

Model robust hybrid likelihood

Ingrid Dæhlen and Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract: The article concerns hybrid combinations of empirical and parametric likelihood functions. Combining the two allows classical parametric likelihood to be crucially modified via the nonparametric counterpart, making possible model misspecification less problematic. Limit theory for the maximum hybrid likelihood estimator is sorted out, also outside the parametric model conditions. Results include consistency of the estimated parameter in the parametric model towards a well-defined limit, as well as asymptotic normality after proper scaling and centring of the same quantity. Our results allow for the presence of plug-in parameters in the hybrid and empirical likelihood framework. Furthermore, the variance and mean squared error of these estimators are studied, with recipes for their estimation. The latter is used to define a focused information criterion, which can be used to choose how the parametric and empirical part of the hybrid combination should be balanced. This allows for hybrid models to be fitted in a context driven way, minimizing the estimated mean squared error for estimating any pre-specified quantity of interest.

Key words and phrases: empirical likelihood, focused model selection, large-sample theory, parametrics and nonparametrics, robustness.

1. Introduction and summary

Parametric models are of course widely used in statistics. Such procedures are often interpretable, and under model conditions they are in some sense asymptotically optimal. When the model fitted is wrongly specified, however, these properties are no longer guaranteed to hold and instead bias can be introduced. Nonparametric methods, on the other hand,

typically have little to no bias. They can, however, be quite variable, which might make the methods too imprecise to be useful in practice. In this article, we study hybrid likelihood theory, a middle ground between these two extreme approaches. We derive new and model agnostic results for the framework and greatly extend its applicability by proposing ways of deciding “how parametric” the method should be.

Hybrid likelihood theory was introduced in Hjort, McKeague, and Van Keilegom (2018) and is a framework for combining the inference of nonparametric and parametric likelihood theory. For a given parametric model, with likelihood function $L_n(\theta)$, this is achieved by studying the hybrid likelihood function

$$\text{HL}_n(\theta) = L_n(\theta)^{1-a} \text{EL}_n[\mu(\theta)]^a, \quad (1.1)$$

where $\text{EL}_n[\mu(\theta)]$ is the empirical likelihood function, described in Section 2, a nonparametric counterpart to $L_n(\theta)$, evaluated at a context-driven control parameter $\mu(\theta)$, and a is a tuning parameter deciding how much weight should be put on the parametric and nonparametric parts of HL_n . The control parameter μ is some pre-specified parameter we wish to estimate robustly, i.e. without necessarily trusting the parametric model. Working with HL_n instead of either of its components separately allows us to choose a middle ground between purely parametric and nonparametric approaches.

For the nonparametric part of HL_n , the empirical likelihood function is used. This concept was first introduced in Owen (1988), and has since found numerous extensions and applications. For an overview, see e.g. Owen (2001) or Hjort et al. (2009). So far, however, there have been few attempts to combine parametric and empirical likelihoods. Certain works study ways of combining the two methods when a parametric model is available for a subset of the data only, see e.g. Qin (1994). This is, however, quite different from the

framework of Hjort, McKeague, and Van Keilegom (2018), where information from empirical and parametric likelihoods constructed using the whole sample are combined. Because of this, our hybrid likelihood framework turns out to be more related with semi-parametric methods like those of e.g. Olkin and Spiegelman (1987) or Mazo and Portier (2021). Furthermore, the framework bears certain similarity with methods developed from a rather different Bayesian semi-parametric setup in Hjort (1986). There a given parametric model has its likelihood modified by certain control sets, via a pinned down Dirichlet process on well-defined residuals.

In this article, we derive new results for both empirical and hybrid likelihood theory. In Section 2, we use arguments similar to those of Molanes Lopez, Van Keilegom, and Veraverbeke (2009) to derive an alternative characterization of the empirical likelihood function. This allows us to give an alternative characterization of the maximum empirical likelihood estimator (MEL estimator), defined and studied in Qin and Lawless (1994). Rather than arguing as in that article, however, we use the framework of generalized estimating equation (see e.g. Hansen (1982), Hall and Inoue (2003) or Shao (2003, chapter 5)) and reach results similar to those of Imbens, Spady, and Johnson (1998), Schennach (2007) and Chib et al. (2018), where the authors show and utilize that the MEL estimator can be expressed as a Z-estimator, the zero of a mean of functions. This alternative characterization and method of proof allows us to also study the limiting behavior of the MEL estimator when plug-in estimators are used in the empirical likelihood function, i.e. $\hat{\mu}(\hat{P})$, the maximizer of $\mu \mapsto \text{EL}_n(\mu, \hat{P})$ where $\hat{P}$ is some estimator. The behavior of $\log \text{EL}_n[\hat{\mu}(\hat{P}), \hat{P}]$ has been studied in Hjort et al. (2009) and Molanes Lopez, Van Keilegom, and Veraverbeke (2009) worked with $\hat{\mu}(\hat{P})$ when the plug-in estimator is a function of μ , equal to the maximizer

of $EL_n(\mu, P)$ for each fixed μ . So far, however, no general result concerning the limiting behavior of $\hat{\mu}(\hat{P})$ for a general plug-in estimator has been derived. We believe our results fill this gap in literature.

In Section 3 we study the asymptotic properties of the hybrid likelihood function and its maximizer. Hjort, McKeague, and Van Keilegom (2018) did this, to some degree, by deriving asymptotic results for the hybrid likelihood function and its maximizer when the parametric model used in construction of the hybrid likelihood function is correctly specified, but their results do not hold true under model misspecification. Furthermore, we also derive results covering the case with plug-in estimators, which was not studied in Hjort, McKeague, and Van Keilegom (2018). In Section 3 we use the theorems of Section 2 to show that the MHL estimator aims for the minimizer of a well-defined divergence $d_{a,\mu}$ and extend the results of Hjort, McKeague, and Van Keilegom (2018) to the case of possible model misspecification. We also mention that some preliminary results were reached in the master thesis Dæhlen (2021).

For hybrid likelihood to be useful in practice, we need to choose how much weight should be put on the parametric and nonparametric part of HL_n . In Section 4, we construct a focused information criterion (FIC) for models fitted with hybrid likelihood. The FIC was first introduced in Claeskens and Hjort (2003) for misspecified models where the degree of misspecification disappears in the limit, and in Hjort, McKeague, and Van Keilegom (2018) the authors derive a FIC similar in spirit to those of Claeskens and Hjort (2003). Our version of the FIC, however, is based upon the relatively newer version of the criterion introduced in Jullum and Hjort (2017) with important modifications in Dæhlen, Hjort, and Hobæk Haff (2023). The criterion directly estimates the mean squared error of some pre-

specified parameter of interest, which may or may not be equal to the control parameter. Using our proposed strategy to set a should therefore, in theory, be asymptotically optimal and lead to more precise estimation than either of the methods can achieve on their own. This holds true without requiring model misspecification to disappear in the limit or to be absent altogether. Our derivations are based on the model robust results derived in Section 3.1, and would not be possible with previous methods like those of Qin and Lawless (1994), White (1982), or Hjort, McKeague, and Van Keilegom (2018). We therefore believe we offer a new way to fit a hybrid parametric-nonparametric model in a way that is theoretically justified and guaranteed to lead to improved asymptotic inference.

After having discussed also regression settings, in Section 5, we go on to illustrate the hybrid likelihood machinery in Section 6, pertaining to the important battle death series for all major interstate wars from 1823 to the present; see the Correlates of War data set, starting with Sarkees and Wayman (2010) and later curated by other scholars. Our analysis is a contribution to the ongoing discussion of whether wars have become somewhat less deadly over time.

2. Empirical likelihood

Before we can study the hybrid likelihood function and its maximizer, we define and take a closer look at the empirical likelihood function. In Section 3, the arguments will be extended to the context of hybrid likelihood.

Empirical likelihood is a nonparametric tool for making inference about solutions to equations on the form $0 = E[m(Y, \mu)]$, where $Y \in \mathbb{R}^d$ is a random variable, $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$ is some function and $\mu \in \mathbb{R}^s$ is the parameter we are solving the equation with respect to. Such

expressions are called estimating equations, and m is referred to as the estimating function. Many quantities can be expressed as solutions to equations on the form of $E[m(Y, \mu)] = 0$ for some function m . Examples include quantiles, moments, probabilities, and regression coefficients.

Empirical likelihood theory is based on the study of the empirical likelihood function first introduced in Owen (1988). For an independent and identically distributed (i.i.d.) sample of random variables $Y_1, \dots, Y_n \sim F$ in $\mathbb{R}^d$ and a fixed $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$, the empirical likelihood function for the solution to $E[m(Y, \mu)] = 0$ takes the following form

$$\text{EL}_n(\mu) = \max \left\{ \prod_{i=1}^n w_i : \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i m(Y_i, \mu) = 0, w_i \geq 0 \right\}. \quad (2.2)$$

The main result of empirical likelihood theory states that if μ_0 is the solution to $E[m(Y, \mu)] = 0$ and the second moment of $m(Y, \mu_0)$ with $Y \sim F$ is finite, $-2 \log \text{EL}_n(\mu_0)$ converges in distribution to a chi-squared distributed variable with q degrees of freedom. This result can be used to construct approximate non-parametric confidence intervals and hypothesis tests. For proofs, consult Owen (2001) and Hjort, McKeague, and Van Keilegom (2018).

Using the Lagrange method, one can show

$$\text{EL}_n(\mu) = \prod_{i=1}^n [1 + \lambda_n(\mu)^T m(Y_i, \mu)]^{-1}, \quad (2.3)$$

where $\lambda_n(\mu)$ is the solution to

$$0 = n^{-1} \sum_{i=1}^n \frac{m(Y_i, \mu)}{1 + \lambda_n^T m(Y_i, \mu)} \quad (2.4)$$

under the condition that $1 + \lambda_n^T m(Y_i, \mu) \leq 0$ for all $i = 1, \dots, n$. When no such $\lambda_n(\mu)$ exists, $\text{EL}_n(\mu) = 0$. For derivations, see e.g. Owen (2001).

