Statistica Sinica **28** (2018), 2389-2407
doi:https://doi.org/10.5705/ss.202017.0291

HYBRID COMBINATIONS OF PARAMETRIC AND EMPIRICAL LIKELIHOODS

Nils Lid Hjort, Ian W. McKeague and Ingrid Van Keilegom

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, robust methods, semiparametric estimation.

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 of 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 unstressed working atmosphere. Among the pluses for my Down Under adventures were 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.

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

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).

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.

1. 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 the usual likelihood function for the parametric model, say $L_n(\theta)$, with n referring to sample size, and 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 (1.2) below, is 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 is estimated using $\widehat{\psi}_{\rm hl} = \psi(\widehat{\theta}_{\rm hl})$.

If the working parametric model is correct, these hybrid estimators lose a certain amount in terms of efficiency, when compared to the usual maximum likelihood estimator. We show, however, that the efficiency loss under ideal model conditions is typically a small one, and that the hybrid estimator often outperforms the maximum likelihood when the working model is not correct. Thus the hybrid likelihood is 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 $Y_1, \ldots, Y_n$ 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

$$R_n(\mu) = \max\left\{\prod_{i=1}^n (nw_i)\colon\ \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i m(Y_i, \mu) = 0, \text{each } w_i > 0\right\}. \quad (1.1)$$

This is the empirical likelihood function for μ, see Owen (2001), with further discussions in e.g. Hjort, McKeague and Van Keilegom (2009) and Schweder and Hjort (2016, Chap. 11). One might e.g. choose $m(Y, \mu) = g(Y) - \mu$ for suitable $g = (g_1, \ldots, g_q)$, 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*

$$H_n(\theta) = L_n(\theta)^{1-a} R_n(\mu(\theta))^a, \quad (1.2)$$

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. The general approach (1.2) 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 (1.2) provides a bridge from the purely parametric to the nonparametric empirical likelihood. The a parameter dictates the degree of balance. One can view (1.2) 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.

Example 1. Let f_θ be the normal density with parameters (ξ, σ^2), and take $m_j(y, \mu_j) = I\{y \le \mu_j\} - j/4$ for $j = 1, 2, 3$. Then (1.1), with the ensuing $\mu_j(\xi, \sigma) = \xi + \sigma\,\Phi^{-1}(j/4)$ for $j = 1, 2, 3$, can 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 can 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. 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 the 'control equation' above, with $p = P\{Y \in A\} = \int_A f(y, \theta)\, dy$. The effect is 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 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 that the hybrid likelihood estimation strategy in this case is large-sample equivalent to maximising

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

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 $p_j(\widehat{\theta})$ is the model-based estimate.

In Section 2 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 3 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 4 examines aspects related to fine-tuning the balance parameter a of (1.2), and we provide a recipe for its selection. An illustration of our HL methodology is given in Section 5, 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, Guo and Tang (2017). We like to mention in particular the paper by DiCiccio, Hall and Romano (1989), in which the features and behaviour of parametric and empirical likelihood functions are compared. We mention that Peter also made very influential contributions to the somewhat related area of likelihood tilting, see e.g. Choi, Hall and Presnell (2000).

2. 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 the Wilks theorem. The HL estimator $\widehat{\theta}_{\rm hl}$ maximises

$$h_n(\theta) = \log H_n(\theta) = (1-a)\ell_n(\theta) + a \log R_n(\mu(\theta)) \tag{2.1}$$

over θ (assumed here to be unique), 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 arbitrary compact $S \subset \mathbb{R}^p$. For simplicity 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 $V_n(s) = n^{-1/2}\sum_{i=1}^n m_{i,n}(s)$ and $W_n(s) = n^{-1}\sum_{i=1}^n m_{i,n}(s) m_{i,n}(s)^{\rm t}$. Hence 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, Chap. 11); for a detailed proof, along with further discussion, see Hjort, McKeague and Van Keilegom (2009, Remark 2.7). The following 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. *For a compact $S \subset \mathbb{R}^p$, 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 $W = \mathrm{Var}\, m(Y, \mu(\theta_0))$ 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 $\widehat{\lambda}_n(s) = \mathrm{argmax}_\lambda\, G_n(\lambda, s)$ and $\lambda_n^*(s) = \mathrm{argmax}_\lambda G_n^*(\lambda, s) = W_n^{-1}(s)V_n(s)$ 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, 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, 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 (1.1), and also on how $\mu = \mu(f_\theta) = \mu(\theta)$ behaves close to θ_0. In addition to $\mathrm{E}\, m(Y, \mu(\theta_0)) = 0$, we assume that

