

Model robust hybrid likelihood

Ingrid Dæhlen^{1,2} and Nils Lid Hjort¹

¹*Department of Mathematics, University of Oslo*

²*Norwegian Computing Center*

Supplementary Material

The supplementary material contains proofs of Theorem 1 and 2 in the paper (Section S2 and S3 with preliminaries in Section S1), a simulation study (Section S5) and a multivariate example (Section S6).

We will first give formal proofs for the theorems in the article. The arguments utilize techniques for proofs concerning estimators fitted by two-stage procedures and are similar to those given by e.g. Imbens, Johnson, and Spady (1998) and Schennach (2007). Before we begin, however, we will state and prove a result concerning consistency of Z-estimators. The result is a modification of Theorem 5.42 in van der Vaart (2000).

Lemma 1. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and let $\psi: \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^q$ be a function with $\Theta \subseteq \mathbb{R}^p$. Assume there is a unique solution θ_n to the equation $0 = \Psi_n(\theta) = n^{-1} \sum_{i=1}^n \psi(Y_i, \theta)$ for all n and a unique solution θ^* to the limit equation $0 = \Psi(\theta) = E[\psi(Y, \theta)]$ when $Y \sim F$. Then θ_n converges in probability to θ^* , provided*

(C1) the function Ψ is continuously differentiable in a neighborhood of θ^* and $\Psi'(\theta^*)$ is invertible;

(C2) there exists a neighborhood $\mathcal{N}$ of θ^* and an F -integrable function p_0 such that $\|\psi(y, \theta)\| \leq$

$p_0(y)$ for all $\theta \in \mathcal{N}$ and F -almost all y .

Proof. Let Ψ and Ψ_n be defined as in the lemma. Inspection of the proof of Theorem 5.42 in van der Vaart (2000) shows that the conditions of the theorem are slightly too strict and that the only thing needed for the arguments to go through is that Ψ is continuously differentiable with non-singular derivative in a neighborhood of θ^* and that there exists a closed ball around θ^* inside which Ψ_n converges uniformly to Ψ . The first condition is ensured by (C1) and continuity of the determinant function. The second condition follows from (C2) and the uniform law of large numbers. Hence, there exists a sequence of estimators converging in probability to θ^* . By assumption, $\Psi_n(\theta) = 0$ has a unique solution θ_n . Because of this, θ_n must be this consistent sequence, proving that θ_n converges in probability to θ^* . $\square$

S1 The maximum empirical likelihood estimator

We start by proving a limit result for the maximum empirical likelihood estimator (the MEL estimator). This will make the proofs concerning the maximum hybrid likelihood estimator more transparent.

Lemma 2. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^d \times \mathcal{M} \rightarrow \mathbb{R}^q$, where $\mathcal{M} \subseteq \mathbb{R}^s$ is an open set, be a function twice continuously differentiable in the last s arguments. Let $\hat{\mu}$ be the maximizer of the empirical likelihood function and its unique critical point in $\mathcal{M}$. Furthermore, assume $\hat{\lambda}$ is the unique solution to*

$$0 = n^{-1} \sum_{i=1}^n \frac{m(Y_i, \hat{\mu})}{1 + \lambda^T m(Y_i, \hat{\mu})},$$

and $\hat{\alpha} = (\hat{\lambda}, \hat{\mu})$. Set

$$\psi_{\text{EL}}(y, \alpha) = [1 + \lambda^T m(y, \mu)]^{-1} (m(y, \mu)^T, \lambda^T \partial m(y, \mu))^T$$

with ∂m as short hand for $\partial m / \partial \mu$. Lastly assume $\Gamma^{\text{EL}}(\alpha) = \mathbb{E} \psi(Y, \alpha)$ has a unique root $\alpha^* = (\lambda^*, \mu^*)$. Then $\hat{\alpha}$ is consistent for α^* and

$$\hat{\alpha} = \alpha^* - n^{-1} J \Gamma^{\text{EL}}(\alpha^*)^{-1} \sum_{i=1}^n \psi_{\text{EL}}(Y_i, \alpha^*) + o_{\text{pr}}(1/\sqrt{n}) \quad (\text{S1.1})$$

if there exists a neighbourhood $\mathcal{N}$ on which

(1) Γ^{EL} is continuously differentiable with non-singular Jacobian matrix at α^* ;

(2) $\Pr(1 + \lambda^T m(Y, \mu) > L) = 1$ for some $L > 0$ and all $(\lambda, \mu) \in \mathcal{N}$;

(3) the norm of m and all first and second order partial derivatives of m are bounded by some functions p_j for $j = 0, 1, 2$ respectively, with $\mathbb{E}[p_0(Y)^4]$, $\mathbb{E}[p_1(Y)^4] < \infty$ and $\mathbb{E}[p_2(Y)^2] < \infty$.

In particular, $\sqrt{n}(\hat{\alpha} - \alpha^*) \xrightarrow{d} N[0, J \Gamma^{\text{EL}}(\alpha^*)^{-1} K_{\text{EL}} J \Gamma^{\text{EL}}(\alpha^*)^{-1}]$, with $K_{\text{EL}} = \text{Var}[\psi_{\text{EL}}(Y, \alpha^*)]$, $Y \sim F$.

Proof. Since m is differentiable and $\mathcal{M}$ is open, the MEL estimator satisfies the equation

$$0 = \frac{\partial}{\partial \mu} \log \text{EL}_n(\mu) = - \sum_{i=1}^n \frac{\lambda'_n(\mu)^T m(Y_i, \mu) + \partial m(Y_i, \mu)^T \lambda_n(\mu)}{1 + \lambda_n(\mu)^T m(Y_i, \mu)} = - \sum_{i=1}^n \frac{\partial m(Y_i, \mu)^T \lambda_n(\mu)}{1 + \lambda_n(\mu)^T m(Y_i, \mu)},$$

where the last equality follows from definition of $\lambda_n(\mu)$. Hence, $(\lambda_n(\hat{\mu}), \hat{\mu})$ is the solution to $0 = n^{-1} \sum_{i=1}^n \psi_{\text{EL}}(Y_i, \alpha)$, where ψ_{EL} is as defined in the lemma. This shows that $\hat{\alpha}$ is a Z-estimator, and hence, standard results concerning such estimators can be used to study limiting properties of $\hat{\alpha}$. Furthermore, since the critical point of $\log \text{EL}_n$ was assumed to be unique. α_n is the unique root of $n^{-1} \sum_{i=1}^n \psi_{\text{EL}}(Y_i, \alpha)$.

We start by showing consistency towards α^* . This is achieved by proving condition (C2) of Lemma 1. (C1) holds by condition (1) of this lemma. By assumption (2) there exists $L > 0$ such that $1 + \lambda^T M(\mu) > L$ with probability 1 for all $\alpha \in \mathcal{N}$. Combining this with

condition (3) ensures $\|\psi_{\text{EL}}(y, \alpha)\| \leq C_1 p_0(y) + C_2 p_1(y)$ for some constants $C_1, C_2 \in \mathbb{R}$. The right hand side of this inequality is a sum of F -integrable functions. This shows condition (C2) of Lemma 1, and guarantees that $\hat{\alpha}$ converges in probability to α^* .

We will use Theorem 5.21 in van der Vaart (2000) to show (S1.1). We start with the Lipschitz condition. Direct computations show $\|\frac{\partial \psi}{\partial \alpha}(y, \alpha)\| \leq C_3 p_0(y)^2 + C_4 p_0(y) p_1(y) + C_5 p_1(y) + C_6 p_1(y)^2 + C_7 p_2(y)$, for some constants C_j for $j = 3, \dots, 7$. This is a consequence of condition (3) of the lemma. Since all functions on the right hand side of this equation are square integrable with respect to the measure F , the sum is square integrable as well. This in conjunction with the mean value theorem for multivariate functions, ensures the Lipschitz property of Theorem 5.21 in van der Vaart (2000). Furthermore, $E[\|\psi(Y, \alpha^*)\|^2] < \infty$ by a similar argument. The rest of the conditions of Theorem 5.21 in van der Vaart (2000) hold by assumption. (S1.1) follows.