The right-hand side of (2.4) is a mean of functions. Because of this, $\lambda_n(\mu)$ is a so-called Z-estimator for each fixed value of μ . A Z-estimator is an estimator which can be

expressed as the zero of an expression on the form $\sum_{i=1}^n \psi(Y_i, \theta)$ for some known function ψ . Such estimators are widely studied in the literature (see e.g. Van der Vaart (2000, Ch. 5)), and we will now use standard theory to derive an asymptotically equivalent expression for $\log \text{EL}_n(\mu)$. For simplicity, we will leave μ out of the notation and replace $m(Y_i, \mu)$ by M_i and $m(Y, \mu)$ for a general $Y \sim F$ by M .

Firstly, standard theory guarantees that a solution to (2.4) exists with probability tending to one. In addition, the sequence of solutions has the following properties,

$$\lambda_n \xrightarrow{\text{pr}} \lambda \quad \text{and} \quad \sqrt{n}(\lambda_n - \lambda) = S^{-1}V_n + o_{\text{pr}}(1) \xrightarrow{d} N(0, S^{-1}), \quad (2.5)$$

where λ is the solution to the equation $0 = E[M/(1 + \lambda^T M)]$ such that $0 = \Pr(1 + \lambda^T M \leq 0)$ and $S = E[MM/(1 + \lambda^T M)^2]$ and $V_n = n^{-1/2} \sum_{i=1}^n M_i/(1 + \lambda^T M_i)$. For the above to hold, a relatively standard set of conditions needs to be satisfied. See e.g. Van der Vaart (2000, Thms. 5.31 and 5.42). For these results to be applicable, however, we need to assume there exists $L > 0$ and an open ball, Λ , around λ such that $\Pr(1 + x^T M > L) = 1$ for all $x \in \Lambda$. This ensures that the function $(y, \lambda) \mapsto m(y, \lambda)/[1 + \lambda^T m(y, \lambda)]$ is smooth in λ with probability 1 and that the theorems can be applied.

Since a solution to (2.4) exists with probability tending to one, we can without loss of generality assume that EL_n can be written as in (2.3). Taking logarithms and Taylor expanding (2.2), reveals $\log \text{EL}_n(\mu) + \sum_{i=1}^n \log(1 + \lambda^T M_i) = -\sqrt{n}(\lambda_n - \lambda)^T V_n + \frac{1}{2}\sqrt{n}(\lambda_n - \lambda)^T S_n \sqrt{n}(\lambda_n - \lambda) + \epsilon_n$, where $S_n = n^{-1} \sum_{i=1}^n M_i M_i^T / (1 + \lambda^T M_i)^2$. By (2.5) and the law of large numbers, this implies $\log \text{EL}_n(\mu) + \sum_{i=1}^n \log(1 + \lambda^T M_i) = -\frac{1}{2}V_n S^{-1} V_n + \epsilon_n + o_{\text{pr}}(1)$. Since open balls are convex, Taylor's theorem and the triangle inequality ensures $|\epsilon_n| \leq \|\lambda_n - \lambda\|^3 \sum_{i=1}^n \|M_i\|^3 / 3L^3$. If the third moment of $\|M\|$ is finite, the law of large numbers and (2.5) guarantees that the right-hand side of this inequality is of order $O_{\text{pr}}(1/\sqrt{n})$. In

particular, this shows

$$\log \text{EL}_n(\mu) = - \sum_{i=1}^n \log[1 + \lambda(\mu)^T m(Y_i, \mu)] - \frac{1}{2} V_n(\mu)^T S(\mu)^{-1} V_n(\mu) + o_{\text{pr}}(1), \quad (2.6)$$

with probability tending to one. Here we have added μ to the notation for clarity and easier reference.

Let μ_0 solve the estimating equation $E[M(\mu)] = 0$. Then $E\{M(\mu_0)/[1 + 0 \cdot M(\mu_0)]\} = 0$. Hence, $\lambda(\mu_0) = 0$. Plugging this into (2.6), shows $-2 \log \text{EL}_n(\mu_0) = V_n(\mu_0)^T S(\mu_0)^{-1} V_n(\mu_0) + o_{\text{pr}}(1) \xrightarrow{d} \chi_q^2$ as $V_n(\mu_0)$ converges in distribution to a $N[0, S(\mu_0)]$ -distribution by the central limit theorem. Hence, (2.6) allows us to rediscover the main theorem of empirical likelihood theory. This is not the only application of the alternative characterization in (2.6), however. After dividing by n we notice that the right-hand side of the equation is asymptotically equivalent to $-n^{-1} \sum_{i=1}^n \log[1 + \lambda(\mu)^T m(Y_i, \mu)]$. Because of this, we would expect its maximizer to be ‘almost’ an M-estimator, a maximizer of a mean of functions (see e.g. Van der Vaart (2000, Ch. 5) for definitions and properties of M-estimators). The following theorem shows that this intuition is indeed correct. The proof is given in the supplementary material.

Theorem 1. *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 continuously differentiable in the last s arguments. Let $\hat{\mu}$ be the unique maximizer of the empirical likelihood function, constructed using m and the data $Y_1, \dots, Y_n$. Let $\lambda(\mu)$ be the solution to $0 = E\{m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]\}$ and μ_0 the unique maximizer of $\mu \mapsto -E\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$. Then the MEL estimator $\hat{\mu}$ is consistent for μ_0 and*

$$\hat{\mu} = \mu_0 + n^{-1} J_{\text{EL}}^{-1} \sum_{i=1}^n \psi_{\mu}(Y_i, \mu_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (2.7)$$

where $\psi_\mu(y, \mu) = [\partial m(y, \mu)^T \lambda(\mu) + \lambda'(\mu)^T m(y, \mu)] / [1 + \lambda(\mu)^T m(y, \mu)]$, $J_{\text{EL}} = H_{\mu_0} \log[1 + \lambda(\mu)^T m(Y, \mu)]$ and ∂m is short-hand for $\partial m / \partial \mu$, provided the following conditions hold true:

- (1) the function $(\lambda, \mu) \mapsto \mathbb{E}\{[1 + \lambda^T m(y, \mu)]^{-1} (m(y, \mu)^T, \lambda^T \partial m(y, \mu))^T\}^T$ is continuously differentiable in a neighborhood of $(\mu_0^T, \lambda(\mu_0)^T)^T$ and its Jacobian at $(\mu_0^T, \lambda(\mu_0)^T)^T$ is non-singular;
- (2) there exists $L > 0$ and a neighborhood $\mathcal{N}$ of $(\mu_0^T, \lambda(\mu_0)^T)^T$ on which $\Pr(1 + \lambda^T m(Y, \mu) > L) = 1$ for all $(\lambda^T, \mu^T)^T \in \mathcal{N}$;
- (3) there exists a neighborhood $\mathcal{N}$ of μ_0 and functions p_j for $j = 0, 1, 2$ such that for F -almost all y and all $\mu \in \mathcal{N}$, the norm of m and all first and second order partial derivatives of m are bounded by p_j for $j = 0, 1, 2$ respectively with $\mathbb{E}[p_0(Y)^4]$, $\mathbb{E}[p_1(Y)^4]$ and $\mathbb{E}[p_2(y)^2]$ being finite.

In particular, $\sqrt{n}(\hat{\mu} - \mu_0)$ converges in distribution to a $N(0, J_{\text{EL}}^{-1} K_{\text{EL}} J_{\text{EL}}^{-1})$ distribution, where $K_{\text{EL}} = \text{Var}[\psi_\mu(Y, \mu_0)]$.

Remark 1. The quantity $\lambda'(\mu)$ is a well-defined quantity, but except for in a very few cases, it is only defined implicitly. By the implicit function theorem, however, $\lambda'(\mu) = S_1(\mu, \lambda)^{-1} S_2(\mu, \lambda)$ where $S_1(\mu, \lambda) = \mathbb{E}\{M(\mu)M(\mu)^T / [1 + \lambda^T M(\mu)]^2\}$ and $S_2(\mu, \lambda) = \mathbb{E}\{[M'(\mu) + M'(\mu)\lambda^T M(\mu) - M(\mu)\lambda^T M'(\mu)] / [1 + \lambda^T M(\mu)]^2\}$. Here $M(\mu)$ is short-hand for $m(Y, \mu)$ where $Y \sim F$. In Definition 1 we will use this to define estimators of the variance of $\hat{\mu}$.

Remark 2. Conditions (1) and (3) of Theorem 1 are more or less standard conditions ensuring that all functions involved are sufficiently regular. Condition (2) on the other hand warrants some discussion, as it excludes situations where $m(Y, \mu)$ has unbounded support. Because of this, condition (2) actually leads to several important situations not being covered.

The perhaps most obvious example is $m(y, \mu) = y - \mu$ when F is the standard normal distribution. The assumption is, however, necessary for the proof to go through. In practice, these considerations matter little as one can often truncate data to lie within some arbitrary large and fixed interval to avoid problems with unbounded support, ensuring that condition (2) holds.

Assume that μ_T solves the estimating equation $E[m(Y, \mu)] = 0$ for $Y \sim F$. Then $E\{m(Y, \mu)/[1 + 0]\} = 0$, and hence $\lambda(\mu_T) = 0$. Furthermore, by utilizing $\lambda(\mu_T) = 0$ direct computation shows that the derivative of $-E\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$ at μ_T is zero. Hence, $\mu_0 = \mu_T$ and multiple formulas simplifies significantly. We state this in the following corollary.

