the HL construction $(1-a)\ell_n(\beta)+a\log R_n(\gamma(\beta))$. Here $\ell_n(\beta)$ is the ordinary parametric log-likelihood; $R_n(\gamma)$ is the EL associated with m_0; and $\gamma(\beta)$ is $({\rm E}\, X_0X_0^{\rm t})^{-1}\,{\rm E}\, X_0Y$ seen through the lens of the smaller regression, where ${\rm E}\, X_0Y=X_0X^{\rm t}\beta$. This leads to inference about β where it is taken into account that regression with respect to the x_0 components is of particular importance.

S.6 Confidence curve for a focus parameter

For a focus parameter $\psi=\psi(\theta)$, consider the profiled log-hybrid-likelihood function $h_{n,{\rm prof}}(\psi)=\max\{h_n(\theta)\colon \psi(\theta)=\psi\}$. Note that $h_{n,\max}=h_n(\widehat{\theta}_{\rm hl})$ is also the same as $h_{n,{\rm prof}}(\widehat{\psi}_{\rm hl})$. We shall find use for the hybrid deviance function associated with ψ,

$$\Delta_n(\psi)=2\{h_{n,{\rm prof}}(\widehat{\psi}_{\rm hl})-h_{n,{\rm prof}}(\psi)\}.$$

Essentially relying on Theorem 1, which involves matrices J^* and K^* and the limit variable $U^*\sim {\rm N}_p(0,K^*)$, we show below that

$$\Delta_n(\psi_0)\to_d \Delta=\frac{\{c^{\rm t}(J^*)^{-1}U^*\}^2}{c^{\rm t}(J^*)^{-1}c}\sim k\chi^2_1, \tag{23}$$

where $k=c^{\rm t}(J^*)^{-1}K^*(J^*)^{-1}c/c^{\rm t}(J^*)^{-1}c$. Here $c=\partial\psi(\theta_0)/\partial\theta$, as in (10). Estimating this k via plug-in then leads to the full confidence curve ${\rm cc}(\psi)=\Gamma_1(\Delta_n(\psi)/\widehat{k})$, see Schweder and Hjort (2016, Chs. 2–3), often improving on the usual symmetric normal-approximation based confidence intervals. Here $\Gamma_1(\cdot)$ is the

distribution function of the χ^2_1.

To show (23), we go via a profiled version of $A_n(s)$ in (5), namely $B_n(t) = h_{n,\text{prof}}(\psi_0+t/\sqrt{n})-h_{n,\text{prof}}(\psi_0)$, where $\psi_0 = \psi(\theta_0)$. For $B_n(t)$ and $\Delta_n(\psi)$ we have the following.

Theorem 3. *Assume the conditions of Theorem 1 are in force. With $\psi_0 = \psi(\theta_0)$ the true parameter value, and $c = \partial\psi(\theta_0)/\partial\theta$, we have $B_n(t) \to_d B(t) = \{c^{\rm t}(J^*)^{-1}U^*t - \frac{1}{2}t^2\}/c^{\rm t}(J^*)^{-1}c$. Also,*

$$\Delta_n(\psi_0) = 2\max B_n \to_d \Delta = 2\max B = \frac{\{c^{\rm t}(J^*)^{-1}U^*\}^2}{c^{\rm t}(J^*)^{-1}c}.$$

It is clear that $\Delta \sim k\chi^2_1$, with the k given above. Proving the theorem is achieved via Theorem 1, along the lines of a similar type of result for log-likelihood profiling given in Schweder and Hjort (2016, Section 2.4), and we leave out the details.