Strictly speaking, we have only showed that the sequence of solutions to $n^{-1} \sum_{i=1}^n \psi(Y_i, \alpha)$ satisfies the properties of the lemma. Since, $\log \text{EL}_n(\mu)$ is only equal to $-\sum_{i=1}^n \log[1 + \lambda_n(\mu)^T m(Y_i, \mu)]$ when $1 + \lambda_n(\mu)^T m(Y_i, \mu) > 0$ for all $i = 1, \dots, n$, $\partial \log \text{EL}_n(\hat{\mu}) / \partial \mu = 0$ is not immediately equivalent to $n^{-1} \sum_{i=1}^n \psi(Y_i, \hat{\alpha}) = 0$. However, by construction the empirical likelihood function is always non-negative and equal to zero when $1 + \lambda_n(\mu)^T m(Y_i, \mu) \leq 0$ for some $i = 1, \dots, n$. Hence, the maximizer of $-n^{-1} \sum_{i=1}^n \log[1 + \lambda_n(\mu)^T m(Y_i, \mu)]$ is equal to the maximizer of EL_n as long as it satisfies $1 + \lambda_n(\mu)^T m(Y_i, \mu) > 0$ for all $i = 1, \dots, n$. Furthermore, since $1 + \lambda^T m(Y_i, \mu) > 0$ for all $i = 1, \dots, n$ holds on $\mathcal{N}$ by assumption (2), $\hat{\alpha}$ will satisfy $1 + \lambda_n(\mu)^T m(Y_i, \mu) > 0$ for all $i = 1, \dots, n$ and be the maximizer of $\log \text{EL}_n(\mu)$ with probability tending to one. $\square$

Next we show a result concerning the limiting properties of $\hat{\mu}$ alone.

Theorem 1. *Let all assumptions of Theorem 2 hold true and all quantities be as defined there. Furthermore, assume that $\mu \rightarrow E \log[1 + \lambda(\mu)^T m(Y, \mu)]$ where $\lambda(\mu)$ denotes the solution to $0 = E\{m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]\}$ has a maximizer μ^* . Then the MEL estimator $\hat{\mu}$ is consistent for μ^* , and*

$$\hat{\mu} = \mu^* + n^{-1} J_{\text{EL}}^{-1} \sum_{i=1}^n \psi_{\mu}(Y_i, \mu^*) + o_{\text{pr}}(1/\sqrt{n}), \quad (\text{S1.2})$$

with $J_{\text{EL}} = H_{\mu^*} E\{\log[1 + \lambda^T M(\mu)]\}$ and $\psi_{\mu}(y, \mu) = -\nabla_{\mu} \log[1 + \lambda(\mu)^T m(y, \mu)]$, ∇_{μ} being the gradient with respect to μ .

Proof. By Lemma 2, $\hat{\mu}$ converges in probability to the last s components of α^* , the parameter solving $0 = E[\psi_{\text{EL}}(Y, \alpha)]$. By definition $\lambda(\mu)$ is the root of $E\{m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]\}$, the first q component functions in $E[\psi_{\text{EL}}(Y, \alpha)]$. Hence, $\alpha^* = (\lambda(\mu^*), \mu^*)$ where μ^* is the root of $E\{\partial m(Y, \mu)^T \lambda(\mu)/[1 + \lambda(\mu)^T m(Y, \mu)]\}$. Furthermore, the maximizer of $-E\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$ satisfies $0 = E\{[\lambda'(\mu)^T m(Y, \mu) + \partial m(Y, \mu)^T \lambda(\mu)]/[1 + \lambda(\mu)^T m(Y, \mu)]\}$, since $\mathcal{M}$ is open and m is differentiable. As $\lambda(\mu)$ is the root of $E\{m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]\}$, the right hand side of this equation is equal to $E\{\partial m(Y, \mu)^T \lambda(\mu)/[1 + \lambda(\mu)^T m(Y, \mu)]\}$. Since the root of $\Gamma^{\text{EL}}(\alpha)$ is unique, μ^* must therefore be the the maximizer of $-E\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$.

For (S1.2), let $(0, I)$ be the $(q + s) \times (q + s)$ block matrix where I is the $s \times s$ -identity matrix. Then, by (S1.1),

$$\hat{\mu} = (0, I)\hat{\alpha} = \mu^* - (0, I)\Gamma^{\text{EL}}(\alpha^*)^{-1}n^{-1} \sum_{i=1}^n \psi(Y_i, \alpha^*) + o_{\text{pr}}(1/\sqrt{n}).$$

Let H_{xy} be the matrix obtained by first differentiating $E\{\log[1 + \lambda^T M(\mu)]\}$ with respect to y and afterwards with respect to x evaluated at (λ^*, μ^*) . Then $J\Gamma^{\text{EL}}(\alpha^*)$ can be written as

the following block matrix

$$J \Gamma^{\text{EL}}(\alpha^*) = \begin{pmatrix} H_{\lambda\lambda} & H_{\mu\lambda} \\ H_{\lambda\mu} & H_{\mu\mu} \end{pmatrix}.$$

Furthermore, $\lambda'(\mu^*) = -H_{\lambda\lambda}^{-1}H_{\mu\lambda}$ by the implicit function theorem. Utilizing this and the formula for inverses of block matrices, we find $(0, I)\Gamma^{\text{EL}}(\alpha^*) = (A^{-1}\lambda'(\mu^*)^T, A^{-1}) = A^{-1}(\lambda'(\mu^*)^T, I)$, where $A = H_{\mu\mu} - H_{\lambda\mu}H_{\lambda\lambda}^{-1}H_{\mu\lambda}$. Direct computation shows that $A = J_{\text{EL}}$. Putting all of this together shows $-(0, I)\Gamma^{\text{EL}}(\alpha^*)^{-1}\psi(y, \alpha^*) = J_{\text{EL}}^{-1}(\lambda'(\mu^*)^T, I)\psi(y, \lambda^*, \mu^*) = J_{\text{EL}}^{-1}\psi_\mu(y, \mu^*)$, proving (S1.2). $\square$

We also state and show a result for plug-in estimators.

Lemma 3. *Let $Y_1, \dots, Y_n$ be i.i.d. from some distribution F and $m: \mathbb{R}^d \times \mathcal{M} \times \mathbb{R}^r \rightarrow \mathbb{R}^q$, where $\mathcal{M} \subseteq \mathbb{R}^s$ is open, be a function continuously differentiable in the last $s+r$ arguments. Assume furthermore, that $\hat{P}$ is some estimator satisfying $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_{\text{pr}}(1/\sqrt{n})$ and that the empirical likelihood function is constructed using $\hat{P}$ as a plug-in for P , i.e. using the estimating function $(y, \mu) \mapsto m(y, \mu, \hat{P})$. Furthermore, let $\hat{\mu}(\hat{P})$ be the MEL estimator and unique critical point of $\log \text{EL}_n$. Lastly, let $\hat{\lambda}$ be the unique solution to*

$$0 = n^{-1} \sum_{i=1}^n \frac{m(Y_i, \hat{\mu}, \hat{P})}{1 + \lambda m(Y_i, \hat{\mu}, \hat{P})}$$

and set

$$\psi_{\text{EL}}^P(y, \lambda, \mu, P) = [1 + \lambda^T m(y, \mu, P)]^{-1} (m(y, \mu, P)^T, \lambda^T \partial_\mu m(y, \mu, P))^T$$

with $\partial_\mu m$ being shorthand for $\partial m / \partial \mu$. Then, if $\alpha_{P,0} = (\mu^*, \lambda^*, P_0)$ is the unique solution to $0 = \Gamma_P^{\text{EL}}(\mu, \lambda, P) = (\mathbb{E}[\psi_{\text{EL}}^P(Y, \alpha_P)]^T, \mathbb{E}[h(Y, P)]^T)^T$, the vector $\hat{\alpha}_P = (\hat{\mu}^T, \hat{\lambda}^T, \hat{P}^T)^T$ converges

in probability to $\alpha_{P,0}$ and

$$\hat{\alpha}_P = \alpha_{P,0} - n^{-1} J \Gamma_P^{\text{EL}}(\alpha_{P,0})^{-1} \sum_{i=1}^n \begin{pmatrix} \psi_{\text{EL}}^P(Y_i, \alpha_{P,0}) \\ h(Y_i, P_0) \end{pmatrix} + o_p(1/\sqrt{n}), \quad (\text{S1.3})$$

provided there exists a neighbourhood $\mathcal{N}$ of $\alpha_{0,P}$ on which:

(1) the function Γ_P^{EL} is continuously differentiable with non-singular Jacobian matrix at $\alpha_{P,0}$;

(2) $\Pr(1 + \lambda^T m(Y, \mu, P) > L) = 1$ for all $(\mu, \lambda, P) \in \mathcal{N}$ and some $L > 0$;

(3) the norm of m and all first and second order partial derivatives of m with respect to μ and

P are bounded by functions p_j for $j = 0, 1, 2$ respectively, with $E[p_0(Y)^4], E[p_1(Y)^4] < \infty$

and $E[p_2(y)^2] < \infty$.

(4) the norm of $h(y, P)$ and all first order partial derivatives of h with respect to P is bounded

by functions q_1 and q_2 not depending on P and with $E[q_1(Y)], E[q_2(Y)^2] < \infty$.

Proof. By assumption, the plug-in estimator $\hat{P}$ satisfies $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_p(1/\sqrt{n})$.

Hence, the vector $\hat{\alpha}_P$ is the solution to

$$0 = n^{-1} \sum_{i=1}^n \begin{pmatrix} \psi_{\text{EL}}^P(Y_i, \lambda, \mu, P) \\ h(Y_i, P) \end{pmatrix} + \delta_n$$

where δ_n is a term of order $o_p(1/\sqrt{n})$ which does not depend on α_P . The lemma now follows from arguments completely analogue to those given in the proof of Lemma 2. $\square$

As with Lemma 2, the above result give asymptotic properties of the full vector $\hat{\alpha}_P$. In practice, however, $\hat{\mu}$ alone is of interest, and because of this, a marginal result for $\hat{\mu}$ alone is given below.

Theorem 2. Assume the conditions of Theorem 2 hold true and let all quantities be as defined here. Furthermore, let $\lambda(\mu, P)$ and P_0 denote the solutions to $0 = E\{m(Y, \mu, P)/[1 +$

$\lambda^T m(Y, \mu, P)]\}$ and $Eh(Y, P) = 0$ respectively, and assume that the function $\mu \mapsto E \log[1 + \lambda(\mu, P)^T m(Y, \mu, P)]$ has a maximizer $\mu(P)$ for each P . Then the MEL estimator $\hat{\mu}(\hat{P})$ is consistent for $\mu^* = \mu(P_0)$ and

$$\hat{\mu}(\hat{P}) = \hat{\mu}(P_0) + \mu'(P_0)(\hat{P} - P_0) + o_{\text{pr}}(1/\sqrt{n}), \quad (\text{S1.4})$$

where $\hat{\mu}(P_0)$ is the MEL estimator when P is fixed at P_0 .

Proof. Let $\alpha_{P_0} = ((\lambda^*)^T, (\mu^*)^T, P_0^T)^T$ be the root of

$$E[\Gamma_P^{\text{EL}}(\alpha_P)] = (E[\psi_{EL}^P(Y, \alpha_P)]^T, E[h(Y, P)]^T)^T.$$

Then, by definition, P_0 is the solution to $E[h(Y, P)] = 0$ and λ^* the solution to

$$E\left(\frac{m(Y, P_0, \mu^*)}{1 + \lambda^T m(Y, P_0, \mu^*)}\right) = 0.$$

Now let μ_1 be the maximizer of $-\log[1 + \lambda(\mu, P_0)^T m(Y, P_0, \mu)]$ in $\mathcal{M}$. Then μ_1 satisfies

$$0 = E\left(\frac{\partial_{\mu_1} \lambda(\mu, P_0)^T m(Y, P_0, \mu_1) + \lambda(\mu, P_0)^T \partial_{\mu_1} m(Y, P_0, \mu)}{1 + \lambda(\mu, P_0)^T m(Y, P_0, \mu_1)}\right). \quad (\text{S1.5})$$

By definition of $\lambda(\mu_1, P)$ the second term equals zero, implying

$$0 = E\left(\frac{\lambda(\mu, P_0)^T \partial_{\mu_1} m(Y, P_0, \mu)}{1 + \lambda(\mu, P_0)^T m(Y, P_0, \mu_1)}\right), \quad (\text{S1.6})$$

and ensuring that $\Gamma^{\text{EL}}(\mu_1, \lambda(\mu_1, P_0), P_0) = 0$, but by assumption this equation has a unique root at α_{P_0} . Hence, $\mu_1 = \mu^*$ and μ^* is the maximizer of $-E\{\log[1 + \lambda(\mu, P_0)^T m(Y, P_0, \mu)]\}$.

Secondly, consider the matrix $J \Gamma_P^{\text{EL}}(\alpha)$. By definition of the Jacobian matrix, we can write $J \Gamma_P^{\text{EL}}(\alpha)$ as the following block matrix:

$$J \Gamma_P^{\text{EL}}(\alpha) = \begin{pmatrix} J \Gamma^{\text{EL}}(\lambda, \mu), & \partial_P \Gamma_P^{\text{EL}}(\lambda, \mu, P) \\ 0, & \partial_P Eh(Y, P) \end{pmatrix},$$

where $J \Gamma^{\text{EL}}(\lambda, \mu)$ is equal to $J \Gamma^{\text{EL}}(\lambda, \mu)$ defined in Lemma 3 where the estimating function used is $(y, \mu) \mapsto m(y, P, \mu)$ for each fixed P and $\partial_P f$ is short-hand for $\partial f / \partial P$ for a general

function f . By inversion of block matrices and condition (1) in Lemma 2, this matrix is continuously differentiable in a neighborhood of $((\lambda^*)^T, (\mu^*)^T, P_0^T)^T$ with

$$[J \Gamma_P^{\text{EL}}(\alpha_{P,0})]^{-1} = \begin{pmatrix} [J \Gamma_P^{\text{EL}}(\lambda, \mu)]^{-1}, & \alpha'(P_0)[\partial_P \text{Eh}(Y, P)]^{-1} \\ 0, & [\partial_P \text{Eh}(Y, P)]^{-1} \end{pmatrix}$$

non-singular. In the above $\alpha(P)$ is the root of $(\lambda, \mu) \mapsto \Gamma_P^{\text{EL}}(\lambda, \mu, P)$ for each fixed P . Hence, by arguments similar to those given in the proof of Theorem 1,

$$\begin{aligned} & - (0 \ I_s \ 0) (J \Gamma_P^{\text{EL}}(\alpha_{P,0}))^{-1} \psi_P(y, \alpha_{P,0}) = \\ & J_{\text{EL}}^{-1} \psi_\mu^{P_0}(y, \mu^*) - \mu'(P_0)[\partial_P \text{Eh}(Y, P)]^{-1} h(y, P_0), \end{aligned}$$

where $\psi_\mu^{P_0}$ is equal to ψ_μ defined in Theorem 1 in the article when replacing m by $(y, \mu) \mapsto m(y, \mu, P_0)$. Lastly, by condition (4) in Lemma 2 and theorem 5.21 in van der Vaart (2000), $\hat{P} - P_0 = -[\partial_P \text{Eh}(Y, P)]^{-1} \sum_{i=1}^n h(Y_i, P_0) + o_p(1/\sqrt{n})$. Hence, $-\mu'(P_0)n^{-1}[\partial_P \text{Eh}(Y, P)]^{-1} n^{-1} \sum_{i=1}^n h(Y_i, P_0) = \mu'(P_0)(\hat{P} - P_0) + o_p(1/\sqrt{n})$. This concludes the proof. $\square$

S2 Proving Theorem 1 in the article

We first show a lemma.