Corollary 1. *Assume the conditions of Theorem 1 hold true and that μ_0 solves the estimating equation $E[m(Y, \mu)] = 0$. Then the MEL estimator converges in probability to μ_0 , $\psi_\mu(y, \mu_0) = \lambda'(\mu_0)^T m(y, \mu)$ and $J_{EL} = \lambda'(\mu_0)^T E[\partial m(Y, \mu)]$ where $\lambda'(\mu_0) = -\{E[m(Y, \mu_0)m(Y, \mu_0)]^T\}^{-1} E[\partial m(Y, \mu_0)]$. In particular, $\sqrt{n}(\hat{\mu} - \mu_0) \xrightarrow{d} N(0, J_{EL}^{-1})$.*

In certain situations, it can be beneficial to use a plug-in estimator inside the empirical likelihood framework. Say that we, for instance, want to make inference about the variance σ^2 in a distribution. This can be achieved by using the empirical likelihood framework by using the estimating function $m(y, \mu, \sigma^2) = (y - \mu, (y - \mu)^2 - \sigma^2)$ and applying Theorem 1 or Corollary 1 in conjunction with the delta method. A perhaps more intuitive way, however, could be to replace μ with the empirical mean of the observations $\hat{\mu}$ and use $m(y, \sigma^2) = (y - \hat{\mu})^2 - \sigma^2$. This allows us to avoid adding a parameter for the mean in the distribution and only define estimating functions for the variance. The following result allows for such constructions. The proof is given in the supplementary material.

Theorem 2. Let $Y_1, \dots, Y_n$ be i.i.d. from some distribution F and let $m: \mathbb{R}^{d+r} \times \mathcal{M} \rightarrow \mathbb{R}^q$, where $\mathcal{M}$ is an open set, be a function continuously differentiable in the last s arguments. Assume furthermore, that $\hat{P} \in \mathbb{R}^r$ is some estimator satisfying $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_p(1/\sqrt{n})$ and that $\hat{\mu}(\hat{P})$ is the unique maximizer of the empirical likelihood function constructed using the estimating function $(y, \mu) \mapsto m(y, \hat{P}, \mu)$. Set $\Gamma(\mu, \lambda, P) = -E\{\log[1 + \lambda^T m(Y, P, \mu)]\}$. Furthermore, let $\lambda(\mu, P)$ be the root of $E\{m(Y, P, \mu)/[1 + \lambda^T m(Y, P, \mu)]\}$ for each fixed μ and P and let $\mu(P)$ be the unique maximizer of $\mu \mapsto \Gamma[\mu, \lambda(\mu, P), P]$ for each P . Then $\hat{\mu}(\hat{P})$ is consistent for $\mu_0 = \mu(P_0)$ where P_0 is the solution to $E[h(Y, P)] = 0$, and

$$\hat{\mu}(\hat{P}) = \hat{\mu}(P_0) + \mu'(P_0)(\hat{P} - P_0) + o_p(1/\sqrt{n}), \quad (2.8)$$

provided the following conditions hold true:

- (1) The function $(\mu, \lambda, P) \mapsto E\{(\partial_\mu \Gamma(\lambda, \mu, P)^T, \partial_\lambda \Gamma(\lambda, \mu, P)^T, h(Y, P)^T)\}^T$ is continuously differentiable in a neighborhood of $(\mu_0^T, \lambda(\mu_0, P_0)^T, P_0^T)^T$ with non-singular derivative at $(\mu_0^T, \lambda(\mu_0, P_0)^T, P_0^T)^T$;
- (2) there exists $L > 0$ and a neighborhood $\mathcal{N}$ of $\alpha_{P_0} = (\mu_0^T, \lambda(\mu_0, P_0)^T, P_0^T)^T$ on which $\Pr(1 + \lambda^T m(Y, P, \mu) > L) = 1$ for all $(\mu, \lambda, P) \in \mathcal{N}$;
- (3) there exists a neighborhood $\mathcal{N}$ of α_{P_0} and functions p_j for $j = 0, 1, 2$ such that for F -almost all y and all $\alpha \in \mathcal{N}$, the norm of m and all first and second order partial derivatives of m are bounded by p_j for $j = 0, 1, 2$ respectively, with $E[p_0(Y)^4]$, $E[p_1(Y)^4]$ and $E[p_2(Y)^2]$ finite;
- (4) there exists a neighborhood of P_0 on which the norm of $h(y, P)$ and all first order partial derivatives of h with respect to P are bounded by functions $q_0(y)$ and $q_1(y)$ respectively, with $E[q_0(Y)]$, $E[q_1(Y)^2] < \infty$.

Remark 3. Although $\mu'(P_0)$ is well-defined, it can be hard to compute in practice as $\mu(P)$ is only implicitly defined. The implicit function theorem can, however, be used to find an explicit expression for the quantity, similar as for $\lambda'(\mu)$ discussed in the remark following Theorem 1.

We have $\hat{P} - P_0 = n^{-1}V^{-1}\sum_{i=1}^n h(Y_i, P_0) + o_p(1/\sqrt{n})$, by the conditions above and Van der Vaart (2000, Thm. 5.21), . Furthermore, $\hat{\theta}(P_0)$ satisfies (2.7) with $m(y, \mu)$ replaced by $(y, \mu) \mapsto m(y, P_0, \mu)$ by Theorem 1. Hence, the central limit theorem guarantees that $\sqrt{n}[\hat{\mu}(P_0) - \mu_0(P_0)]$ and $\sqrt{n}(\hat{P} - P_0)$ and converges jointly in distribution to a central normal distribution with covariance matrix

$$\Sigma_\mu = \begin{pmatrix} J_{\text{EL}}^{-1} K_{\text{EL}} J_{\text{EL}}^{-1}, & J_{\text{EL}}^{-1} C_\mu (V^{-1})^T \\ V^{-1} C_\mu^T (J_{\text{EL}}^{-1})^T, & V^{-1} \text{Var}[h(Y, P_0)] V^{-1} \end{pmatrix}$$

where $V = E[\partial_{P_0} h(Y, P)]$ and C_μ is the covariance between $\psi_\mu(Y, \mu_0, P_0)$ and $h(Y, P_0)$ where $Y \sim F$. This together with (2.8) can be used to show that $\sqrt{n}[\hat{\mu}(\hat{P}) - \mu(P_0)]$ converges to a normal distribution with mean zero and variance $(I_s, \mu'(P_0)^T) \Sigma_\mu (I_s, \mu'(P_0)^T)^T$ where I_s is the identity matrix in $\mathbb{R}^{s \times s}$ and $(I_s, \mu'(P_0)^T)$ is a $s \times (s+r)$ matrix written as a block matrix.

Now that we have investigated the empirical likelihood function and the MEL estimator, we will turn our attention towards hybrid likelihood, the main topic of this article.

3. Hybrid likelihood

Hybrid likelihood was introduced in Hjort, McKeague, and Van Keilegom (2018) and is based on study of the hybrid likelihood function. As explained in the introduction, this is a combination of the likelihood function of the data, L_n , and the empirical likelihood function of a pre-specified parameter of interest. This pre-specified parameter is called the

control parameter, and we will denote it by μ . We will assume that the control parameter can be identified by the estimating equation $0 = E[m(Y, \mu)] = 0$. Furthermore, for a given parametric model f_θ with $\theta \in \Theta \subseteq \mathbb{R}^p$, we will let $\mu(\theta)$ refer to the value the control parameter takes in the model f_θ . With this notation, we define the hybrid likelihood function as $HL_n(\theta) = L_n(\theta)^{1-a} EL_n[\mu(\theta)]^a$, where $EL[\mu(\theta)]$ is the empirical likelihood function for μ evaluated at $\mu(\theta)$ and constructed using the estimating function m , and $a \in [0, 1)$ is some tuning parameter deciding how much weight should be put on the parametric and nonparametric part of HL_n . The quantity a is called the balance parameter.

The hybrid likelihood function can be interpreted as a more model robust version of the standard likelihood function, penalizing θ s for which $\mu(\theta)$ is far from solving $0 = n^{-1} \sum_{i=1}^n m(Y_i, \mu)$. If we, for instance, wish to fit a Weibull(λ, k) model (parameterized such that the cumulative distribution function takes the form $F_{\lambda, k}(y) = 1 - \exp(-(y/\lambda)^k)$) to some data while making sure that the mean is estimated robustly, we can use the hybrid likelihood function in place of the standard likelihood function. The mean satisfies $E[m(Y, \mu)] = 0$ with $m(y, \mu) = y - \mu$. Hence, the empirical likelihood function can be constructed using this estimating function. Furthermore, in a Weibull(λ, k)-distribution, the mean takes the form $\lambda\Gamma(1 + 1/k)$ where Γ is the Gamma function. Hence, in the hybrid likelihood function we would use $\mu(k, \lambda) = \lambda\Gamma(1 + 1/k)$. This procedure generalizes easily to higher moments, as we, in general, for the ℓ -th moment can use $m(y, \mu) = y^\ell - \mu$ and $\mu(k, \lambda) = \lambda^\ell\Gamma(1 + \ell/k)$. More generally, for a parametric model $f_\theta(y)$, the hybrid likelihood framework can ensure that estimation of $E[h(Y)]$ is extra robust by using $m(y, \mu) = h(y) - \mu$ and letting $\mu(\theta)$ be the value of $E[h(Y)]$ in the model f_θ . A choice which might be of interest in many applications is to use $m(y, \mu) = I(y \in [b, c)) - \mu$ for some pre-specified numbers

$-\infty \leq b < c \leq \infty$ and $\mu(\theta) = F_\theta(c) - F_\theta(b)$ where F_θ is the cumulative distribution function in the model f_θ . This would ensure that the fitted model estimates the probability of landing within $[b, c)$ extra robustly. This could be beneficial in situations where the fit in certain regions of the domain of y are more important than others.