$$\sup_{s\in S} \|V_n(s) - V_n(0) - \xi_n s\| = o_{\rm pr}(1), \tag{2.2}$$

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 (2.2), 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:

$$\begin{aligned} A_n(s) &= h_n\left(\theta_0 + \frac{s}{\sqrt{n}}\right) - h_n(\theta_0) \\ &= (1-a)\left\{\ell_n\left(\theta_0 + \frac{s}{\sqrt{n}}\right) - \ell_n(\theta_0)\right\} \\ &\quad + a\left\{\log R_n\left(\mu\left(\theta_0 + \frac{s}{\sqrt{n}}\right)\right) - \log R_n(\mu(\theta_0))\right\}. \end{aligned} \tag{2.3}$$

Below we show that $A_n(s)$ converges weakly to a quadratic limit $A(s)$, uniformly in s over compacta, which then leads to our most important insights concerning HL-based estimation and inference. By the multivariate Central Limit Theorem,

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

$$\text{where } \Sigma = \begin{pmatrix} J & C \\ C^{\rm t} & W \end{pmatrix}. \tag{2.4}$$

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 $p \times p$, $C = \mathrm{E}\, u(Y,\theta_0) m(Y,\mu(\theta_0))^{\mathrm{t}}$ 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 smootheness conditions on* $\log f(y,\theta)$ *hold, as spelled out in Section S.2; the conditions of Lemma* 1 *are in force, along with condition* (2.2) *with the appropriate* ξ_0*, for each compact* $S \subset \mathbb{R}^p$*; and that* Σ *has full rank. Then, for each compact* S*,* $A_n(s)$ *converges weakly to* $A(s) = s^{\mathrm{t}} U^* - (1/2) s^{\mathrm{t}} J^* s$*, in the function space* $\ell^\infty(S)$ *endowed with the uniform topology, where* $U^* = (1-a)U_0 - a\xi_0^{\mathrm{t}} W^{-1} V_0$ *and* $J^* = (1-a)J + a\xi_0^{\mathrm{t}} W^{-1}\xi_0$*. Here* $U^* \sim \mathrm{N}_p(0, K^*)$*, with variance matrix* $K^* = (1-a)^2 J + a^2 \xi_0^{\mathrm{t}} W^{-1}\xi_0 - a(1-a)(CW^{-1}\xi_0 + \xi_0^{\mathrm{t}} W^{-1} C^{\mathrm{t}})$.

The theorem, proved in Section S.2 of the supplementary material, is valid for each fixed balance parameter a in (1.2), with J^* and K^* also depending on a. We discuss ways of fine-tuning a in Section 4.

The $p \times q$-dimensional block component C of the variance matrix Σ of (2.4) 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 $\partial\mu/\partial\theta$, 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 $j = 1,\ldots,q$. From $\int m(y,\mu(\theta)) f(y,\theta)\,\mathrm{d}y = 0$ for all θ follows the $q \times p$-dimensional equation $\int m^*(y,\mu(\theta_0)) f(y,\theta_0)\,\mathrm{d}y\,(\partial\mu/\partial\theta)^{\mathrm{t}} + \int m(y,\mu(\theta_0)) f(y,\theta_0) u(y,\theta_0)^{\mathrm{t}}\,\mathrm{d}y = 0$, where $m^*(y,\mu) = \partial m(y,\mu)/\partial\mu$, in case m is differentiable with respect to μ. This means $C = -(\partial\mu/\partial\theta)\,\mathrm{E}_\theta\, m^*(Y,\mu(\theta_0))$. If $m(y,\mu) = g(y) - \mu$, for example, corresponding to parameters $\mu = \mathrm{E}\, g(Y)$, we have $C = \partial\mu/\partial\theta$. Also, using (2.2) we have $\xi_0 = -(\partial\mu/\partial\theta)^{\mathrm{t}}$, of dimension $q \times p$. Applying Theorem 1 yields $U^* = (1-a)U_0 + a(\partial\mu/\partial\theta)W^{-1}V_0$, along with

