

Model robust hybrid likelihood

Ingrid Dæhlen*

Department of Mathematics, University of Oslo

and

Nils Lid Hjort

Department of Mathematics, University of Oslo

December 23, 2023

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 centering of the same quantity. 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 prespecified quantity of interest.

Keywords: empirical likelihood, focused model selection, large-sample theory, parametrics and nonparametrics, robustness

*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 Center 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.

1 Introduction and summary

Parametric models are of course widely used in statistics. This is partially due to the fact that maximum likelihood estimators enjoy properties like consistency, asymptotic normality, and optimality when the parametric model is specified correctly. When the true underlying distribution is not a member of the chosen parametric family, however, these desirable properties of maximum likelihood estimators are no longer guaranteed to hold, and instead bias can be introduced. This can potentially weaken or even invalidate an analysis. A way to safeguard against potential bias is to abandon parametric methods entirely and instead use nonparametric alternatives. In general, such techniques lead to asymptotically unbiased results. Typically, however, this unbiasedness comes at the price of increased variability, and in certain situations higher variances might make the methods too imprecise to derive anything meaningful from data. In this article, we study hybrid likelihood theory, a middle ground between these two extreme approaches and a way to borrow strengths from both the parametric and nonparametric schools of thought. 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 et al. (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

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

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 non-

parametric 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. As a result the method borrows strengths from both paradigms, allowing us to be more precise than the empirical likelihood method and more robust than classical likelihood inference.

For the nonparametric part of HL_n , the empirical likelihood function is used. This concept was first introduced in Owen (1988), where the author showed that a scaled version of the log-empirical likelihood function converges to a chi-squared limit at the true value of the parameter of interest. This can be seen as a nonparametric counterpart to Wilks's theorem used in parametric likelihood theory, and allows for construction of approximate nonparametric confidence intervals and hypothesis tests using the empirical in place of the parametric likelihood function. Because of this the empirical likelihood function can really be considered a nonparametric counterpart to the parametric likelihood function, and the hybrid likelihood framework a way of combining nonparametric and parametric likelihood theory.

Since the publication of Owen (1988), many extensions and applications of the theory of empirical likelihood have been made. See e.g. Mykland (1995) and Kitamura (1997) for dependent data, Li (1995) and Adimari (1997) for censored data, and Owen (1991), Chen (1994) and Kolaczyk (1994) for regression settings. In addition, Owen's original version of the empirical likelihood function was only formulated for parameters on the form $Eh(Y)$, where h is some function, and extensions to the more general framework of multidimensional estimating equations were made in Owen (1991) and Qin & Lawless (1994). Furthermore, the theory has had success in advanced applications. Examples

include change point detection (Zou et al. (2007)), fitting of copulas (Chen et al. (2009) and Molanes Lopez et al. (2009)), and Bayesian analysis (Lazar (2003)). Important extensions were also made in Hjort et al. (2009), where plug-in estimates of nuisance parameters and large numbers of estimating equations compared to the amount of data are discussed. The textbook Owen (2001) gives a comprehensive review of the methods, algorithms and the literature of empirical likelihood up to that time, along with numerous illustrations.

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 et al. (2018), where information from empirical and parametric likelihoods constructed using the whole sample are combined. Because of this, hybrid likelihood theory has conceptual similarities with Basu et al. (1998) and Hjort & Glad (1995), more than previous attempts of combining nonparametric and parametric methods.

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 et al. (2009) to derive an alternative characterization of the empirical likelihood function. This allows us to rediscover the main theorem of empirical likelihood theory. In addition, our results give an alternative characterization of the maximum empirical likelihood estimator (MEL estimator), defined and studied in Qin & Lawless (1994). Rather than arguing as in that article, however, we show results similar to those of Imbens et al. (1998) and Schennach (2007), 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 become crucial when we in Section 3 study the asymptotic properties of the hybrid likelihood

function and its maximizer. Hjort et al. (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. 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 et al. (2018) to the case of possible model misspecification. Such results are necessary to make valid inference in situations where the model fit is lacking, and hence, as hybrid likelihood attempts to correct possible bias in the model, crucial for achieving the goals of hybrid likelihood.