In generalization of the above, we may construct a histogram controlled hybrid likelihood framework as follows. For bins $C_1, \dots, C_k$, associated with a histogram, use estimating equations $m_j(u, \mu_j) = I(y \in C_j) - \mu_j$ for $j = 1, \dots, k-1$. This corresponds to using the probabilities $\mu_j = \Pr(Y \in C_j)$ for $j = 1, \dots, k$ as control parameters, and hence fit into the hybrid likelihood framework by using $m(y, \mu) = (m_1(y, \mu), \dots, m_{k-1}(y, \mu))^T$ and $\mu = (\mu_1, \dots, \mu_{k-1})^T$ where $m_j(y, \mu) = I(y \in C_j) - \mu_j$ and $\mu_j(\theta) = \Pr_\theta(Y \in C_j)$ for $j = 1, \dots, k-1$. Here $\Pr_\theta(Y \in C_j)$ is the probability of Y landing within C_j when $Y \sim f_\theta$ for each j . By using these control parameters and corresponding estimating functions, the hybrid likelihood framework can be used to fit the parametric model f_θ in a way ensuring that the probability of landing within each cell $C_1, \dots, C_{k-1}$ is robustly estimated. This corresponds to ensuring that the histogram of the data with bins $C_1, \dots, C_k$ fits well with the fitted model. We illustrate this idea on a data set taken from Pearson (1902) containing life lengths of 141 men and women in Roman era Egypt. The data ranges from 1.5 to 96.0, and for this illustration we will fit a Weibull model to the life lengths. We will do this using the hybrid likelihood framework with $C_1 = [0, 20), \dots, C_5 = [80, 100)$ and use m as described above. In this case, $\mu_j(k, \lambda) = \exp[-(20(j-1)/\lambda)^k] - \exp[-(20j/\lambda)^k]$. The fit obtained using maximum likelihood is displayed in Figure 1, together with the fits resulting from maximizing the hybrid likelihood function for different values of the balance parameter. From the figure we notice that as a increases in value, the fitted density moves

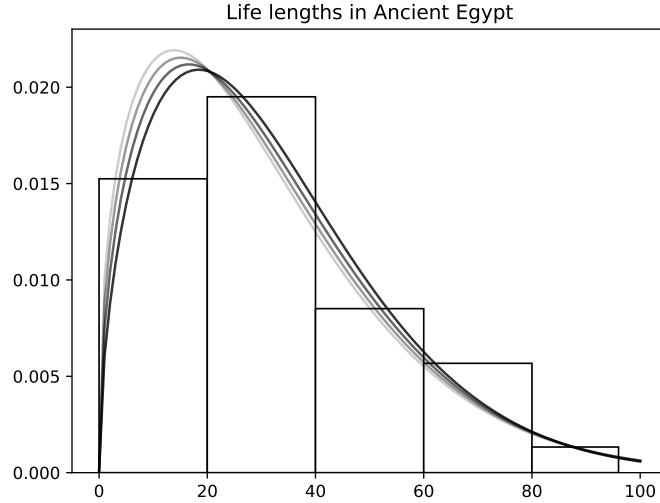


Figure 1: The figure shows a histogram of the life lengths for 141 men and women in Roman era Egypt together with the density of four fitted Weibull models. The models were all fitted by maximizing the hybrid likelihood function for $a = 0, 0.25, 0.5, 0.75, 0.99$. The value of a is indicated by the colour of the line with light values corresponding to low values of a .

from the maximum likelihood estimate towards a curve which agrees slightly more with the histogram of the data. This is a consequence of the empirical likelihood part of the hybrid likelihood function becoming more dominant as a increases.

Now that we have introduced the framework of hybrid likelihood, we will study the asymptotic behavior of the maximizer of the hybrid likelihood function. We will refer to this as the maximum hybrid likelihood estimator (the MHL estimator). Hjort, McKeague, and Van Keilegom (2018) consider the case when the parametric model f_θ is correctly specified. When this assumption holds true, the MHL estimator converges in probability to the true value of θ_0 and at the speed of $O(1/\sqrt{n})$. As a result, the MHL estimator of μ also converges to the true value of μ in the underlying distribution of the data. Classic results concerning

the empirical likelihood function in shrinking neighborhoods of the true value (e.g. Hjort, McKeague, and Van Keilegom (2009) and Molanes Lopez, Van Keilegom, and Veraverbeke (2009)) can therefore be used to derive limit results. When, on the other hand, the true underlying density is not assumed to be a member of the parametric family, there is no guarantee that the MHL estimator of the control parameter converges to the true solution of the estimating equation. Because of this, the results of e.g. Molanes Lopez, Van Keilegom, and Veraverbeke (2009) and Hjort, McKeague, and Van Keilegom (2009) are not sufficient. Instead, we will use the arguments of the previous section.

3.1 The hybrid likelihood function and its maximizer

Similarly as we did for the empirical likelihood function, we start by deriving a pointwise asymptotically equivalent expression for $n^{-1} \log \text{HL}_n(\theta)$. Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. following a distribution F . Furthermore, fix $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$ and $\mu_0 \in \mathbb{R}^s$ such that $E[m(Y, \mu_0)] = 0$ for $Y \sim F$. Assume f_θ for $\theta \in \Theta$ is some parametric model such that $E\{m[Y, \mu(\theta)]\} = 0$ for $Y \sim f_\theta$. To ease the notation, we will let $M_i(\theta)$ act as short-hand for $m[Y_i, \mu(\theta)]$ and $M(\theta)$ for $m[Y, \mu(\theta)]$ for a general $Y \sim F$. Lastly, fix $a \in [0, 1)$ and let $h_n: \mathbb{R}^p \rightarrow \mathbb{R}$ be the hybrid log-likelihood function, defined as $h_n(\theta) = (1 - a)\ell_n(\theta) + a \log \text{EL}_n[\mu(\theta)]$, where ℓ_n is the log-likelihood of the data. Applying (2.6) shows $h_n(\theta) = \sum_{i=1}^n \psi[Y_i, \mu(\theta)] + V_n[\mu(\theta)]^T S[\mu(\theta)]^{-1} V_n[\mu(\theta)] + o_{\text{pr}}(1)$, where $\psi(y, \theta) = a \log f_\theta(y) - (1 - a) \log\{1 + \lambda[\mu(\theta)]^T m[y, \mu(\theta)]\}$. In particular, the above shows that $n^{-1}h_n(\theta)$ is asymptotically equivalent to a mean. Similarly as for the MEL estimator we would expect that this implies that the MHL estimator is ‘almost’ an M-estimator. Once again this intuition is correct, as can be seen from the following theorem. A proof can be found in the supplementary material.

Theorem 3. Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^d \times \mathcal{M} \rightarrow \mathbb{R}^q$ a function continuously differentiable in the last s arguments. Assume that m 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 $\mu: \mathbb{R}^p \rightarrow \mathbb{R}^s$ a function 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 unique maximizer of h_n in Θ . Let $\lambda(\theta)$ be the solution to $0 = \text{E}\{M(\theta)/[1 + \lambda^T M(\theta)]\}$ for each θ . Assume furthermore that conditions (1-3) of Theorem 1 hold true at $(\lambda_0^T, \mu(\theta_0)^T)^T$, in addition to:

- (4) the function $\mu(\theta)$ is twice continuously differentiable in a neighborhood of θ_0 ;
- (5) there exists a neighborhood of θ_0 on which all components $u(y, \theta)$ and $\partial u / \partial \theta$ are bounded by some F -square integrable functions q_1 and q_2 respectively not depending on θ .

Lastly, let θ_0 be the unique maximizer of $\theta \mapsto (1 - a)\text{E}[\log f_\theta(Y)] - a\text{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$ in Θ . Then the MHL estimator $\hat{\theta}$ is consistent for θ_0 and

$$\hat{\theta} = \theta_0 + n^{-1}(J_a^*)^{-1} \sum_{i=1}^n \psi_\theta(Y_i, \theta_0) + o_{\text{pr}}(1/\sqrt{n})$$

where $\psi_\theta(y, \theta) = (1 - a)u(y, \theta) - a\mu'(\theta)^T \psi_\mu[y, \mu(\theta)]$ with ψ_μ as defined in Theorem 1 and $J_a^* = (1 - a)J + aH_{\theta_0}\text{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$ where J is the Fisher matrix of the model f_θ at θ_0 . In particular, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (J_a^*)^{-1}K_a^*(J_a^*)^{-1})$, where $K_a^* = \text{Var}[\psi(Y, \theta_0)]$.

If the true underlying distribution is a member of the parametric family fit to the data, there is a vector $\theta_0 \in \mathbb{R}$ such that the distribution F has density function f_{θ_0} . In this case, $\mu(\theta_0)$ solves the estimating equation $\text{E}\{m[Y, \mu(\theta_0)]\} = 0$ and hence maximizes $-\text{E}\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$ by the arguments preceding Corollary 1. As a result, both

$E[\log f_\theta(Y)]$ and $-E\{\log[1 + \lambda(\theta)^T M(\theta)]\}$ are maximized by θ_0 , making this vector the maximizer of $\theta \mapsto (1 - a)E[\log f_\theta(Y)] - aE\{\log[1 + \lambda(\theta)^T M(\theta)]\}$. Furthermore,

$$\psi_\theta(y, \theta_0) = (1 - a)u(y, \theta_0) - aE[\partial m(Y, \theta_0)]E\{m(Y, \theta_0)m(Y, \theta_0)^T\}^{-1}m[y, \mu(\theta_0)],$$

by Corollary 1. Plugging this into Theorem 3 shows $K_a^* = (1 - a)^2 J + a^2 \xi_0^T W^{-1} \xi_0 - a(1 - a)(C W^{-1} \xi_0 + \xi_0^T W^{-1} C^T)$, where J is the Fisher matrix, W is the covariance matrix of $m(Y, \theta_0)$, C is the covariance between $u(Y, \theta_0)$ and $m(Y, \mu_0)$, and $\xi_0 = E[\partial m(Y, \theta_0)]$. In addition, $J_a^* = (1 - a)J + a \xi_0^T W^{-1} \xi_0$ by the chain rule and Corollary 1. Hence, if the model is correctly specified, Theorem 3 simplifies to the corresponding results given in Hjort, McKeague, and Van Keilegom (2018). Because of this, we believe Theorem 3 to be a proper extension of the theorems given in Hjort, McKeague, and Van Keilegom (2018) and a tool for making model robust inference in the hybrid likelihood setting.

Also in the hybrid likelihood case, we would like to allow for plug-in parameters. The following result gives the asymptotic behavior of $\hat{\theta}$ also in this case. The proof is given in the supplementary material.