Lemma 4. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$ a function continuously differentiable in the last s arguments. Let f_θ for $\theta \in \Theta \subseteq \mathbb{R}^p$, where Θ is an open set, be some parametric family with score function u and $\mu: \mathbb{R}^p \rightarrow \mathbb{R}^s$ such that $\text{E}\{m[Y, \mu(\theta)]\} = 0$ when $Y \sim f_\theta$. Fix $a \in [0, 1)$ and assume that $\hat{\theta}$ is the maximizer and unique critical point of h_n . Furthermore, let $\hat{\lambda}(\hat{\theta})$ be the unique solution to*

$$0 = n^{-1} \sum_{i=1}^n \frac{m[Y_i, \mu(\hat{\theta})]}{1 + \lambda^T m[Y_i, \mu(\hat{\theta})]} = 0$$

Lastly, let $\beta = (\theta, \lambda)$, $\hat{\beta} = (\hat{\theta}, \hat{\lambda})$, and

$$\psi_{\text{HL}}(y, \beta) = (1 - a) \begin{pmatrix} 0 \\ u(y, \theta) \end{pmatrix} - \begin{pmatrix} I, & 0 \\ 0, & a\mu'(\theta)^T \end{pmatrix} \psi_{\text{EL}}[y, \lambda, \mu(\theta)].$$

Assume that $\beta^* = (\lambda^*, \mu^*)$ is the unique root of $\Gamma^{\text{HL}}(\lambda, \mu) = \mathbb{E}[\psi(Y, \beta)]$, where $Y \sim F$. Then $\hat{\beta}$ is consistent for β^* and

$$\hat{\beta} = \beta^* - n^{-1} J \Gamma^{\text{HL}}(\beta^*)^{-1} \sum_{i=1}^n \psi_{\text{HL}}(Y_i, \beta^*) + o_{\text{pr}}(1/\sqrt{n}) \quad (\text{S2.7})$$

provided condition (2) and (3) of Lemma 2 hold true with $\alpha^* = ((\lambda^*)^T, \mu(\theta^*)^T)^T$ and the following additional assumptions

(4) the function $(\lambda, \theta) \mapsto \mathbb{E}\{(1 - a)u(Y, \theta)^T - aM'(\theta)/[1 + \lambda^T M(\theta)], M^T/[1 + \lambda^T M(\theta)]\}$ is continuously differentiable in a neighbourhood of β^* with non-singular derivative at β^* .

(5) the function $\mu(\theta)$ is twice continuously differentiable in a neighbourhood of θ^* .

(6) in a neighbourhood of θ^* , all components of u and $\partial u/\partial \theta$ are bounded by some F -square integrable functions q_1 and q_2 respectively not depending on θ .

In particular, $\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, J \Gamma^{\text{HL}}(\beta^*)^{-1} K_{\text{HL}} J \Gamma^{\text{HL}}(\beta^*)^{-1})$, with $K_{\text{HL}} = \text{Var}[\psi_{\text{HL}}(Y, \beta^*)]$, $Y \sim F$.

Proof. Notice that

$$\begin{aligned} \frac{\partial h_n(\theta)}{\partial \theta} &= \sum_{i=1}^n \left\{ (1 - a)u(Y_i, \theta) - a\mu'(\theta)^T \frac{\partial m[Y_i, \mu(\theta)]^T \lambda_n(\theta) + \lambda'_n(\theta)^T m[Y_i, \mu(\theta)]}{1 + \lambda_n(\theta)^T m[Y_i, \mu(\theta)]} \right\} \\ &= \sum_{i=1}^n \left\{ (1 - a)u(Y_i, \theta) - a\mu'(\theta)^T \frac{\partial m[Y_i, \mu(\theta)]^T \lambda_n(\theta)}{1 + \lambda_n(\theta)^T m[Y_i, \mu(\theta)]} \right\}, \end{aligned}$$

where the last equality follows from the definition of $\lambda_n(\theta)$ and ∂m is short hand for $\partial m/\partial \mu$.

Hence, since Θ is open and the maximizer of h_n in Θ is the unique critical point of the function, $\hat{\beta}$ is the unique solution to $0 = n^{-1} \sum_{i=1}^n \psi_{\text{HL}}(Y_i, \beta)$ where ψ_{HL} is as defined in the

lemma. Because of this $\hat{\beta}$ is a Z-estimator and, just as for the MEL estimator, standard results can be used to study its limiting behavior.

Let $m_2: \mathbb{R}^{d+p} \rightarrow \mathbb{R}^q$ be defined as $m_2(y, \theta) = m[y, \mu(\theta)]$. Since μ is twice continuously differentiable as a function of θ , m_2 is twice continuously differentiable as a function of θ . By the proof of Lemma 2 this ensures that there exists a neighborhood on which the norm of

$$\begin{pmatrix} I, & 0 \\ 0, & \mu'(\theta)^T \end{pmatrix} \psi_{\text{EL}}[y, \lambda, \mu(\theta)] = \begin{pmatrix} -m_2(y, \theta) \\ -\partial m_2(y, \theta)^T \lambda \end{pmatrix} [1 + \lambda^T m_2(y, \theta)]^{-1} \quad (\text{S2.8})$$

is bounded some square F -integrable function depending on neither λ nor μ . Combined with assumption (5), this proves that the norm of $\psi(y, \beta)$ is bounded by some square integrable function not depending on β . In particular, (C2) of Lemma 1 holds true. As (C1) holds by assumption (4), consistency of $\hat{\beta}$ towards β^* follows.

To show (S2.7), we will again use Theorem 5.21 in van der Vaart (2000). The only condition which remains to show is the Lipschitz condition. By the chain rule and proof of Lemma 2, the norm of all partial derivatives of (S2.8) are bounded by some square integrable function. Combining this with assumption (6) of this lemma and the mean value theorem, ensures that there exists a square integrable function r such that $\|\psi_{\text{HL}}(y, \beta_1) - \psi_{\text{HL}}(y, \beta_2)\| \leq r(y)\|\beta_1 - \beta_2\|$. This proves the Lipschitz condition of Theorem 5.21 in van der Vaart (2000) and concludes the proof. $\square$

S2.1 Proof of Theorem 1 in the article

The proof of Theorem 1 in the article now follows directly from Lemma 4 by performing computations similar to those in Section S1. The conditions needed are those stated in Lemma 4 The case of $a = 1$ follows from Theorem 1.

S3 Proof of Theorem 2 in the article

We aim to prove Theorem 2 in the paper. To do this, we first show a lemma.