$$J^* = (1-a)J + a\frac{\partial\mu}{\partial\theta} W^{-1} \left(\frac{\partial\mu}{\partial\theta}\right)^{\mathrm{t}}$$

$$\text{and } K^* = (1-a)^2 J + \{1-(1-a)^2\}\frac{\partial\mu}{\partial\theta} W^{-1}\left(\frac{\partial\mu}{\partial\theta}\right)^{\mathrm{t}}. \qquad (2.5)$$

For the following 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

$$\Gamma(\theta) = -(1-a)\,\mathrm{KL}(f_{\theta_0}, f_\theta) - a\,\mathrm{E}\,\log\big(1 + \xi(\theta)^{\mathrm{t}} m(Y,\mu(\theta))\big), \qquad (2.6)$$

with $\widehat{\theta}_{\mathrm{hl}}$ as the argmax of $\Gamma_n(\cdot)$. Here $\mathrm{KL}(f, f_\theta) = \int f \log(f/f_\theta)\,\mathrm{d}y$ is the

Kullback–Leibler divergence, in this case from f_{θ_0} to f_θ, and with $\xi(\theta)$ the solution of $\mathrm{E}\, m(Y,\mu(\theta))/\{1+\xi^{\mathrm{t}} m(Y,\mu(\theta)\} = 0$ (that this solution exists and is unique is a consequence of the proof of Corollary 1 below).

Corollary 1. *Under the conditions of Theorem* 1 *and under conditions* (*A*1)–(*A*3) *given in Section S.*3 *of the supplementary material,* (*i*) *there is consistency of* $\widehat{\theta}_{\mathrm{hl}}$ *towards* θ_0*;* (*ii*) $\sqrt{n}(\widehat{\theta}_{\mathrm{hl}}-\theta_0) \to_d (J^*)^{-1}U^* \sim \mathrm{N}_p\big(0,(J^*)^{-1}K^*(J^*)^{-1}\big)$*; and* (*iii*) $2\{h_n(\widehat{\theta}_{\mathrm{hl}})-h_n(\theta_0)\} \to_d (U^*)^{\mathrm{t}}(J^*)^{-1}U^*$.

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 them, 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 $\widehat{\psi}_{\mathrm{hl}} = \psi(\widehat{\theta}_{\mathrm{hl}})$. With $\psi_0 = \psi(\theta_0)$ as the true parameter value, we then have via the delta method that

$$\sqrt{n}(\widehat{\psi}_{\mathrm{hl}}-\psi_0) \to_d c^{\mathrm{t}}(J^*)^{-1}U^* \sim \mathrm{N}(0,\kappa^2), \text{ where } \kappa^2 = c^{\mathrm{t}}(J^*)^{-1}K^*(J^*)^{-1}c. \quad (2.7)$$

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 $\sqrt{n}(\widehat{\mu}_{j,\mathrm{hl}}-\mu_{0,j})$ has a normal limit with variance $(\partial\mu_j/\partial\theta)^{\mathrm{t}}(J^*)^{-1}K^*(J^*)^{-1}(\partial\mu_j/\partial\theta)$, in terms of $(\partial\mu_j/\partial\theta) = \partial\mu_j(\theta_0)/\partial\theta$. Armed with Corollary 1, 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 $A_1 = \xi_0^{\mathrm{t}}W^{-1}\xi_0 - J$ and $A_2 = -2J - CW^{-1}\xi_0 - \xi_0^{\mathrm{t}}W^{-1}C^{\mathrm{t}}$. For the class of functions of the form $m(y,\mu) = g(y) - T(\mu)$, corresponding to $\mu = T^{-1}(\mathrm{E}\, g(Y))$, 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. Writing $(J^*)^{-1}K^*(J^*)^{-1}$ as $(J^{-1} - aJ^{-1}A_1J^{-1})(J+aA_2)(J^{-1} - aJ^{-1}A_1J^{-1}) + O(a^2)$, therefore, one finds that this is $J^{-1} + O(a^2)$, which in particular means that the efficiency loss is very small when a is small.