Theorem 4. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^{d+r+s} \rightarrow \mathbb{R}^q$ a function continuously differentiable in the last s arguments. Assume furthermore, that $\hat{P} \in \mathbb{R}^r$ is some estimator with the property $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_p(1/\sqrt{n})$ and that that $(y, \mu) \mapsto m(y, \hat{P}, \mu)$ is used as estimating function 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 $\mu: \mathbb{R}^{p+r} \rightarrow \mathbb{R}^s$ a function such that $E\{m[Y, P, \mu(\theta, P)]\} = 0$ when $Y \sim f_\theta$. Fix $a \in [0, 1)$ and assume that $\hat{\theta}(\hat{P})$ is the unique maximizer of h_n . Let $\lambda(\theta, P)$ be the solution to $0 = E\{M(\theta, P)/[1 + \lambda^T M(P, \theta)]\}$ for each θ where $M(\theta, P)$ is short-hand for $m[Y, P, \mu(\theta, P)]$ with $Y \sim F$, furthermore, assume that the maximizer of*

$\theta \mapsto (1-a)\mathbb{E}[\log f_\theta(Y)] - a\mathbb{E}\{\log[1 + \lambda(\theta, P)^T M(P, \theta)]\}$ is unique and denote it by $\theta(P)$ for each P . Then the MHL estimator $\hat{\theta}(\hat{P})$ is consistent for $\theta_0 = \theta(P_0)$ and

$$\hat{\theta}(\hat{P}) = \hat{\theta}(P_0) + \theta'(P_0)(\hat{P} - P_0) + o_{\text{pr}}(1/\sqrt{n})$$

provided conditions (4) and (5) of Theorem 3 hold true in addition to conditions (1-4) of Theorem 2 at $\alpha_{p,0} = (\mu(\theta_0, P_0)^T, \lambda(\theta_0, P_0)^T, P_0^T)^T$.

Remark 4. Just as $\mu'(P_0)$ was only implicitly defined in Theorem 2, the same is true for $\theta'(P_0)$. Again, however, we can find an expression for the quantity by applying the implicit function theorem similarly as discussed in the remark following Theorem 1.

Similarly as for Theorem 2, the above result can be used to show that $\sqrt{n}[\hat{\theta}(\hat{P}) - \theta(P_0)]$ has a limiting normal distribution with mean zero and covariance matrix equal to

$$\Sigma_\theta = \begin{pmatrix} (J_a^*)^{-1} K_a^* (J_a^*)^{-1}, & (J_a^*)^{-1} C_\theta (V^{-1})^T \\ V^{-1} C_\theta^T [(J_a^*)^{-1}]^T, & V^{-1} \text{Var}[h(Y, P_0)] V^{-1} \end{pmatrix}$$

where $V = \mathbb{E}[\partial_{P_0} h(Y, P)]$ and C_θ is the covariance between $\psi_\theta(Y, \theta(P_0), P_0)$ and $h(Y, P_0)$ where $Y \sim F$ and ψ_θ^P is the function ψ_θ defined in Theorem 3 when $m(y, \mu)$ is replaced by $m(y, P, \mu)$ for each fixed P . This together with (2.8) ensures that $\sqrt{n}[\hat{\theta}(\hat{P}) - \theta(P_0)]$ converges to a normal distribution with mean zero and variance $(I_p^T, \mu'(P_0)^T) \Sigma_\theta (I_p^T, \mu'(P_0)^T)^T$.

3.2 The hybrid divergence

By Theorem 3, the maximum hybrid likelihood estimator converges in probability to the maximizer of $(1-a)\mathbb{E}[\log f_\theta(Y)] - a\mathbb{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$. This function is useful for showing consistency of the MHL estimator, but the expression is somewhat hard to interpret, and it is not clear from the formula why we should want to maximize it. In this section, we will

take a closer look at $(1-a)\mathbb{E}[\log f_\theta(Y)] - a\mathbb{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$ and show that maximizing this expression is equivalent to minimizing a divergence taking both model fit and precise estimation of the control parameter into account.

We start by considering $-\mathbb{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$. Using the Lagrange method of multipliers, one can show that for each fixed μ , the expression is the solution to the following optimization problem:

$$\max_{w: \mathbb{R}^d \rightarrow \mathbb{R}} \{ \mathbb{E}[\log w(Y)] : \mathbb{E}[w(Y)M(\mu)] = 0 \text{ and } \mathbb{E}[w(Y)] = 1 \}. \quad (3.9)$$

Assume now that the true distribution of Y has a density (pdf) or point mass function (pmf) g and let $h(y) = w(y)g(y)$. Then solving (3.9) is equivalent to minimizing $\text{KL}(g, h)$ with respect to h under the condition that h is a pmf/pdf and $\mathbb{E}_h[M(\mu)] = 0$. Here $\text{KL}(g, h)$ is the Kullback-Leibler divergence from h to g defined as $\mathbb{E}\{\log[h(Y)/g(Y)]\}$. Since subtracting a constant does not change the extrema of an expression, the above shows that the maximizer of $(1-a)\mathbb{E}[\log f_\theta(Y)] - a\mathbb{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$, i.e. the limit of the MHL estimator, is also the minimizer of $\theta \mapsto d_a(g, f_\theta)$ where

$$d_a(g, h) = (1-a)\text{KL}(g, h) + a \min_h \{ \text{KL}(g, h) : \mathbb{E}_h\{M[\mu(h)]\} = 0 \}. \quad (3.10)$$

with $\mu(h)$ being the solution to the estimating equation $\mathbb{E}_h[M(\mu)] = 0$. The Kullback-Leibler divergence is a divergence. Hence, for every $h \neq g$, the first term in $d_a(g, h)$ is positive. Since the second term is non-negative by the previous arguments, $d_a(g, h) > 0$ for all $h \neq g$. Furthermore, $\text{KL}(g, g) = 0$ and $\mathbb{E}_g\{M[\mu(g)]\} = 0$. Hence, $d_a(g, h) = 0$ only when $h = g$. Because of this, d_a is a proper divergence, measuring the “distance” from the true density g to h .

The first term in (3.10) is small when the overall model fit is good, i.e. when h is close to g overall. The latter term, on the other hand, does not care about overall model fit, but

is small when $\mu(h)$ is close to $\mu(g)$. Hence, d_a measures the distance from g to h by a convex combination of the lack of overall model fit and error involved when using $\mu(h)$ in place of $\mu(g)$. By Theorem 3, the MHL estimator converges to θ_0 such that f_{θ_0} minimizes d_a in the class f_θ for $\theta \in \Theta$. Hence, the limit of the MHL estimator is the value of θ making f_θ as close as possible to the true density g , and distance is measured as a trade-off between lack of model fit and bias of the control parameter, and how much weight should be put on each term in the divergence is determined by the value of the tuning parameter.

Figure 2 illustrates the effect of the balance parameter on the divergence d_a and the resulting minimizer. We have created a plot of the limit of the MHL estimator for a selection of a values in a specific situation. The control parameter used is the probability of not exceeding one and the model fitted is a Weibull model with scale λ and shape k . The true underlying distribution is a Gamma distribution with shape 10 and scale 5. From the figure, we see that for small values of a the limit of the MHL estimator is close to (2.23, 3.33). This is the minimizer of the Kullback-Leibler divergence from the true distribution to the model. In Figure 2 we have plotted $\mu_0(a)$, the limit of the MHL estimator of the control parameter μ as a function of a . From the figure, we see that low values of a ensure $\mu_0(a)$ close to 0.504. This is the limit of the ML estimator of μ . Larger values of the balance parameter on the other hand ensure $\mu(a)$ being closer to 0.542, the true probability of a Gamma(10, 5) variable not exceeding one.

Remark 5. When $a = 1$, d_a is no longer a divergence as all distributions satisfying $E_g\{m[Y, \mu_0]\} = 0$ will maximize $d_0(f, g)$. Hence, $d_0(f, g)$ can be equal to 0 even when $g \neq f$, violating one of the criteria for being a divergence. As the hybrid likelihood framework is defined for $a \in [0, 1)$ only, this is not a problem in our context.

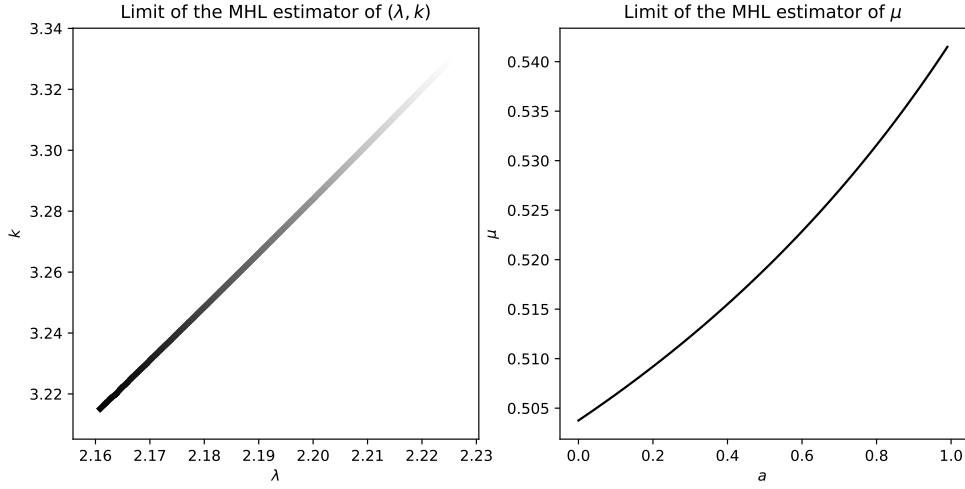


Figure 2: The figure to the left shows the limit of the MHL estimator of the parameters (λ, k) for different values of a . The balance parameter is indicated by the color of the line. Light colors corresponds to small values of a , while darker colors indicate a close to one. In the plot to the right we have displayed the limit of the MHL estimator of the control parameter μ as a function of the balance parameter a .

4. The balance parameter

To use the hybrid likelihood theory in practice, we need to choose a control parameter, μ , and set the value of the balance parameter, a . The choice of the former should reflect what parameters are of particular interest in the specific application. In some situations, the same might be true for a . Usually, however, how much weight should be put on the parametric and nonparametric part of HL_n is dependent on the quality of the model fit and estimator of the control parameter. In this section, we will propose a way to choose a taking these considerations into account.