Remark 1. The special case of $a = 0$ for the HL construction corresponds to parametric ML estimation, and results reached above specialise to the classical results $\sqrt{n}(\widehat{\theta}_{\rm ml}-\theta_0) \to_d {\rm N}_p(0, J^{-1})$, $2\{\ell_{n,\max}-\ell_n(\theta_0)\} \to_d \chi^2_p$, and $\sqrt{n}(\widehat{\psi}_{\rm ml} - \psi_0) \to_d {\rm N}(0, c^{\rm t}J^{-1}c)$. Theorem 3 is then the Wilks theorem for the profiled log-likelihood function. The other extreme case is that of $a \to 1$, with the EL applied to $\mu = \mu(\theta)$. Here Theorem 1 yields $U^* = -\xi_0^{\rm t}W^{-1}V_0$, and with both J^* and K^* equal to $\xi_0^{\rm t}W^{-1}\xi_0$. This case corresponds to a version of the classic EL chi-squared result, now filtered through the parametric model, and with $-2\log R_n(\mu(\theta_0)) \to_d (U^*)^{\rm t}(J^*)^{-1}U^* \sim \chi^2_p$. Also, $\sqrt{n}(\widehat{\psi}_{\rm el} - \psi_0) \to_d {\rm N}(0, \kappa^2)$, with $\kappa^2 = c^{\rm t}\xi_0^{\rm t}W^{-1}\xi_0 c$; here $\widehat{\psi}_{\rm el} = \psi(\widehat{\theta}_{\rm el})$ in terms of the EL estimator, the maximiser of $R_n(\mu(\theta))$.

S.7 An implied goodness-of-fit test for the parametric model

Methods developed in Section 5, in particular those associated with estimating the mean squared error of the final estimator, lend themselves nicely to a goodness-of-fit test for the parametric working model, as follows. We accept the parametric model if the fic(a) criterion of Subsection 5.3 tells us that $\widehat{a} = 0$ is the best balance, and if $\widehat{a} > 0$ the model is rejected. This model test can be accurately examined, by working out an expression for the derivative of fic(a) at $a = 0$, say $\widehat{Z}^0_n$; we reject the model if $\widehat{Z}^0_n > 0$ (since then and only then is $\widehat{a}$ positive).

Here $\widehat{Z}^0_n$ is the estimated version of the limit experiment variable Z^0, which we shall identify below, as a function of $D \sim {\rm N}_q(\delta, Q)$, cf. (20). Let us write $\omega_{\rm hl}(a) = \omega + a\nu + O(a^2)$. Since $\tau_{0,\rm hl}(a)^2 = \tau_0^2 + O(a^2)$, the derivative of

$$\text{fic}(a) = (\omega + a\nu)^{\rm t}(DD^{\rm t} - Q)(\omega + a\nu) + \tau_0^2 + O(a^2)$$

with respect to a, at zero, is seen to be $Z^0 = 2\omega^{\mathrm{t}}(DD^{\mathrm{t}} - Q)\nu$. Hence the limit experiment version of the test is to reject the parametric model if $(\omega^{\mathrm{t}}D)(\nu^{\mathrm{t}}D) > \omega^{\mathrm{t}}Q\nu$, or

$$Z = \frac{\omega^{\mathrm{t}}D}{(\omega^{\mathrm{t}}Q\omega)^{1/2}}\frac{\nu^{\mathrm{t}}D}{(\nu^{\mathrm{t}}Q\nu)^{1/2}} > \rho = \frac{\omega^{\mathrm{t}}Q\nu}{(\omega^{\mathrm{t}}Q\omega)^{1/2}(\nu^{\mathrm{t}}Q\nu)^{1/2}}.$$

Under the null hypothesis of the model, Z is equal in distribution to X_1X_2, where (X_1, X_2) is a binormal pair, with zero means, unit variances, and correlation ρ. The implied significance level, of the implied goodness of fit test, is hence $\alpha = P\{X_1X_2 > \rho\}$, which can be read off via numerical integration or simulation, for a given ρ.