3. Hybrid Likelihood Outside Model Conditions

In Section 2 we investigated the hybrid likelihood estimation strategy under the conditions of the parametric model. Under suitable conditions, the HL is consistent and asymptotically normal, with a certain mild loss of efficiency under model conditions, compared to the parametric ML method, 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 turns 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 (1.2), and we shall deal with this problem too, in Section 4.

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 now examine behaviour of the ML and the HL schemes when γ is in the neighbourhood of γ_0. Suppose in fact that

$$f_{\rm true}(y) = f\left(y, \theta_0, \gamma_0 + \frac{\delta}{\sqrt{n}}\right), \tag{3.1}$$

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

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d {\rm N}_p(B\delta, \Omega), \tag{3.2}$$

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

$$\sqrt{n}(\widehat{\psi} - \psi_{\rm true}) \to_d {\rm N}(b^{\rm t}\delta, \tau^2), \tag{3.3}$$

with $b = B^{\rm t}(\partial\psi/\partial\theta) - \partial\psi/\partial\gamma$ and $\tau^2 = (\partial\psi/\partial\theta)^{\rm t}\Omega(\partial\psi/\partial\theta)$, 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 ${\rm mse}(\delta) = (b^{\rm t}\delta)^2 + \tau^2$. Among the consequences of using the narrow working model when it is moderately wrong, at the level of $\gamma = \gamma_0 + \delta/\sqrt{n}$, is the bias $b^{\rm t}\delta$. The size of this bias depends on the focus parameter, and it may be zero for some foci, even when the model is incorrect.

We 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

(3.3). 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

$$J_{\rm wide}=\begin{pmatrix}J_{00} & J_{01}\\ J_{10} & J_{11}\end{pmatrix} \tag{3.4}$$

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 (2.4) and appearing in Theorem 1 etc. Here one can 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 (3.3) followed from (3.2), one finds for $\widehat\psi_{\rm ml}=\psi(\widehat\theta_{\rm ml})$ that

$$\sqrt{n}(\widehat\psi_{\rm ml}-\psi_{\rm true})\to_d {\rm N}(\omega^{\rm t}\delta,\tau_0^2), \tag{3.5}$$

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

In such a situation, with a clear interest parameter ψ, we use the HL to get $\widehat\psi_{\rm hl}=\psi(\widehat\theta_{\rm hl},\gamma_0)$. We work out what happens with $\widehat\theta_{\rm hl}$ in this framework, generalising what is found in the previous section. Introduce $S(y)=\partial\log f(y,\theta_0,\gamma_0)/\partial\gamma$, the score function in direction of these extension parameters, and let $K_{01}=\int f(y,\theta_0)m(y,\mu(\theta_0))S(y)\,{\rm d}y$, of dimension $q\times r$, along with $L_{01}=(1-a)J_{01}-a(\partial\psi/\partial\theta)^{\rm t}W^{-1}K_{01}$, of dimension $p\times r$, and with transpose $L_{10}=L_{01}^{\rm t}$. The following is proved in Section S.4.

Theorem 2. *Assume data stem from the extended $p+r$-dimensional model* (3.1)*, 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* (1.2)*, results* (3.2)–(3.3) *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 (3.3) follows from (3.2):

$$\sqrt{n}(\widehat\psi_{\rm hl}-\psi_{\rm true})\to_d {\rm N}(\omega_{\rm hl}^{\rm t}\delta,\tau_{0,\rm hl}^2), \tag{3.6}$$

with $\omega_{\rm hl}=L_{10}(J^*)^{-1}(\partial\psi/\partial\theta)-\partial\psi/\partial\gamma$ and $\tau_{0,\rm hl}^2=(\partial\psi/\partial\theta)^{\rm t}(J^*)^{-1}K^*(J^*)^{-1}(\partial\psi/\partial\theta)$. The quantities involved in these large-sample properties of the HL estimator depend on the balance parameter a employed in the basic HL construction (1.2). For $a=0$ we are back to the ML, with (3.6) specialising to (3.5). As a moves away from zero, more emphasis is placed on the EL part, in effect pushing θ so $n^{-1}\sum_{i=1}^n m(Y_i,\mu(\theta))$ gets 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.