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. This criterion is designed to choose between different parametric models available, but we propose an additional use by suggesting to use FIC to decide how much weight should be put on the parametric and nonparametric parts of the hybrid likelihood function. FIC was first introduced in Claeskens & Hjort (2003) for misspecified models where the degree of misspecification disappears in the limit, and in Hjort et al. (2018) the authors derive a FIC similar to that of Claeskens & Hjort (2003). Our version of the FIC, however, is based upon the relatively newer version of the criterion introduced in Jullum & Hjort (2017) with important modifications in Dæhlen et al. (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 & Lawless (1994), White (1982), or Hjort et al. (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 & 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 need to define and take a closer look at the empirical likelihood function.

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

$$\text{EL}_n(\mu) = \max \left\{ \prod_{i=1}^n w_i : \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i m(Y_i, \mu) = 0, w_i \geq 0 \text{ for } i = 1, \dots, n \right\}. \quad (2)$$

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 et al. (2018).

Using Lagrange method of multipliers, one can show

$$\text{EL}_n(\mu) = \prod_{i=1}^n [1 + \lambda_n(\mu)^T m(Y_i, \mu)]^{-1}, \quad (3)$$

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

$$0 = n^{-1} \sum_{i=1}^n \frac{m(Y_i, \mu)}{1 + \lambda^T m(Y_i, \mu)} \quad (4)$$

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

The right-hand side of (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 a function on the form $\theta \mapsto \sum_{i=1}^n \psi(Y_i, \theta)$ for some known function ψ . Such estimators are widely studied in the literature (see e.g. Chapter 5 of Van der Vaart (2000) for an introduction), and standard theory can therefore be used to study the pointwise asymptotic behavior of $\lambda_n(\mu)$. We will now use this observation 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 (4) exists with probability tending to one. In addition, the sequence of solutions has the following properties,

$$\lambda_n \xrightarrow{\text{pr}} \lambda \quad \text{and} \quad \sqrt{n}(\lambda_n - \lambda) = S^{-1}V_n + o_{\text{pr}}(1) \xrightarrow{d} N(0, S^{-1}), \quad (5)$$

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. Theorem 5.41 and 5.42 in Van der Vaart (2000). 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 (4) exists with probability tending to one, we can without loss of generality assume that EL_n can be written as in (3). Taking logarithms and Taylor expanding (2), reveals $\log EL_n(\mu) + \sum_{i=1}^n \log(1 + \lambda^T M_i) = -\sqrt{n}(\lambda_n - \lambda)^T V_n + \frac{1}{2}\sqrt{n}(\lambda_n - \lambda)^T S_n \sqrt{n}(\lambda_n - \lambda) + \epsilon_n$, where $S_n = n^{-1} \sum_{i=1}^n M_i M_i^T / (1 + \lambda^T M_i)^2$. By (5) and the law of large numbers, this implies $\log EL_n(\mu) + \sum_{i=1}^n \log(1 + \lambda^T M_i) = -\frac{1}{2}V_n S^{-1}V_n + \epsilon_n + o_{\text{pr}}(1)$. 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 (5) guarantees that the right-hand side of this inequality is of order $O_{\text{pr}}(1/\sqrt{n})$. In particular, this shows

$$\log EL_n(\mu) = - \sum_{i=1}^n \log[1 + \lambda(\mu)^T m(Y_i, \mu)] - \frac{1}{2}V_n(\mu)^T S(\mu)^{-1}V_n(\mu) + o_{\text{pr}}(1), \quad (6)$$

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 (6), shows $-2 \log EL_n(\mu_0) = V_n(\mu_0)^T S(\mu_0)^{-1}V_n(\mu_0) +$

$o_{\text{pr}}(1) \xrightarrow{d} \chi_q^2$ as $V_n(\mu_0) \xrightarrow{d} N[0, S(\mu_0)]$ by the central limit theorem. Hence, (6) allows us to immediately rediscover the main theorem of empirical likelihood theory. This is, however, not the only application of the alternative characterization in (6). After dividing by the sample size, 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. Chapter 5 of Van der Vaart (2000) for definitions and properties of M-estimators). The following theorem shows that this intuition is indeed correct.