4.1 FIC for models fitted with hybrid likelihood

In certain applications, estimation of a single parameter is the main point of the analysis. In such cases, models resulting in precise estimators of this parameter of interest are more desirable than those achieving a good overall fit. This is the main idea behind the focused information criterion.

The focused information criterion (see Claeskens and Hjort (2008) for an introduction) ranks models by the quality of their estimate of the focus parameter. The information criterion is created to choose between different parametric models, but the estimators defined can in principle be used for other purposes. We propose to use the information criterion to choose the optimal value for the balance parameter a used in construction of the hybrid likelihood function. This idea was briefly visited in Hjort, McKeague, and Van Keilegom (2018), where the authors derive formulas for a focused information criterion for models in shrinking neighborhoods of the true underlying density. More precisely, Hjort, McKeague, and Van Keilegom (2018) formulate an FIC for models on the form $f(y, \theta)$ when the true

underlying distribution of the data is $f_{\text{true}}(y, \theta, \gamma_0 + \delta/\sqrt{n})$ and $f_{\text{true}}(y, \theta, \gamma_0) = f(y, \theta)$. Their version of the information criterion therefore requires the model to be asymptotically correctly specified, and to be able to compute their FIC one needs to define a wider model $f_{\text{true}}(y, \theta, \gamma)$, which is then assumed to be correct. This is not too problematic in e.g. the context of variable selection, but can be quite restrictive for more general use. We will therefore instead create a fully model agnostic FIC, based upon the formulas and results of Jullum and Hjort (2017) with important modifications in Dæhlen, Hjort, and Hobæk Haff (2023). These articles find asymptotically unbiased approximations to the mean squared error of estimators of the focus parameter in a large number of cases. Their results can be used to find estimators of the mean square error of maximum hybrid likelihood estimators of focus parameters. In the following theorem, we state a version of the results of Jullum and Hjort (2017) and Dæhlen, Hjort, and Hobæk Haff (2023) valid in the hybrid likelihood context.

Theorem 5. *Let $Y_1, \dots, Y_n$ be i.i.d. from some distribution F and assume we fit a parametric model f_θ for $\theta \in \Theta \subseteq \mathbb{R}^p$ using maximum hybrid likelihood with some fixed control parameter μ and balance parameter a . Let $\xi \in \mathbb{R}$ be some focus parameter and $\hat{\xi}$ a consistent estimator of ξ satisfying*

$$\hat{\xi} = \xi + n^{-1} \sum_{i=1}^n \phi(Y_i, \xi) + \epsilon_n \quad (4.11)$$

for some function $\phi: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\epsilon_n = O_{\text{pr}}(1/n)$. Assume further that all conditions of Theorem 3 hold true and that there is a continuously differentiable function $g: \mathbb{R}^p \rightarrow \mathbb{R}$ such that $\xi = g(\theta)$ in the model f_θ . Then $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$ is only $o(1/n)$ away from the true MSE of $g(\hat{\theta})$ in expected value where $\hat{b} = g(\hat{\theta}) - \hat{\xi}$, $c = \lim_{n \rightarrow \infty} nE(\epsilon_n)$, $\kappa = \text{Var}[\nabla g(\theta_0)^T (J_a^*)^{-1} \psi_\theta(Y, \theta_0) - \phi(Y, \xi)]$ and $\tau = (J_a^*)^{-1} K_a^* (J_a^*)^{-1}$ provided the assumptions

of Jullum and Hjort (2017) and Dæhlen, Hjort, and Hobæk Haff (2023) hold.

The crucial part of Theorem 5 is the fact that $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$ is only $o(1/n)$ away from the true MSE of $g(\hat{\theta})$ in expected value. By Theorem 3 the variance of $\hat{\theta}$ is of order $O(1/n)$. Hence, we need to approximate the MSE to a smaller order than $O(1/n)$. Otherwise the variance will be on the same scale as the error of the estimator of the MSE and might be greatly downplayed in the approximation.

In practice, the quantities c , κ and τ are usually unknown. To approximate the MSE of the MHL estimator using Theorem 5, we therefore need to first define estimators of these parameters and approximations to the expression $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$. This is provided in the following definition.

Definition 1. Let all quantities be as defined in Theorem 5. We define the following approximation to the MSE of $g(\hat{\theta})$,

$$\text{HFIC} = \max \left\{ \hat{b}^2 - n^{-1}(2\hat{b}\hat{c} + \hat{\kappa}) \right\} + n^{-1}\hat{\tau},$$

where $\hat{c}$ is either of the estimators given in Dæhlen, Hjort, and Hobæk Haff (2023), $\hat{\tau} = \nabla g(\hat{\theta})^T (\hat{J}_a^*)^{-1} \hat{K}_a^* (\hat{J}_a^*)^{-1} \nabla g(\hat{\theta})$ and $\hat{\kappa}$ is given by $\hat{\tau} + n^{-1} \sum_{i=1}^n \phi(Y_i, \hat{\xi})^2 - 2 \nabla g(\hat{\theta})^T (\hat{J}_a^*)^{-1} \hat{V}$, with $\hat{J}_a^* = n^{-1} H h_n(\hat{\theta})$, $\hat{K}_a^* = n^{-1} \sum_{i=1}^n \hat{\psi}_\theta(Y_i, \hat{\theta}) \hat{\psi}_\theta(Y_i, \hat{\theta})^T$ and $\hat{V} = n^{-1} \sum_{i=1}^n \hat{\psi}_\theta(Y_i, \hat{\theta}) \phi(Y_i, \hat{\xi})$. Here $\hat{\psi}_\theta(y, \theta)$ is equal to $\psi_\theta(y, \theta)$ with $\lambda(\mu)$ replaced by $\hat{\lambda}$, the solution to $0 = \sum_{i=1}^n M_i[\mu(\theta)] / (1 + \lambda^T M_i[\mu(\theta)])$ and $\lambda'[\mu(\theta)]$ by $-(S_n^1)^{-1} S_n^2$ where $S_n^1 = n^{-1} \sum_{i=1}^n \hat{M}_i \hat{M}_i^T / (1 + \hat{\lambda}^T \hat{M}_i)^2$ and $S_n^2 = n^{-1} \sum_{i=1}^n (\hat{\lambda}^T \hat{M}_i + \hat{M}_i' \hat{\lambda}^T \hat{M}_i - \hat{M}_i \hat{\lambda}^T \hat{M}_i') / (1 + \hat{\lambda}^T \hat{M}_i)^2$ with $\hat{M}_i = m[Y_i, \mu(\hat{\theta})]$ and $\hat{M}_i' = \partial_{\mu(\hat{\theta})} m(Y_i, \mu)$ for each $i = 1, \dots, n$.

Remark 6. In most applications the control and focus parameter will be the same, but there is no technical reason for this, and, if useful, μ and ψ can be chosen to be different parameters. See Section 6.

The quantities $\hat{J}_a^*$, $\hat{K}_a^*$, $\hat{V}$ and $n^{-1} \sum_{i=1}^n \phi(Y_i, \hat{\xi})^2$ are consistent for J_a^* , K_a^* , $\text{Var}[\phi(Y, \xi)]$ and $E[\psi_\theta(Y, \theta_0)\phi(Y)]$ respectively under weak conditions. See e.g. Lemma 16 of Dæhlen, Hjort, and Hobæk Haff (2023) for a set of sufficient conditions. When this is the case, $\hat{\tau}$ and $\hat{\kappa}$ converge in probability towards τ and κ by the continuous mapping theorem and Theorem 3, making HFIC approximately unbiased for the mean squared error of the MHL estimator. Hence, to choose the balance parameter in a way minimizing the asymptotic mean squared error, we can simply compute the MHL estimator and corresponding HFIC score for some values of a in a grid of points between zero and one. Afterwards, the balance parameter resulting in the smallest estimated mean squared error of the focus parameter can be chosen.

The above HFIC is only valid when plug-in estimators are not present. Thanks to Theorem 4, however, the criterion can also be defined in this context. We give the formulas below.

Definition 2. Let all quantities be defined as in Theorem 4 and assume the conditions of this theorem hold. Furthermore, let $\xi \in \mathbb{R}$ be some focus parameter taking the value $g(\theta)$ in the parametric model and $\hat{\xi}$ a consistent estimator of ξ satisfying (4.11) for some ϕ . Define the following function

$$\psi(y, \theta, \lambda, P) = \begin{pmatrix} m(y, P, \mu)/(1 + \lambda^T m(y, P, \mu)) \\ (1 - a)u(y, \theta) - a\lambda^T \partial_\theta m[y, P, \mu(\theta)]/(1 + \lambda^T m[y, P, \mu(\theta)]) \\ h(y, P) \end{pmatrix},$$

and let $J\psi(y, \theta, \lambda, P)$ denote its Jacobian matrix with respect to all parameters but y . Let $\hat{A} = n^{-1} \sum_{i=1}^n J\psi(Y_i, \hat{\theta}, \hat{\lambda}, \hat{P})$ and $\hat{B} = n^{-1} \sum_{i=1}^n \psi(Y_i, \hat{\theta}, \hat{\lambda}, \hat{P}) \psi(Y_i, \hat{\theta}, \hat{\lambda}, \hat{P})^T$ and $\hat{C} = \hat{A}^{-1} \hat{B} (\hat{A}^{-1})^T$. With this we define the HFIC as in Definition 1 with $\hat{b} = g[\hat{\theta}(\hat{P})] - \hat{\xi}$, $\hat{c}$ equal

to either of the estimators given in Dæhlen, Hjort, and Hobæk Haff (2023) and

$$\hat{\tau} = (\nabla g[\hat{\theta}(\hat{P})], 0)^T \hat{C} (\nabla g[\hat{\theta}(\hat{P})]^T, 0)^T \text{ and}$$

$$\hat{\kappa} = \hat{\tau}_P + n^{-1} \sum_{i=1}^n \phi(Y_i, \hat{\xi})^2 - 2(\nabla_{\hat{\theta}(\hat{P})} g(\theta)^T, 0) \hat{C} n^{-1} \sum_{i=1}^n \psi[Y_i, \hat{\theta}(\hat{P}), \hat{\lambda}, \hat{P}] \phi(Y_i, \hat{\xi})$$

where $\hat{\lambda}$ is the root of $\sum_{i=1}^n m[Y_i, \hat{P}, \hat{\theta}(\hat{P})] / \{1 + \lambda^T m[Y_i, \hat{P}, \hat{\theta}(\hat{P})]\}$ and the matrix $(\nabla g[\hat{\theta}(\hat{P})], 0)^T$ is a block matrix of dimension $1 \times (p + q + r)$ with zeros in the latter $q \times r$ components.

