Supplementary material to '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

This supplementary material 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

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

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

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

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,

$$\begin{aligned}
&P\Big(\sup_{s\in S}|\max_\lambda G_n(\lambda,s) - \max_\lambda G_n^*(\lambda,s)| \ge \varepsilon\Big)\\
&\le P\Big(\sup_{s\in S}\sup_{\|\lambda\|\le c}|G_n(\lambda,s) - G_n^*(\lambda,s)| \ge \varepsilon\Big)\\
&\le P\Big((4/3)c^3\sup_{s\in S}(E_n(s)e_{n,\max}(s)) \ge \varepsilon\Big) + P\Big(\sup_{s\in S}\|\widehat{\lambda}_n(s)\| > c\Big)\\
&\qquad + P\Big(\sup_{s\in S}\|\lambda_n^*(s)\| > c\Big) + P\Big(c\sup_{s\in S}E_n(s) > \tfrac{1}{2}\Big).
\end{aligned}$$

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, with $U_n = n^{-1/2}\sum_{i=1}^n u(Y_i,\theta_0)$, which tends to $U_0\sim {\rm N}_p(0,J)$, cf. (6),

$$\ell_n(\theta_0+s/\sqrt{n}) - \ell_n(\theta_0) = s^{\rm t}U_n - \tfrac{1}{2}s^{\rm t}Js + \varepsilon_n(s), \quad \text{with } \sup_{s\in S}|\varepsilon_n(s)| \to_{\rm pr} 0, \tag{19}$$

under various sets of mild regularity conditions. If $\log f(y,\theta)$ is concave in θ, no other conditions are required, beyond finiteness of the Fisher information matrix J, see Hjort and Pollard (1994). Without concavity, but assuming the existence of third order derivatives $D_{i,j,k}(y,\theta) = \partial^3\log f(y,\theta)/\partial\theta_i\partial\theta_j\partial\theta_k$, it is straightforward via Tayor expansion to verify (19) under the condition that $\sup_{\theta\in\mathcal{N}}\max_{i,j,k}|D_{i,j,k}(Y,\theta)|$ has finite mean, with $\mathcal{N}$ a neighbourhood around θ_0. This condition is met for most of the usually employed parametric families. We finally point out that (19) can be established without third order derivatives, with a mild continuity condition on the second derivatives, see e.g. Ferguson (1996, Ch. 18).

Secondly, we shall see that Lemma 1 may be applied, implying

$$\log R_n(\mu(\theta_0 + s/\sqrt{n})) = -\tfrac{1}{2} V_n(s)^{\mathrm{t}} W_n(s)^{-1} V_n(s) + o_{\mathrm{pr}}(1), \tag{20}$$

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

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

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

From these efforts we find

$$\begin{aligned} \log R_n(\mu(\theta_0 + s/\sqrt{n})) - \log R_n(\mu(\theta_0)) \;&\to_d\; -\tfrac{1}{2}(V_0 + \xi_0 s)^{\mathrm{t}} W^{-1}(V_0 + \xi_0 s) + \tfrac{1}{2} V_0^{\mathrm{t}} W^{-1} V_0 \\ &=\; -V_0^{\mathrm{t}} W^{-1} \xi_0 s - \tfrac{1}{2} s^{\mathrm{t}} \xi_0^{\mathrm{t}} W^{-1} \xi_0 s. \end{aligned}$$

This convergence also takes place jointly with (19), 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, where $\Gamma(\cdot)$ is defined in (8):

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

(A2) The class $\{y \mapsto \frac{\partial}{\partial \theta} \log f(y, \theta) \colon \theta \in \Theta\}$ is P-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 $m(Y, \mu(\theta))$, with θ playing the role of ν, except that instead of demanding boundedness of our function $m(Y, \mu)$ we assume merely that the class

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

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

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

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

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 $\xi(\theta)$. That the $\widehat{\xi}(\theta)$ solution is unique follows from considerations along the lines of Molanes López et al. (2009, p. 415). 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 statement (ii) of the corollary, we make use of Theorems 1 and 2 in Sherman (1993) about the asymptotics for the maximiser of a (not necessarily concave) criterion function, and the results in Molanes López et al. (2009), who use the Sherman (1993) paper to establish asymptotic normality and a version of the 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. We note that Theorem 1 in Sherman (1993) requires consistency of the estimator, which we here have established by arguments above. For the EL part all the work is already done using our Theorem 1 and Lemmas 1–6 in Molanes López et al. (2009), which 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 $\widehat{\theta}_{\mathrm{hl}}$ is asymptotically normal, and its asymptotic variance is equal to $(J^*)^{-1}K^*(J^*)^{-1}$ using Theorem 1.