Theorem 2.1. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$ 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 $0 = E \{m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]\}$. Then the MEL estimator $\hat{\mu}$ is consistent for μ_0 , the maximizer of $\mu \mapsto -E \log[1 + \lambda(\mu)^T m(Y, \mu)]$. Furthermore,*

$$\hat{\mu} = \mu_0 + n^{-1} J_{\text{EL}}^{-1} \sum_{i=1}^n \psi_{\mu}(Y_i, \mu_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (7)$$

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 $(\lambda, \mu) \mapsto E \left\{ [1 + \lambda^T m(y, \mu)]^{-1} (m(y, \mu)^T, \lambda^T \partial m(y, \mu))^T \right\}^T$ 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 second and third partial order derivatives of m are bounded by p_j for $j = 0, 1, 2$ respectively with $p_0(Y)^4$, $p_1(Y)^4$ and $p_2(y)^2$ having finite means.

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 $K_{\text{EL}} = \text{Var } \psi_{\mu}(Y, \mu_0)$.

A proof is given in the supplementary material and is based on arguments similar to those given in e.g. Imbens et al. (1998) and Schennach (2007).

Remark. Conditions (1) and (3) of Theorem 2.1 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 quantity $E \{ [1 + \lambda^T m(y, \mu)]^{-1} (m(y, \mu)^T, \lambda^T \partial m(y, \mu)) \}$ to be defined in a neighborhood of $(\mu_0^T, \lambda(\mu_0)^T)^T$ and hence, for the proof to go through. In practice, these considerations matter little as one can truncate data to lie within some arbitrary large and fixed interval to avoid problems with unbounded support and ensure that condition (2) holds.

Assume that μ_T solves the estimating equation $E m(Y, \mu) = 0$ for $Y \sim F$. Then $E \{ m(Y, \mu) / [1 + 0 \cdot m(Y, \mu_T)] \} = 0$, 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 2.2. *Assume the conditions of Theorem 2.1 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 we have $\lambda'(\mu_0) = -[E m(Y, \mu_0) m(Y, \mu_0)^T]^{-1} E \partial m(Y, \mu_0)$. In particular, $\sqrt{n}(\hat{\mu} - \mu_0) \xrightarrow{d} N(0, J_{EL}^{-1})$.*

3 Hybrid likelihood

Hybrid likelihood was introduced in Hjort et al. (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 prespecified parameter of interest. This prespecified parameter is called the control parameter, and we will denote it by μ . For some 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 . We define the hybrid likelihood function as in (1).

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 $0 = n^{-1} \sum_{i=1}^n m(Y_i, \mu)$. Another way of looking at $HL_n(\theta)$ is as a version of the empirical likelihood function borrowing strength from the parametric model. How much weight is put on each part is decided by the balance parameter, a . Hjort et al. (2018) discuss interpretations and nuances in the definition of the hybrid likelihood function in detail; consult that paper for more discussions and insights.