By Lemma 5 in the supplementary material and the results of Jullum and Hjort (2017) and Dæhlen, Hjort, and Hobæk Haff (2023) the focused information criterion is only $O(1/n)$ away from the true MSE of $g[\hat{\theta}(\hat{P})]$ in expected value. In Section 6 we will use the criterion to choose the balance parameter in hybrid models fitted to the number of battle deaths in interstate wars.

5. Regression

So far, we have only worked with i.i.d. data points. All of the results derived are, however, immediately applicable to regression settings if we are willing to treat the covariates as random variables. We will now illustrate how this can be done in the context of linear regression. These ideas were briefly visited in Hjort, McKeague, and Van Keilegom (2018).

Let $Y \in \mathbb{R}$ be the response variable and X the regressor, and assume that $Z_i = (Y_i, X_i)$ for $i = 1, \dots, n$ are i.i.d. from some distribution F . Suppose furthermore that we wish to fit a linear regression model to these data. This corresponds to maximizing the function $\ell_n(\beta_0, \beta) = -\sum_{i=1}^n (Y_i - \beta_0 - \beta^T X_i)^2$. We will use ℓ_n as the likelihood function in the hybrid likelihood framework.

Suppose we are unsure if a subset of the covariates is relevant for predicting the outcome of Y . In such a context, it could be beneficial to fit the model in a way giving additional

weight to the components we know are important. This can be achieved with the hybrid likelihood framework. Let $\mathcal{S}$ be the set of indices corresponding to the predictors we believe are relevant for prediction, and for a general vector v let $v_{\mathcal{S}}$ denote the components of v with indices in $\mathcal{S}$. Then, the solution to $E[m(Y, X, \beta)] = 0$ where $m(y, x, \beta_0, \beta_{\mathcal{S}}) = (y - \beta_0 - \beta_{\mathcal{S}}^T x_{\mathcal{S}})(1, x_{\mathcal{S}}^T)^T$, is the minimizer of $E[(Y - \beta_0 - \beta_{\mathcal{S}}^T X_{\mathcal{S}})^2]$. Hence, m can be used to construct the empirical likelihood function for $(\beta_0, \beta_{\mathcal{S}}^T)^T$, the coefficients in the best linear fit when regressing Y onto $X_{\mathcal{S}}$.

Direct computations show that for the smaller regression of Y onto $X_{\mathcal{S}}$, $(\beta_0(\beta), \beta_{\mathcal{S}}(\beta)^T)^T = [E(X_{\mathcal{S}+1} X_{\mathcal{S}+1}^T)]^{-1} E(X_{+1, \mathcal{S}} X_{+1}) \beta = (\Sigma_{\mathcal{S}+1, \mathcal{S}+1} + \mu_{+1} \mu_{+1}^T)^{-1} (\Sigma_{\mathcal{S}+1, :} + \mu_{\mathcal{S}+1} \mu_{+1}^T) \beta$, where for any vector v , v_{+1} indicates the vector $(1, v^T)$, $\Sigma_{\mathcal{S}+1, \mathcal{S}+1}$ is the covariance matrix of $X_{\mathcal{S}+1}$ and $\Sigma_{\mathcal{S}+1, :}$ is the covariance between $X_{\mathcal{S}+1}$ and the full component vector X_{+1} . Hence, for fixed a we can define the hybrid log-likelihood function as $h_n(\beta_0, \beta, \mu, \Sigma) = (1 - a) \ell_n(\beta_0, \beta) + a \log \text{EL}_n[\beta_{\mathcal{S}}(\beta, \mu, \Sigma)]$, where EL_n is the empirical likelihood function of $\beta_{\mathcal{S}}$ constructed with the estimating function m . Maximizing h_n for $a > 0$ fits a linear regression model in all covariates, while giving additional attention to the fit of Y regressed onto $X_{\mathcal{S}}$.

In practice, estimation of μ and Σ is rarely of interest. Because of this, it can be beneficial to replace them by $\hat{\mu}$, the empirical mean, and $\hat{\Sigma}$, the empirical covariance, respectively rather than optimizing with respect to these variables. This corresponds to using the hybrid likelihood framework with plug-in estimators for μ and Σ . Indeed, let $h(y, x, \mu, \Sigma) = (x^T - \mu^T, (x - \mu)(x - \mu)^T - \Sigma)^T$, then $(\hat{\mu}, \hat{\Sigma})$ satisfies $n^{-1} \sum_{i=1}^n h(Z_i, \hat{\mu}, \hat{\Sigma}) = o_p(1/\sqrt{n})$. Hence, inference of the hybrid likelihood estimator of β using plug-in estimators for μ and Σ can be made using Theorem 4.

Remark 7. It is worth noting that when $\mathcal{S}$ consists of all indices, both ℓ_n and $\log \text{EL}_n$ are

maximized at $\hat{\beta}_{\text{OLS}}$, the ordinary least square estimator of β . Hence, $\hat{\beta}_{\text{OLS}}$ maximizes h_n for all values of a and is therefore equal to the MHL estimator regardless of the choice of balance parameter. This holds true, whether the parametric model used is specified correctly or not. When $\mathcal{S} \neq \{1, \dots, d\}$, however, $\log \text{EL}_n$ and ℓ_n do not have the same maxima and hence the MHL estimator is not constant as a function of a .

6. Application: Battle deaths in two hundred years of interstate wars

In his *The Better Angels of Our Nature*, Pinker (2011) broadly argues that the world is changing for the better, and that the overall violence is on the decline. This claim quickly stirred up heated discussions in both the media and academic community, see e.g. Gray (2015), Cirillo and Taleb (2016a,b). We will now illustrate how the hybrid likelihood framework can be applied to investigate certain aspects of these claims.

We extracted the number of battle deaths in the 95 most recent (and concluded) interstate wars from the Correlates of War data set (Sarkees and Wayman (2010)). Cunen, Hjort, and Nygård (2020) argue that the Korean War marks a change in the number of battle deaths, and hence, based on their analysis, we divided the data into two: older wars (the Korean war and all prior conflicts) and newer wars (the remaining data points). Our goal is to make inference about the difference between the number of battle deaths in newer and older wars. In Dæhlen, Hjort, and Hobæk Haff (2023) the authors investigate which models are suited to answer this question. They conclude that a shifted log-normal model outperforms the other options, including the Dagum model which is used in Cunen, Hjort, and Nygård (2020). Because of this, we will use shifted log-normal models for both data sets.

We start by choosing a control parameter. Since $\mu \mapsto I(Y \leq \mu) - 1/2$ is discontinuous in μ and Theorem 3 requires smooth estimating functions, we cannot control for the median directly. As a compromise we chose $m(y, \mu) = I(y \leq q_j) - \mu$ for $j = 1, 2$, where q_1 and q_2 are the empirical median number of battle deaths in older and newer wars, respectively. The theorems of this article can then be applied if q_j are treated as fixed numbers and kept constant as $n \rightarrow \infty$.

We chose separate balance parameters for each model. In each case, this was done by computing HFIC for each value of a in a grid of points between 0 and 1. The median was used as the focus parameter for both models, and the empirical median in each data set as the consistent estimator $\hat{\xi}$ in Definition 1. Since the median satisfies (4.11) with $\phi(x, \xi) = (I(Y \leq \xi) - 0.5)/f(\xi)$, where f is the true density of the data, this is a valid choice. We estimated all quantities in HFIC as given in Definition 1, with $f(\mu)$ replaced by non-parametric kernel density estimates. For more details see Dæhlen, Hjort, and Hobæk Haff (2023), in particular their Appendix A. The results are shown in Figure 3. We see that neither minima are attained at 0 or 1, i.e. a genuine hybrid method is chosen for both data sets. The optimal balance values are $a = 0.48$ for the older wars and $a = 0.17$ for the newer conflicts.

With all parameters set, we fitted the models to each data set using maximum hybrid likelihood. The resulting MHL estimators of the median for the two data sets were 8192 and 4750 for the older and newer wars respectively. In both cases this is closer to the empirical median in the data sets (11375 and 5240 respectively) than the maximum likelihood estimates (6799 and 4733). In addition, we estimated the standard deviation of the MHL estimator of the medians in the two populations. This was done by applying Theorem 3 in conjunction

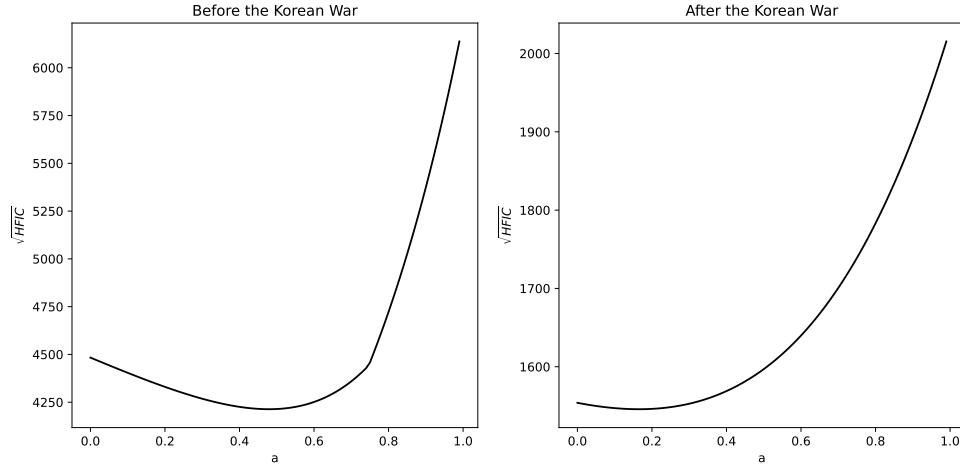


Figure 3: The figure displays root-HFIC scores as functions of a (the balance parameter) for older and newer wars, in the left and right plots respectively.

with the delta method and the estimators defined in Definition 1. Estimated standard deviations were found to be 3488 and 1546 for the older and newer wars respectively. This is only marginally larger than the estimated standard deviations for the ML estimator for the older wars and more or less equal to the that of the the ML estimator for the newer wars (2746 and 1554 respectively). Hence, the added robustness of hybrid likelihood comes at a low price in this case.