Lemma 5. *Let $g: \mathbb{R}^p \rightarrow \mathbb{R}^k$ be a twice smoothly differentiable function of θ , and let*

$$\hat{\theta}(\zeta) = \operatorname{argmax}_{\theta: g(\theta)=\zeta} h_n(\theta) \quad \text{and} \quad \hat{\lambda}(\zeta) = \text{root of } \sum_{i=1}^n \frac{m[Y_i, \mu(\hat{\theta})]}{1 + \lambda^T m[Y_i, \mu(\hat{\theta})]}$$

for every value of ζ . As before, let $\hat{\theta}$ and $\hat{\lambda}$ denote the MHL estimator and solution to $\sum_{i=1}^n m[Y_i, \mu(\hat{\theta})]/\{1 + \lambda^T m[Y_i, \mu(\hat{\theta})]\}$ respectively and assume the conditions of Lemma 4 hold true. Let θ^* and λ^* denote the limits of $\hat{\theta}$ and $\hat{\lambda}$ from that result. We then have

$$\begin{pmatrix} \hat{\theta}(\zeta^*) - \hat{\theta} \\ \hat{\lambda}(\zeta^*) - \hat{\lambda} \end{pmatrix} = MV_n + o_p(1/\sqrt{n})$$

where $\zeta^* = g(\theta^*)$, $V_n = n^{-1} \sum_{i=1}^n \psi_{\text{HL}}(Y_i, \theta^*, \lambda^*)$ and $M = -AB[B^T A^{-1} B]^{-1} B^T A^{-1}$ with $A = J\Gamma_{\text{HL}}(\theta^*, \lambda^*)^{-1}$ and $B = (Jg(\theta^*), 0)^T$.

Proof. By Lagrange method of multipliers, the maximizer of h_n under the condition $g(\theta) = \zeta$ solves $0 = \nabla_{\theta} h_n(\theta) + \eta^T Jg(\theta)$ and $g(\theta) = \zeta$. Hence, $\hat{\theta}(\zeta)$ and $\hat{\lambda}(\zeta)^T$ satisfy

$$\begin{pmatrix} \hat{\theta}(\zeta^*) \\ \hat{\lambda}(\zeta^*) \\ \hat{\eta}(\zeta^*) \end{pmatrix} = \text{zero of } \sum_{i=1}^n \begin{pmatrix} \psi_{\text{HL}}^1(Y_i, \theta, \lambda) + \eta^T Jg(\theta) \\ \psi_{\text{HL}}^2(Y_i, \theta, \lambda) \\ g(\theta) - \zeta^* \end{pmatrix},$$

where ψ_{HL}^1 are the first p and last q components of ψ_{HL} respectively. By arguments completely similar to those given in Lemma 4, we therefore have that

$$\begin{pmatrix} \hat{\theta}(\zeta^*) \\ \hat{\lambda}(\zeta^*) \\ \hat{\eta}(\zeta^*) \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \theta(\zeta^*) \\ \lambda(\zeta^*) \\ \eta(\zeta^*) \end{pmatrix} = \text{zero of E} \begin{pmatrix} \psi_{\text{HL}}^1(Y, \theta, \lambda) + \eta^T Jg(\theta) \\ \psi_{\text{HL}}^2(Y, \theta, \lambda) \\ g(\theta) - \zeta^* \end{pmatrix} \quad (\text{S3.9})$$

and

$$\begin{pmatrix} \hat{\theta}(\zeta^*) \\ \hat{\lambda}(\zeta^*) \\ \hat{\eta}(\zeta) \end{pmatrix} = \begin{pmatrix} \theta(\zeta^*) \\ \lambda(\zeta^*) \\ \eta(\zeta^*) \end{pmatrix} - A^{-1} \sum_{i=1}^n \begin{pmatrix} \psi_{\text{HL}}^1[Y_i, \theta(\zeta^*), \lambda(\zeta^*)] + \eta^T Jg[\theta(\zeta^*)] \\ \psi_{\text{HL}}^2[Y_i, \theta(\zeta^*), \lambda(\zeta^*)] \\ g[\theta(\zeta^*)] - \zeta^* \end{pmatrix} + o_p(1/\sqrt{n}),$$

where A is the Jacobian of the right hand side of (S3.9). Since $\zeta^* = g(\theta^*)$, $\eta(\zeta^*) = 0$, $\theta(\zeta^*) = \theta^*$, $\lambda(\zeta^*) = \lambda^*$ and $g[\theta(\zeta^*)] - \zeta^* = 0$. Utilizing this and formulas for the inverses of block matrices, we get,

$$\begin{pmatrix} \hat{\theta}(\zeta^*) \\ \hat{\lambda}(\zeta^*) \end{pmatrix} = \begin{pmatrix} \theta(\zeta^*) \\ \lambda(\zeta^*) \end{pmatrix} + (M - A)V_n + o_p(1/\sqrt{n}).$$

This in turn, implies

$$\begin{pmatrix} \hat{\theta}(\zeta^*) \\ \hat{\lambda}(\zeta^*) \end{pmatrix} - \begin{pmatrix} \hat{\theta} \\ \hat{\lambda} \end{pmatrix} = (M - A)V_n + AV_n + o_p(1/\sqrt{n}) = MV_n + o_p(1/\sqrt{n}),$$

proving the lemma. □

Now that we have proven Lemma 5, we are ready to prove Theorem 2 in the paper.

Theorem 3. *Let the profile hybrid log-likelihood function be defined as*

$$h_{n,\text{prof}}(\zeta) = \sup_{\theta : g(\theta) = \zeta} h_n(\theta),$$

and assume all conditions of Lemma 5 hold true. Then, if there is an integrable function not depending on θ or λ bounding the norm of all second order partial derivatives of ψ_{HL} in a neighbourhood of $((\theta^*)^T, (\lambda^*)^T)^T$,

$$D_n = -2[h_{n,\text{prof}}(\hat{\zeta}) - h_{n,\text{prof}}(\zeta^*)] \xrightarrow{d} \text{Tr}\{[g'(\theta^*)(J^*)^{-1}g'(\theta^*)^T]^{-1}g'(\theta^*)(J_a^*)^{-1}K_a^*(J_a^*)^{-1}UU^T\} \quad (\text{S3.10})$$

where $U \sim N(0, I_p)$, $\zeta^* = g(\theta^*)$, $\hat{\zeta} = g(\hat{\theta})$ and θ^* and $\hat{\theta}$ are as defined in Lemma 5. In particular, when $g(\theta) \in \mathbb{R}$,

$$D_n \xrightarrow{d} \frac{\nabla g(\theta^*)^T (J_a^*)^{-1} K_a^* (J_a^*)^{-1} \nabla g(\theta^*)}{[\nabla g(\theta^*)^T (J_a^*)^{-1} \nabla g(\theta^*)]} \chi_1^2, \quad (\text{S3.11})$$

where χ_1^2 is the chi-squared distribution with one degree of freedom. Otherwise the limit distribution is a generalized chi-squared distribution.

Proof. By the definitions in Lemma 5, $D_n = -2\{h(\hat{\theta}) - h[\hat{\theta}(\zeta^*)]\}$. Hence,

$$D_n = \sqrt{n} \begin{pmatrix} \hat{\theta}(\zeta^*) - \hat{\theta} \\ \hat{\lambda}(\zeta^*) - \hat{\lambda} \end{pmatrix}^T W_n \sqrt{n} \begin{pmatrix} \hat{\theta}(\zeta^*) - \hat{\theta} \\ \hat{\lambda}(\zeta^*) - \hat{\lambda} \end{pmatrix} + O_p(1/\sqrt{n}),$$

where $W_n = n^{-1} \sum_{i=1}^n J\psi_{\text{HL}}(Y_i, \hat{\theta}, \hat{\lambda})$. This follows from a Taylor expansion and the fact that there is an integrable function bounding the norm of all second order partial derivatives of ψ_{HL} in a neighbourhood of $((\theta^*)^T, (\lambda^*)^T)^T$. This last fact also ensures $W_n \xrightarrow{p} A^{-1}$, see e.g. Lemma 4 in Appendix B.4 in Dæhlen et al. (2024). Combining all of this with Lemma 5, the central limit theorem and Slutsky's theorem, ensures $D_n \xrightarrow{d} V^T M W M M V$, where $V \sim N[0, \text{Var } \psi_{\text{HL}}(\theta^*, \lambda^*)]$. Direct computations show (S3.10) and (S3.11). $\square$

S4 Proving Theorem 3 in the article

We start with a lemma, yet again.