The ν quantity can be identified with a bit of algebraic work, and then estimated consistently from the data. We note that for the special case of $m(y, \mu) = g(y) - \mu$, and with focus on this mean parameter $\mu = \mathrm{E}\, g(Y)$, then ν becomes proportional to ω. The test above is then equivalent to rejecting the model if $(\widehat{\omega}^{\mathrm{t}}D_n)^2/\widehat{\omega}^{\mathrm{t}}\widehat{Q}\widehat{\omega} > 1$, which under the null model happens with probability converging to $\alpha = P\{\chi^2_1 > 1\} = 0.317$.

S.8 A related hybrid estimation method

In earlier sections we have motivated and developed theory for the hybrid likelihood and the HL estimator. A crucial factor has been the quadratic approximation (22). The latter is essentially valid within a $O(1/\sqrt{n})$ neighbourhood around the true data generating mechanism, and has yielded the results of Sections 3 and 5.

A related though different strategy is however to take this quadratic approximation as the starting point. The suggestion is then to define the alternative hybrid estimator as the maximiser $\widetilde{\theta}$ of

$$N_n(\theta) = (1-a)\ell_n(\theta) - \tfrac{1}{2}aV_n(\theta)^{\mathrm{t}}W_n(\theta)^{-1}V_n(\theta). \tag{24}$$

Under and close to the parametric working model, the HL estimator $\widehat{\theta}$ and the new-HL estimator $\widetilde{\theta}$ are first-order equivalent, in the sense of $\sqrt{n}(\widehat{\theta}-\widetilde{\theta}) \to_{\mathrm{pr}} 0$. Of course we could have put up (24) without knowing or caring about EL or HL in the first place, and with different balance weights. But here we are naturally led to the balance weights $1-a$ for the log-likelihood and $-\frac{1}{2}a$ for the quadratic form, from the HL construction.

The advantage of (24) is partly that it is easier computationally, without a layer of Lagrange maximisation for each θ. More importantly, it manages well also outside the $O(1/\sqrt{n})$ neighbourhoods of the working model. The new-HL estimator tends under weak regularity conditions to the maximiser θ_0 of the limit

function of $N_n(\theta)/n$, which may written

$$N(\theta) = (1-a)\int g \log f_\theta \, \mathrm{d}y - \tfrac{1}{2} a \, v_\theta^{\mathrm{t}} (\Sigma_\theta + v_\theta v_\theta^{\mathrm{t}})^{-1} v_\theta,$$

in terms of $v_\theta = \mathrm{E}_f \, m(Y, \mu(\theta))$ and $\Sigma_\theta = \mathrm{Var}_f m(Y, \mu(\theta))$. Note next that $(A+xx^{\mathrm{t}})^{-1} = A^{-1} - A^{-1}xx^{\mathrm{t}}A^{-1}/(1+x^{\mathrm{t}}A^{-1}x)$, for invertible A and vector x of appropriate dimension. This leads to the identity

$$x^{\mathrm{t}}(A+xx^{\mathrm{t}})^{-1}x = \frac{x^{\mathrm{t}}A^{-1}x}{1+x^{\mathrm{t}}A^{-1}x}.$$

Hence the θ_0 associated with the new-HL method is the miminiser of the statistical distance function

$$d_a(f, f_\theta) = (1-a)\mathrm{KL}(f, f_\theta) + \tfrac{1}{2} a \frac{v_\theta^{\mathrm{t}} \Sigma_\theta^{-1} v_\theta}{1 + v_\theta^{\mathrm{t}} \Sigma_\theta^{-1} v_\theta} \tag{25}$$

from the real f to the modelled f_θ; here $\mathrm{KL}(f, f_\theta) = \int f \log(f/f_\theta) \, \mathrm{d}y$ is the Kullback–Leibler distance. For a close to zero, the new-HL is essentially maximising the log-likelihood function, associated with attempting to minimise the KL divergence. For a coming close to 1 the method amounts to minimising an empirical version of $v_\theta^{\mathrm{t}} \Sigma_\theta^{-1} v_\theta$, which means making $v_\theta = \mathrm{E}_f \, m(Y, \mu(\theta))$ close to zero. This is also what the empirical likelihood is aiming at.