We can now make inference about the difference in number of battle deaths between newer and older wars using hybrid likelihood. The MHL estimator of this parameter was found to be 3441, which, relatively speaking, is a lot larger than the corresponding ML estimator of 2066. We can compute variances for this quantity as well by applying the delta method and Theorem 3. We found the standard deviation for the MHL estimator to be 3815, while the ML estimator had a slightly smaller standard deviation of 3155. With these numbers we can construct confidence intervals. A 95% confidence interval obtained with

hybrid likelihood was computed to be $[-4036, 10919]$. With maximum likelihood, we get a slightly smaller interval shifted to the left $([-4135, 8232])$. In addition, we can construct hypothesis tests for testing if the median number of battle deaths is the same for older and newer wars, against the alternative that violence has gone down. The p-value for testing this hypothesis using hybrid likelihood was 0.18, while the test based on maximum likelihood had a p-value of 0.26. These findings are in line with e.g. Clauset (2018).

The above analysis can be repeated for other quantiles than the median. To illustrate this, we performed similar computations as above for the difference between the upper quartiles in the two distribution. We leave out the details, but the p-value for testing if the difference is equal to zero was 0.03 for the hybrid likelihood framework. This is significant on a 5%-level. The same is not true for the classical maximum likelihood method, which had a p-value of 0.06. Hence, the added robustness from the hybrid likelihood framework allows us to draw a more positive conclusion about the world. The difference is even clearer at higher quantile levels. Indeed, for each $p \in [0.50, 0.85]$ we tested if the difference between the p -quantile for newer and older wars was zero against the alternative that violence has decreased. For almost all $p > 0.65$ the null hypothesis is rejected. Hence, we can conclude that the larger wars have indeed become less deadly, which is consistent with the findings of Cunen, Hjort, and Nygård (2020).

The hybrid likelihood machinery may of course also be applied to other foci than the quantile difference, and with other control parameters. In particular, the two cumulative distribution functions, say $F_L(y)$ and $F_R(y)$, for intervals of y of interest. The resulting estimates balance the parametric and nonparametric contributions. The difference is significantly non-zero for certain intervals of y .

It is worth noting that neither the ML estimator nor the MHL estimator are consistent for the true value of the parameter if the model is misspecified. The MHL estimator is, however, likely to aim for a quantity closer to the actual value of the parameter we are trying to estimate because of the added robustness introduced by the empirical likelihood part of HL_n .

7. Concluding remarks

We have worked and significantly extended the hybrid likelihood framework of Hjort, McKee, and Van Keilegom (2018). Firstly, we derived alternative characterization of the empirical and hybrid likelihood functions. We also proved that the both the MEL and the MHL estimator are consistent for well-defined quantities and are asymptotically normal after proper scaling and centering. Our theory also allowed for the presence of plug-in parameters. Afterwards, we used the results to define estimators for the variance and MSE of the MHL estimator. We used this to define a focused information criterion and proposed to set the balance parameter to the minimizer of this criterion. Lastly, we showed how the hybrid likelihood framework could be applied to regression settings and illustrated the developed theory in an application. We will now briefly go through some limitations of our work and possible further research topics.

The perhaps most pressing issue with the theory developed in this article was discussed in Remark 2 following Theorem 1. If m is the estimating function used in the construction of the empirical likelihood function, we require $m(y, \mu)$ to have bounded support for each fixed μ in a neighborhood of μ_0 . Theoretically, this assumption is quite problematic. In practice, however, it is less of a problem, as approximations using truncated distributions

often give very good results. Take for instance, the situation where $m(y, \mu) = y - \mu$ and the data is normally distributed. Replacing each data point Y by $\text{sign}(Y) \max\{K, |Y|\}$ for some large number K makes all theorems go through while making virtually no difference to the analysis. Nevertheless, formal proofs are currently not present, and further research is needed for such procedures to be theoretically justified.

Another avenue for further research is to consider the case of non-smooth estimating functions m . The present theorems require m to be differentiable. This excludes the case of quantiles as $m(y, \mu) = I(y \leq \mu) - p$ for $p \in [0, 1]$ is not even continuous, let alone differentiable in μ . The results do, however, hold true for smoothed approximations to m . Further investigation into the behavior as the smoothed approximations get better and better might therefore lead to results covering the case of quantiles.

The regression setup discussed in Section 5, required us to consider the covariates as random variables. This is more or less standard in literature concerning empirical likelihood (see e.g. Chen and Van Keilegom (2009)), but deviates somewhat from the classical approach in regression, where the covariates are often treated as fixed and non-stochastic. There is, however, ways of extending the empirical likelihood framework also to the non-i.i.d. setting where the covariates are treated as non-random and fixed in regression settings, see e.g. Owen (1991), Wang and Jing (1999) or Yan and Chen (2015), and we therefore believe the framework of hybrid likelihood should be generalizable to this situation as well. In Lemma 3 in the supplementary material, we showed that the MHL estimator can be considered components of a Z-estimator. Combining this with classical results concerning solutions to generalized estimating equations (see e.g. Shao (2003, chapter 5)) could lead to asymptotic theory and inference techniques for the MHLE estimator also in this non-i.i.d. setup.

Supplementary Materials

Proofs of all theorems can be found in the article’s supplementary material.

Acknowledgements

This research has been partially funded by the Norwegian Research Council through the BigInsight Center for Research Driven Innovation (237718). The authors also gratefully acknowledge partial support from Centre for Advanced Study, at the Academy of Science and Letters, Oslo, in connection with the 2022-2023 project Stability and Change, led by H. Hegre and N.L. Hjort.

References

- Chen, S. X. and I. Van Keilegom (2009). A review on empirical likelihood methods for regression. *Test* 18, 415–447.
- Chib, S., M. Shin, and A. Simoni (2018). Bayesian estimation and comparison of moment condition models. *Journal of the American Statistical Association* 113(524), 1656–1668.
- Cirillo, P. and N. N. Taleb (2016a). The decline of violent conflicts: What do the data really say? *Nobel Foundation Symposium 161: The Causes of Peace*, 1–26.
- Cirillo, P. and N. N. Taleb (2016b). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications* 452, 29–45.
- Claeskens, G. and N. L. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association* 98(464), 900–916.
- Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge, England: Cambridge University Press.
- Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances* 4(2), eaao3580.

- Cunen, C., N. L. Hjort, and H. M. Nygård (2020). Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research* 57(2), 221–234.
- Dæhlen, I., N. L. Hjort, and I. Hobæk Haff (2023). Accurate bias estimation with applications to focused model selection. *Scandinavian Journal of Statistics*. Online publication. DOI: 10.1111/sjos.12696.
- Dæhlen, I. (2021). Empirical and hybrid likelihood. Master thesis.
- Gray, J. (2015). Steven Pinker is wrong about violence and war. The Guardian, Friday 13 March 2015. Available at: <https://www.theguardian.com/books/2015/mar/13/john-gray-steven-pinker-wrong-violence-war-declining>.
- Hall, A. R. and A. Inoue (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics* 114(2), 361–394.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society* 50, 1029–1054.
- Hjort, N. L. (1986). Discussion: On the consistency of Bayes estimates. *The Annals of Statistics* 14(1), 49–55.
- Hjort, N. L., I. W. McKeague, and I. Van Keilegom (2009). Extending the scope of empirical likelihood. *The Annals of Statistics* 37, 1079–1111.
- Hjort, N. L., I. W. McKeague, and I. Van Keilegom (2018). Hybrid combinations of parametric and empirical likelihoods. *Statistica Sinica* 28(4), 2389–2407.
- Imbens, G. W., R. H. Spady, and P. Johnson (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, 951–981.
- Mazo, G. and F. Portier (2021). Parametric versus nonparametric: The fitness coefficient. *Scandinavian Journal of Statistics* 48(4), 1344–1383.

- Molanes Lopez, E. M., I. Van Keilegom, and N. Veraverbeke (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics* 36(3), 413–432.
- Olkin, I. and C. H. Spiegelman (1987). A semiparametric approach to density estimation. *Journal of the American Statistical Association* 82(399), 858–865.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2), 237–249.
- Owen, A. B. (1991). Empirical likelihood for linear models. *The Annals of Statistics* 19, 1725–1747.
- Owen, A. B. (2001). *Empirical likelihood*. Boca Raton, FL: CRC press.
- Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika* 1(2), 261–264.
- Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Toronto: Viking.
- Qin, J. (1994). Semi-empirical likelihood ratio confidence intervals for the difference of two sample means. *Annals of the Institute of Statistical Mathematics* 46, 117–126.
- Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *the Annals of Statistics* 22(1), 300–325.
- Sarkees, M. R. and F. Wayman (2010). *Resort to war: 1816 - 2007*. Washington DC: CQ Press.
- Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics* 35, 634–672.
- Shao, J. (2003). *Mathematical statistics*. New York: Springer Science & Business Media.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.
- Wang, Q.-H. and B.-Y. Jing (1999). Empirical likelihood for partial linear models with fixed designs. *Statistics & Probability Letters* 41(4), 425–433.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric*

society 50, 1–25.

Yan, L. and X. Chen (2015). Empirical likelihood for generalized linear models with fixed and adaptive designs.

Statistics 49(5), 978–988.

Ingrid Dæhlen

Department of mathematics, University of Oslo

E-mail: ingrdae@math.uio.no

Nils Lid Hjort

Department of mathematics, University of Oslo

E-mail: nils@math.uio.no