Lemma 6. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^{d+s+r} \rightarrow \mathbb{R}^q$ a function continuously differentiable in the last $s+r$ arguments. Let $\hat{P}$ be some estimator satisfying $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_p(1/\sqrt{n})$ and assume that $(y, \mu) \mapsto m(y, \mu, \hat{P})$ is used in construction of the empirical likelihood function EL_n . Let f_θ for $\theta \in \Theta \subseteq \mathbb{R}^p$, where Θ is an open set, be some parametric family with score function u and $\mu: \mathbb{R}^{p+r} \rightarrow \mathbb{R}^s$ such that*

$E\{m[Y, \mu(\theta, P), P]\} = 0$ when $Y \sim f_\theta$. Fix $a \in [0, 1)$ and assume that $\hat{\theta}(\hat{P})$ is the maximizer of h_n and unique critical value. Furthermore, let $\hat{\lambda}$ be the unique solution to

$$0 = n^{-1} \sum_{i=1}^n \frac{m[Y_i, \mu(\hat{\theta}, \hat{P}), \hat{P}]}{1 + \lambda^T m[Y_i, \mu(\hat{\theta}, \hat{P}), \hat{P}]} = 0$$

Lastly, let $\beta_P = (\theta, \lambda, P)$, $\hat{\beta} = (\hat{\theta}(\hat{P}), \hat{\lambda}, \hat{P})$, and

$$\psi_{\text{HL}}^P(y, \beta) = (1 - a) \begin{pmatrix} 0 \\ u(y, \theta) \\ 0 \end{pmatrix} - a \begin{pmatrix} I, & 0, & 0 \\ 0, & \mu'(\theta)^T, & 0 \\ 0, & 0, & I \end{pmatrix} \begin{pmatrix} \psi_{\text{EL}}^P[y, \lambda, \mu(\theta), P] \\ h(y, P) \end{pmatrix}.$$

Then if $\beta_{P,0} = (\theta^*, \lambda^*, P_0)$ is the unique root of $\Gamma_P^{\text{HL}}(\lambda, \mu) = E[\psi_{\text{HL}}^P(Y, \beta)]$ where $Y \sim F$, $\hat{\beta}_P$ is consistent for $\beta_{P,0}$ and

$$\hat{\beta} = \beta^* - n^{-1} J \Gamma_P^{\text{HL}}(\beta^*)^{-1} \sum_{i=1}^n \psi_{\text{HL}}^P(Y_i, \beta^*) + o_{\text{pr}}(1/\sqrt{n}) \quad (\text{S4.12})$$

provided condition (2-4) of Lemma 3 hold true with $\alpha_{P,0} = (\mu(\theta^*, P_0)^T, (\lambda^*)^T, P_0^T)^T$ and the following additional assumptions

(5) The function Γ_P^{HL} is continuously differentiable in a neighbourhood of $\beta_{P,0}$ with non-singular derivative at $\beta_{P,0}$;

(6) For all P in an neighborhood of P_0 , the function $\mu(\theta, P)$ is twice continuously differentiable with respect to θ and P in a neighbourhood of $(\theta(P_0)^T, P_0^T)^T$.

(7) There exists a neighborhood of θ^* on which all components of u and $\partial u / \partial \theta$ are bounded by some F -square integrable functions q_1 and q_2 respectively not depending on θ .

Proof. The result follows from arguments similar to those given in the proof of Lemma 3 and Lemma 4 and is left out. □

S4.1 Proof of Theorem 3 in the article

Theorem 2 now follows directly from the above lemma by arguing and performing computations similar to those in the proof of Theorem 2. The conditions are those stated in Theorem 2. The case when $a = 1$ follows from Theorem 2.

S5 Simulations

To make our results more trustworthy we performed a small simulation study, and did the following $B = 100$ times. We drew $n = 100$ data points from a Poisson distribution with rate parameter 1.5 and fitted geometric distributions to the data with maximum hybrid likelihood for a selection of a -values. The geometric distribution has density $\log f_p(y) = y \log(1 - p) + \log p$ for $y = 0, 1, \dots$. We used the probability of exceeding 1 as the control parameter and $m(y, \mu) = I(Y \leq 2) - \mu$ as the estimating function in the empirical likelihood function. We will refer to the estimators as $\hat{p}_b(a)$ for $b = 1, \dots, B$ and $a = 0, 0.01, \dots, 0.99$. The mean of $\hat{p}(a)_1, \dots, \hat{p}_B(a)$ for each of the selected a values are shown in Figure 1 together with the theoretical limit, the minimizer of $d_a(g, f_p)$ where g is the pmf in the $\text{Pois}(2)$ distribution. From the figure, we notice that the mean is close to the theoretical limit. In each iteration we also computed the estimated variance of $\sqrt{n}\hat{p}(a)$ for each a using the estimators defined in Definition 1 in the paper. The mean of the B estimator is shown in Figure 1 together with the empirical variance of $\sqrt{n}\hat{p}_1(a), \dots, \sqrt{n}\hat{p}_B(a)$ for each a . Again, the theory and estimators seem to agree with the simulation results.

We also computed the HFIC score for each a in every iteration. We used the control parameter as the focus parameter and the empirical mean of $I(Y_i \leq 1)$ for $i = 1, \dots, n$ as consistent estimator. This satisfies equation (11) in the paper with $\phi(y, \xi) = I(y \leq$

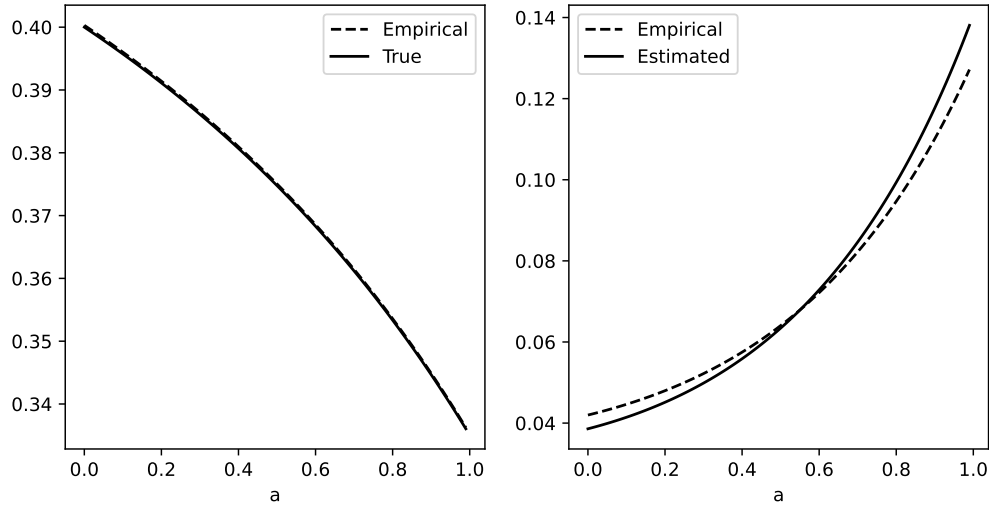


Figure 1: The plot to the left shows the mean of $\hat{p}_1(a), \dots, \hat{p}_B(a)$ for $a = 0, 0.01, \dots, 0.99$ together with the theoretical limit. To the right, the empirical variance of $\sqrt{n}\hat{p}_1(a), \dots, \sqrt{n}\hat{p}_B(a)$ for each a is displayed together with the mean of the estimator of the variance obtained using Definition 1 in the paper.