4. Fine-Tuning the Balance Parameter

The basic HL construction of (1.2) first entails selecting context relevant control parameters μ, 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 here.

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 2 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 2, see e.g. (2.5), and have added the subscript a, for emphasis. If there is special interest in some focus parameter ψ, one may select a so that

$$\kappa_a = \{c^{\rm t}(J_a^*)^{-1}K_a^*(J_a^*)^{-1}c\}^{1/2} \le (1+\varepsilon)\kappa_0 = (1+\varepsilon)(c^{\rm t}J^{-1}c)^{1/2}, \qquad (4.1)$$

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 (2.7) 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 when 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.

Features of the mse(a). The methods above, as with (4.1), rely on the theory developed in Section 2, under the conditions of the parametric working model. In what follows we need the theory given in Section 3, 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

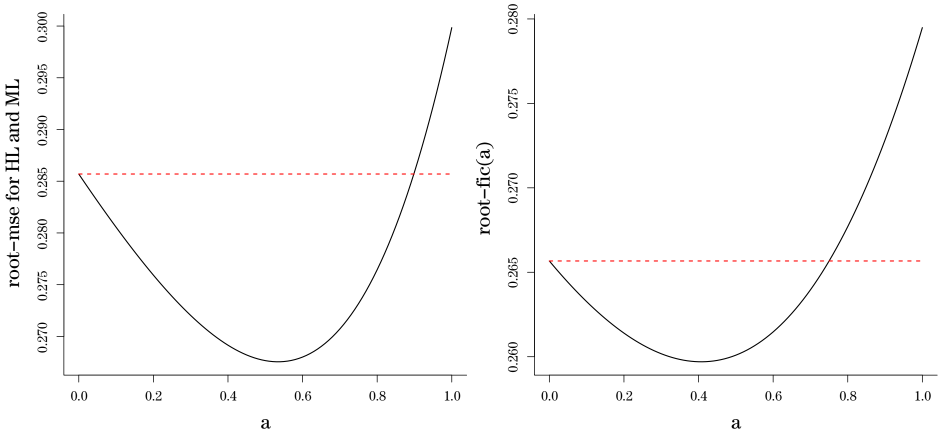

Figure 2. (a) The dotted horizontal line indicates the root-mse for the ML estimator, and the full curve the root-mse for the HL estimator, as a function of the balance parameter a in the HL construction. (b) The root-fic(a), as a function of the balance parameter a, constructed on the basis of $n = 100$ simulated observations, from a case where $\gamma = 1 + \delta/\sqrt{n}$, with δ described in the text.

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 below.

For a given focus parameter ψ, the limit mse when using the HL with parameter a is found from (3.6):

$$\mathrm{mse}(a) = \{\omega_{\mathrm{hl}}(a)^{\mathrm{t}}\delta\}^2 + \tau_{0,\mathrm{hl}}(a)^2. \tag{4.2}$$

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 2(a) displays the root-mse(a) curve in a simple setup, where the parametric start model is Beta($\theta, 1$), with density $\theta y^{\theta-1}$, and the focus parameter used for the HL construction is $\psi = \mathrm{E}\, Y^2$, 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 (3.1). The δ for this illustration is chosen to be $Q^{1/2} = (J^{11})^{1/2}$, from (4.3) 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 (3.6) starts out for $a = 0$ at $\omega = J_{10}J_{00}^{-1}(\partial\psi/\partial\theta) - \partial\psi/\partial\gamma$ in (3.5), 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.

Estimation of mse(a). Concrete evaluation of the mse(a) curves of (4.2) 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 (3.6) 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, Chap. 6–7), in connection with construction of their Focused Information Criterion (FIC), we have

$$D_n \to_d D \sim {\rm N}_r(\delta, Q), \quad \text{with } Q = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}. \qquad (4.3)$$

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 (3.4). This leads to estimating mse(a) using

$$\begin{aligned}{\rm fic}(a) &= \{\widehat\omega_{\rm hl}(a)^{\rm t}(D_nD_n^{\rm t} - \widehat Q)\widehat\omega_{\rm hl}(a)\}_+ + \widehat\tau_{0,{\rm hl}}(a)^2 \\ &= \big[n\widehat\omega_{\rm hl}(a)^{\rm t}\{(\widehat\gamma - \gamma_0)(\widehat\gamma - \gamma_0)^{\rm t} - \widehat Q\}\widehat\omega_{\rm hl}(a)\big]_+ + \widehat\tau_{0,{\rm hl}}(a)^2.\end{aligned}$$

Figure 2(b) displays such a root-fic curve, the estimated root-mse(a). Whereas the root-mse(a) curve shown in Figure 2(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 2(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 2(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 $a = 0.41$.