Finally, claim (iii) of the corollary follows from a combination of Theorem 1 with $s = \sqrt{n}(\widehat{\theta}_{\mathrm{hl}} - \theta_0)$ and

the asymptotic normality of $\sqrt{n}(\widehat{\theta}_{\rm hl}-\theta_0)$ to $(J^*)^{-1}U^*$. Indeed,

$$2\{h_n(\widehat{\theta}_{\rm hl})-h_n(\theta_0)\}\to_d 2\{(U^*)^{\rm t}(J^*)^{-1}U^*-\tfrac12(U^*)^{\rm t}(J^*)^{-1}J^*(J^*)^{-1}U^*\}=(U^*)^{\rm t}(J^*)^{-1}U^*,$$

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

$$\ell_n(\theta_0+s/\sqrt{n})-\ell_n(\theta_0)=U_n s-\tfrac12 s^{\rm t}J_n s+o_{\rm pr}(1)\to_d (U+J_{01}\delta)^{\rm t}s-\tfrac12 s^{\rm t}J_{00}s.$$

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;

$$A_n(s)=h_n(\theta_0+s/\sqrt{n})-h_n(\theta_0)\to_d A(s)=s^{\rm t}U^*_{\rm plus}-\tfrac12 s^{\rm t}J^*s,$$

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 $\mathrm{E}\, m(Y, X, \beta) = 0$, where $m(y, x, \beta) = (y - x^{\rm t}\beta)x$. The underlying regression parameter can be expressed as $\beta = (\mathrm{E}\, XX^{\rm t})^{-1}\,\mathrm{E}\, XY$, 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 $(\mathrm{E}\, X_0X_0^{\rm t})^{-1}\,\mathrm{E}\, X_0Y$ seen through the lens of the smaller regression, where $\mathrm{E}\, X_0Y = X_0X^{\rm t}\beta$. 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 $U^* \sim \mathrm{N}_p(0, K^*)$, we show below that

$$\Delta_n(\psi_0) \to_d \Delta = \frac{\{c^{\rm t}(J^*)^{-1}U^*\}^2}{c^{\rm t}(J^*)^{-1}c} \sim k\chi^2_1, \tag{21}$$

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 (9). Estimating this k via plug-in then leads to the full confidence curve $\mathrm{cc}(\psi) = \Gamma_1(\Delta_n(\psi)/\widehat{k})$, 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 (21), we go via a profiled version of $A_n(s)$ in (5), namely $B_n(t) = h_{n,\rm prof}(\psi_0 + t/\sqrt{n}) - h_{n,\rm prof}(\psi_0)$,

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 4, 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 Section 4 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. (18). 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 $Z^0 = 2\omega^{\rm t}(DD^{\rm t} - Q)\nu$. Hence the limit experiment version of the test is to reject the parametric model if $(\omega^{\rm t}D)(\nu^{\rm t}D) > \omega^{\rm t}Q\nu$, or

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

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 $(\widehat{\omega}^{\rm t}D_n)^2/\widehat{\omega}^{\rm t}\widehat{Q}\widehat{\omega} > 1$, which under the null model happens with probability converging to $\alpha = P\{\chi_1^2 > 1\} = 0.317$.

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 (20). 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 2 and 4.

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

$$N_n(\theta) = (1-a)\ell_n(\theta) - \tfrac{1}{2}aV_n(\theta)^{\rm t}W_n(\theta)^{-1}V_n(\theta). \tag{22}$$

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 $\sqrt{n}(\widehat{\theta}-\widetilde{\theta}) \to_{\rm pr} 0$. Of course we could have put up (22) 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 (22) 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

$$d_a(f,f_\theta) = (1-a)\mathrm{KL}(f,f_\theta) + \tfrac{1}{2}a\frac{v_\theta^{\mathrm{t}}\Sigma_\theta^{-1}v_\theta}{1+v_\theta^{\mathrm{t}}\Sigma_\theta^{-1}v_\theta} \tag{23}$$

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.

References

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

Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *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.

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.

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.