1) $-\xi$. A plot of the mean of the B HFIC curves is displayed in Figure 2 together with the observed mean squared error of the MHL estimator of the focus parameter. As before, the approximation seem to work very well. Since, the HFIC curves were computed in every iteration, we also registered the balance parameter minimising the criterion in each case. A histogram of the a chosen is shown in Figure 2, and reveals that $a \neq 0, 0.99$ is often chosen, indicating that a true hybrid model is often preferred. Furthermore, the values of a being most frequently chosen are those close to the minimizer of the MSE, indicating that HFIC is working as intended when it comes to choosing the balance parameter.

Lastly, we computed the MHL estimate of the focus parameter with the optimal a chosen by HFIC in each iteration. This gave us estimates $\hat{\mu}_1(\hat{a}_1), \dots, \hat{\mu}_B(\hat{a}_B)$. The MSE of these estimators were computed to be 0.003 which is smaller than the MSE of the ML estimators $\hat{\mu}_1(0), \dots, \hat{\mu}_B(0)$ which was 0.007. This supports the claim that using hybrid likelihood in

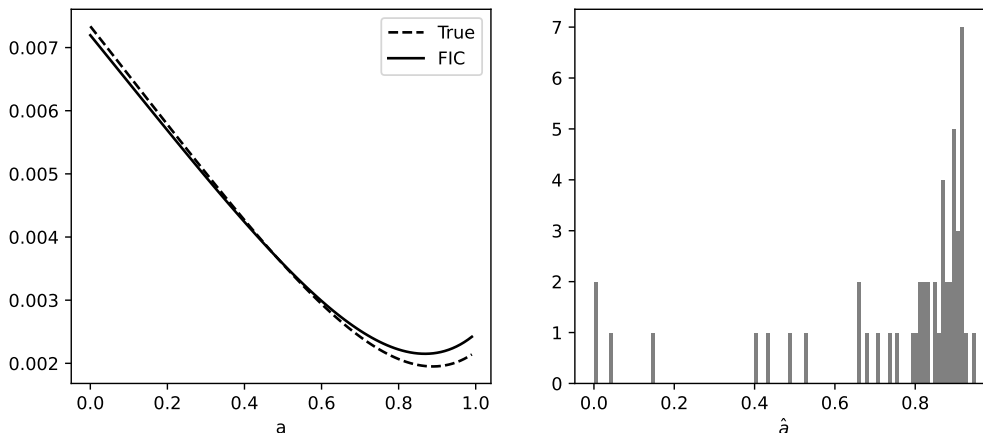


Figure 2: The plot to the left shows the mean of HFIC for $a = 0, 0.01, \dots, 0.99$ together with the true MSE. To the right, a histogram of $\hat{a}_1, \dots, \hat{a}_B$ is displayed.

place of classic maximum likelihood can make results and estimators more precise.

S6 A multivariate example

In this section, we illustrate how the hybrid likelihood framework can be applied in a multivariate setting. We will work with a dataset consisting of the average body and brain weight of 56 different species of mammals (from the dataset `sleep` found in several R-packages).

After taking the logarithm of the weight measurements, the data appears relatively binormal, as can be seen from see Figure 3a. We will therefore work with the log-transformed dataset: $z_i = (x_i, y_i) = (\log x_{0,i}, \log y_{0,i})$ for $i = 1, \dots, n$, and fit a binormal model to this data. Here $x_{0,i}$ and $y_{0,i}$ are the body and brain weight of data point i respectively. We will fit the model using the hybrid likelihood framework.

The binormal model seem to fit the data relatively well, but it has some issues with precise estimation for the untransformed data $z_{0,i} = (x_{0,i}, y_{0,i})$ for $i = 1, \dots, n$. The ML model estimates the expected value of Z_0 to be $(375.148, 0.397)$, which is quite far from the

empirical estimates of (202.982, 0.282). To correct for this, we used the hybrid likelihood framework with control parameter equal to the mean in the original distributions, i.e.

$$m(z, \mu) = \begin{pmatrix} \exp(z_1) - \mu_1 \\ \exp(z_2) - \mu_2 \end{pmatrix} \quad \text{and} \quad \mu(\gamma, \Sigma) = \begin{pmatrix} \exp(\gamma_1 + \Sigma_{1,1}/2) \\ \exp(\gamma_2 + \Sigma_{2,2}/2) \end{pmatrix},$$

where $\Sigma_{i,j}$ is the (i, j) -th entry of Σ . The formula for μ follows directly from the fact that the marginal distributions of X_0 and Y_0 are log-normal when (X, Y) is binormal.

For every $a = 0, 0.01, 0.02, \dots, 0.99$ we fitted binormal models to the data by maximizing the corresponding hybrid likelihood function. A scatter plot of the data together with a selection of level curves in the distributions fitted using hybrid likelihood for different balance parameters can be found in Figure 3a. Although not immediately apparent from the plot, the level curves shift and shrink slightly as $a \rightarrow 1$. Furthermore, the hybrid estimate of the control parameters changes a lot as a grows, from the ML estimate towards the empirical mean of Z_0 . A plot showing how the estimate of the control parameter changes as $a \rightarrow 1$ is shown in Figure 3b.

To select a proper value for a we will use the FIC, but what should we use as the focus parameter in this case? There are of course many options. Humans are the mammals with the largest value of EY_0/X_0 . When trying to make inference about this quantity, the ratio of brain to body weight might be a natural choice. Alternatively, the covariance between these two quantities could be used, or the regression coefficient when regressing Y_0 onto X_0 . All of these options are valid choices and suited for different analyses. We have, however, chosen to do something different, and will use both the expected body and expected brain weight as focus parameters. We chose to do this in an attempt to make the example more generalizable to other types of multivariate data.

The FIC can be used to estimate the MSE of each component of $\hat{\mu}_a$ for every value of

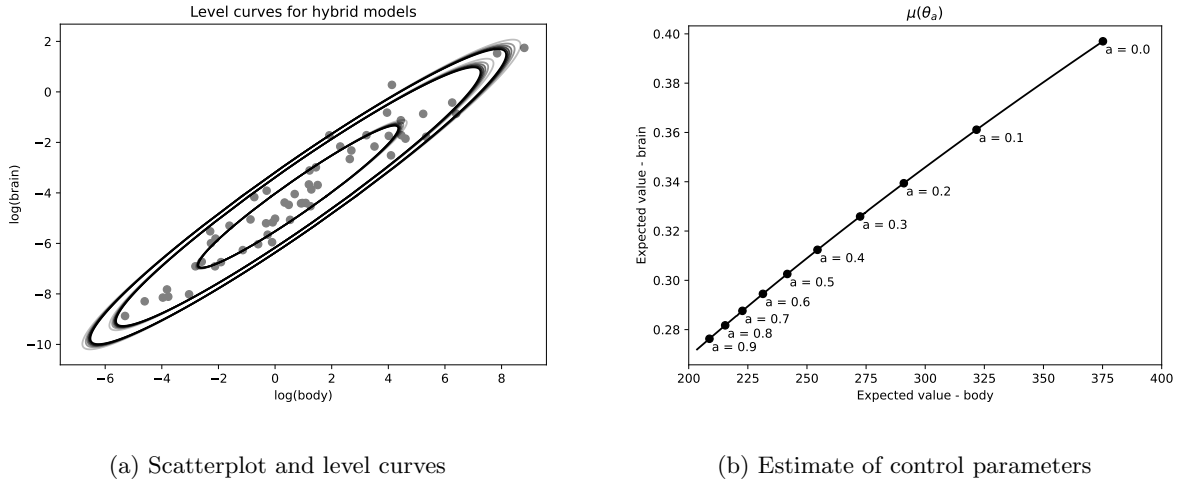


Figure 3: To the left: a scatterplot of (x_i, y_i) for $i = 1, \dots, n$ together with 0.5, 0.9 and 0.95 level curves in the distributions fitted using maximum hybrid likelihood. The values used for a are 0, 0.2, 0.4, 0.6, 0.8, 0.99 and are indicated by the lightness of the lines (from light to dark). We, the homo sapiens, are just on the upper outside of the 0.95 level. To the right: a line plot of the MHL estimator of the control parameter for different values of the balance parameter.