5. 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, Chap. 2).

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 3(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 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 3(b) displays this $\widehat{p}_a$, as a function of a, 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 4 may be appealed to. The relatively speaking simplest of these is that associated with (4.1), where we numerically compute $\kappa_a = \{c^{\mathrm{t}}(J^*)^{-1}K^*(J^*)^{-1}c\}^{1/2}$ for each a, at the ML position in the parameter space of (b,c), and with J^* and K^* from (2.5). 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

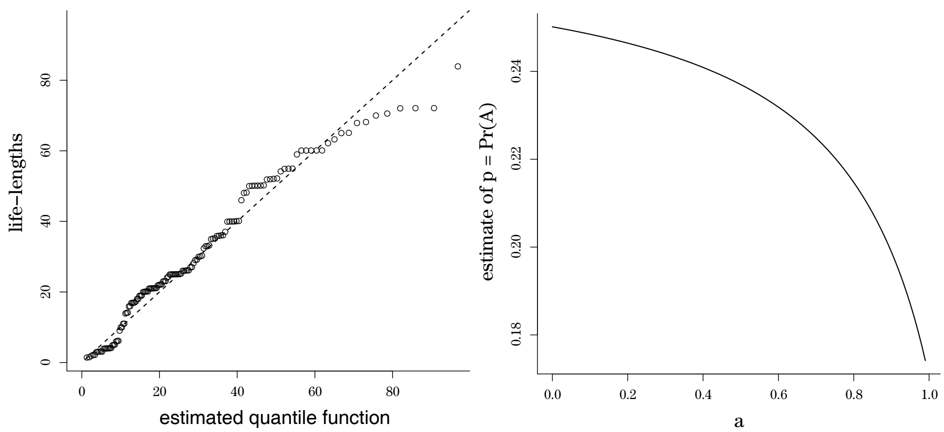

Figure 3. (a) The q-q plot shows the ordered life-lengths $y_{(i)}$ plotted against the ML-estimated gamma quantile function $F^{-1}(i/(n+1), \widehat{b}, \widehat{c})$. (b) The curve $\widehat{p}_a$, with the probability $p = P\{Y \in [9.5, 20.5]\}$ estimated via the HL estimator, is displayed, as a 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.

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 3. The more elaborate recipe of selecting a, developed in Section 4 and using $\mathrm{fic}(a)$, can also be used here.

6. Further Developments and the Supplementary Material

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 extension of the basic HL construction to regression type data, in Section S.5; log-HL-profiling operations and a deviance fuction, leading to a full confidence curve for a focus parameter, in Section S.6, an implicit goodness-of-fit test for the parametric vehicle model, in Section S.7, and a related but different hybrid likelihood construction, in Section S.8.

Supplementary Materials

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.

Acknowledgment

N.L. Hjort is supported via the Norwegian Research Council funded project FocuStat. I.W. McKeague is partially supported by NIH Grant R01GM095722. 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).

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.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, Melbourne.

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. *The Annals of Statistics* **37**, 1079–1111.

Hjort, N. L. and Pollard, D. (1994). Asymptotics for minimisers of convex processes. Technical report, Department of Mathematics, University of Oslo.

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. *The 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.

Department of Mathematics, P.B. 1053 Blindern, University of Oslo, N-0316 Oslo, Norway.

E-mail: nils@math.uio.no

Department of Biostatistics, Columbia University, 722 West 168th Street, MSPH Box 12, New York, NY 10032-2603, USA.

E-mail: im2131@columbia.edu

ORSTAT KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium.

E-mail: ingrid.vankeilegom@kuleuven.be

(Received June 2017; accepted July 2017)