the balance parameter. As the consistent estimator $\hat{\xi}_0$ we used $\hat{\xi}_1 = \overline{\exp(X)}$ and $\hat{\xi}_2 = \overline{\exp(Y)}$ with corresponding ϕ -functions $\phi_1(x) = \exp(x) - \hat{\xi}_1$ and $\phi_2(y) = \exp(y) - \hat{\xi}_2$ and c -s both equal to zero. The FIC-scores for each focus parameter are shown in Figure 4. When using the expected brain weight as focus parameter, $\hat{a}_2 = 0.85$ is chosen. Focusing on body leaves the largest a -value checked ($\hat{a}_1 = 0.99$) as optimal. We therefore get different choices of balance parameters, and hence different models, with different focus parameters. But what if we want to focus on both the expected body and the expected brain weight? Ideally we would like to choose the balance parameter in a way which takes the MSE of both focus parameters into account. One way to do this, is to use the average focused information criterion, or AFIC for short.

The AFIC is thoroughly explained in Claeskens and Hjort (2008), but the version of the

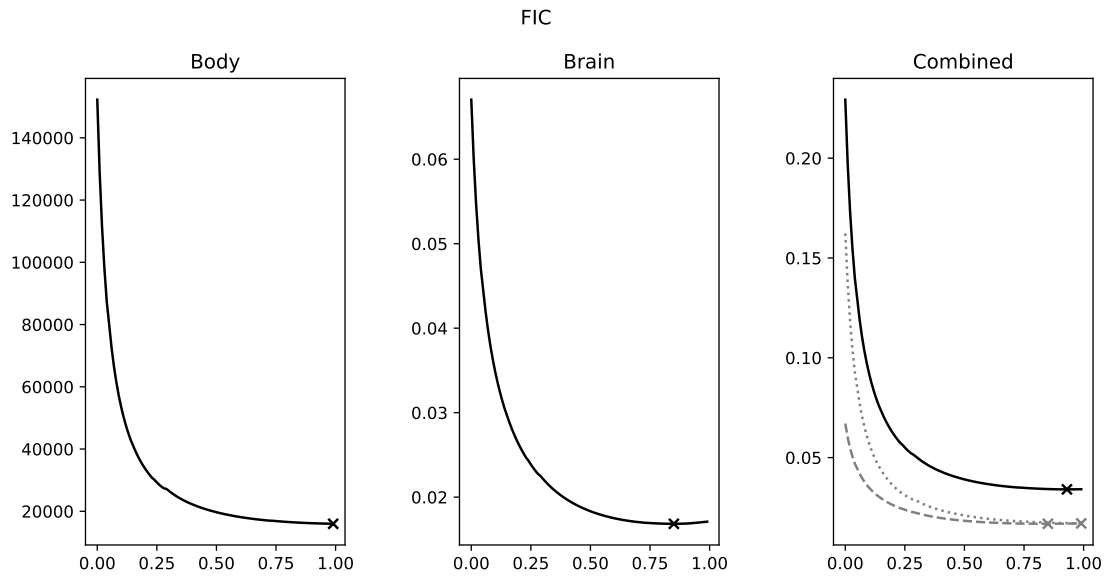


Figure 4: The two leftmost plots show FIC-values as functions of the balance parameter when using EX_0 and EY_0 as focus parameters respectively. The plot to the right displays the AFIC together with the reweighed FIC curves from the two plots to the left (dotted for body and dashed for brain). In each plots, the minimum is indicated by a cross.

AFIC we will use is based on the presentation given in Jullum and Hjort (2017). Assume we have a set of focus parameters $\xi(t)$ for $t \in T$ and suppose we have a weight function $w(t)$ describing the relative importance of each $\xi(t)$ and integrating to one. In this case, the weighted MSE is given by

$$\text{wMSE} = \int_T w(t) \text{bias}(t)^2 dt + \int_T w(t) \text{var}(t) dt.$$

Inspired by this, we define the AFIC in the following way

$$\text{AFIC} = \max \left\{ 0, \int_T w(t) \left[\hat{b}(t)^2 + n^{-1} \left(2\hat{b}(t)\hat{c}(t) - \hat{\kappa}(t) \right) \right] dt \right\} + n^{-1} \int_T w(t) \hat{\tau}(t) dt, \quad (\text{S6.13})$$

where $\hat{b}(t)$, $\hat{c}(t)$, $\hat{\kappa}(t)$ and $\hat{\tau}(t)$ are the estimators defined in Definition 1 in the paper, for each $t \in T$. Notice that the AFIC is not equal to $\int w(t) \text{FIC}(t) dt$, as the truncation in the bias term is done on the estimator of the integrated squared bias, rather than on each individual bias estimate. For mathematical details and further insights about the AFIC, consult Jullum and Hjort (2017).

When T is discrete, the integral in (S6.13) is equal to a sum. Hence, for our illustration with the body and brain weight of mammals,

$$\text{AFIC} = \max \left\{ 0, w_1 \hat{B}_1 + w_2 \hat{B}_2 \right\} + n^{-1} (w_1 \hat{\tau}_1 + w_2 \hat{\tau}_2), \quad (\text{S6.14})$$

where $\hat{B}_j = \hat{b}_j^2 + n^{-1}(2\hat{b}_j\hat{c}_j - \hat{\kappa}_j)$ is the non-truncated estimate of the squared bias for focus parameter j . For weights we chose $w_1 = w_1^*/(w_1^* + w_2^*)$ and $w_2 = w_2^*/(w_1^* + w_2^*)$, where w_1^* and w_2^* are the empirical variance of the (non-transformed) body and brain weights respectively. This ensures that the two focus parameters are put on the same scale and given similar weight in the AFIC. The resulting AFIC curve is shown in Figure 4 together with $w_1 \text{FIC}_1$ and $w_2 \text{FIC}_2$ for comparison. The minimum of each curve is indicated by a cross, and the minimizer of the AFIC was 0.93, which is in-between the minima of FIC_1 and FIC_2 .

Although the focus parameters chosen in this example might not be ideal for all types of analysis, we believe this example illustrates well how the hybrid likelihood framework can be effectively used in a multivariate setting. When working with this type of data, transforming the variables into something relatively simple (e.g. approximately binormally distributed data) before starting the modelling might be highly useful as visually inspecting multivariate scatter plots and deciding on full multivariate distributions can be very challenging. When transforming the data, however, the model fitted might be lacking in certain aspects for the untransformed data, and in such settings, the hybrid likelihood framework can be used effectively to correct for this. We also believe that the concept of AFIC, although not exclusively useful in the multivariate context, offers a natural way of dealing with multiple relevant focus parameters, which might be more common in multivariate settings.

Bibliography

- Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge, England: Cambridge University Press.
- Dæhlen, I., N. L. Hjort, and I. Hobæk Haff (2024). Accurate bias estimation with applications to focused model selection. *Scandinavian Journal of Statistics* 51(2), 724–759.
- Imbens, G. W., P. Johnson, and R. H. Spady (1998). Information theoretic approaches to inference in moment condition models.
- Jullum, M. and N. L. Hjort (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica* 27(3), 951–981.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics* 35(2), 634–672.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.