We will study the maximizer of the hybrid likelihood function, the maximum hybrid likelihood estimator (the MHL estimator). Hjort et al. (2018) study its behavior under the condition that the parametric model f_θ is correctly specified. When this is the case, 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. Classical results concerning the empirical likelihood function in shrinking neighborhoods of the true value (e.g. Hjort et al. (2009) and Molanes Lopez et al. (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 et al. (2009) and Hjort et al. (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 point-wise 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 (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 3.1. *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. Assume that m is used in construction of the empirical likelihood function EL_n . Let f_θ for $\theta \in \Theta \subseteq \mathbb{R}^p$ 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 . 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 2.1 hold true at $(\lambda_0^T, \mu(\theta_0)^T)^T$, in addition to the following assumptions:*

- (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 .*

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

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

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 2.1 and $J^* = (1 - a)J + aH_{\theta_0}\text{E} \log[1 + \lambda(\theta)^T M(\theta)]$ where J is the Fisher matrix of the model f_θ at θ_0 . In particular, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (J^*)^{-1}K^*(J^*)^{-1})$, where $K^* = \text{Var} \psi(Y, \theta_0)$.

Estimators of J^* and K^* will be discussed in Section 4; see in particular Definition 4.2.

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

the expression $-\mathbb{E} \log [1 + \lambda(\mu)^T m(Y, \mu)]$ by the arguments preceding Corollary 2.2. As a result, both $\mathbb{E} \log f_\theta(Y)$ and $-\mathbb{E} \log[1 + \lambda(\theta)^T M(\theta)]$ are maximized by θ_0 , making this vector the maximizer of $\theta \mapsto (1-a)\mathbb{E} \log f_\theta(Y) - a\mathbb{E} \log[1 + \lambda(\theta)^T M(\theta)]$. By Corollary 2.2, we also have $\psi_\theta(y, \theta_0) = (1-a)u(y, \theta_0) - a\mathbb{E} \partial m(Y, \theta_0) \mathbb{E} \{m(Y, \theta_0)m(Y, \theta_0)^T\}^{-1} m[y, \mu(\theta_0)]$. Plugging this into Theorem 3.1 shows $K^* = (1-a)^2 J + a^2 \xi_0^T W^{-1} \xi_0 - a(1-a)(C W^{-1} \xi_0 + \xi_0^T W^{-1} C^T)$, 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 $\xi_0 = \mathbb{E} \partial m(Y, \theta_0)$. In addition, $J^* = (1-a)J + a\xi_0^T W^{-1} \xi_0$ by the chain rule and Corollary 2.2. Hence, if the model is correctly specified, Theorem 3.1 simplifies to the corresponding results given in Hjort et al. (2018). Because of this, we believe Theorem 3.1 to be a proper extension of the theorems given in Hjort et al. (2018) and a tool for making model robust inference in the hybrid likelihood setting.

3.2 The hybrid divergence

By Theorem 3.1, 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 minimize 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:

$$\max_{w: \mathbb{R}^d \rightarrow \mathbb{R}} \{E \log w(Y) : E w(Y)M(\mu) = 0 \text{ and } E w(Y) = 1\}. \quad (8)$$

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 (8) is equivalent to minimizing $KL(g, h)$ with respect to h under the condition that h is a pmf/pdf and $E_h M(\mu) = 0$. Here $KL(g, h)$ is the Kullback-Leibler divergence from h to g defined as $E \log h(Y)/g(Y)$.

Since subtracting a constant does not change the extrema of an expression, the above shows that the maximizer of $(1 - a)E \log f_\theta(Y) - aE \log[1 + \lambda(\theta)^T M(\theta)]$, i.e. the limit of the MHL estimator, is also the minimizer of $\theta \mapsto d_a(g, f_\theta)$ where

$$d_a(g, h) = (1 - a)KL(g, h) + a \min_h \{KL(g, h) : E_h M[\mu(h)] = 0\}. \quad (9)$$

with $\mu(h)$ being the solution to the estimating equation $E_h M(\mu) = 0$. 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, $KL(g, g) = 0$ and $E_g M[\mu(g)] = 0$. 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 (9) 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 3.1, 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 1 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 1 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 $\text{Gamma}(10, 5)$ variable not exceeding one.

Remark. 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.

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 not be guided by general rules, and should rather be set to reflect what parameters are of particular interest in the specific application. In some situations, the same might be true

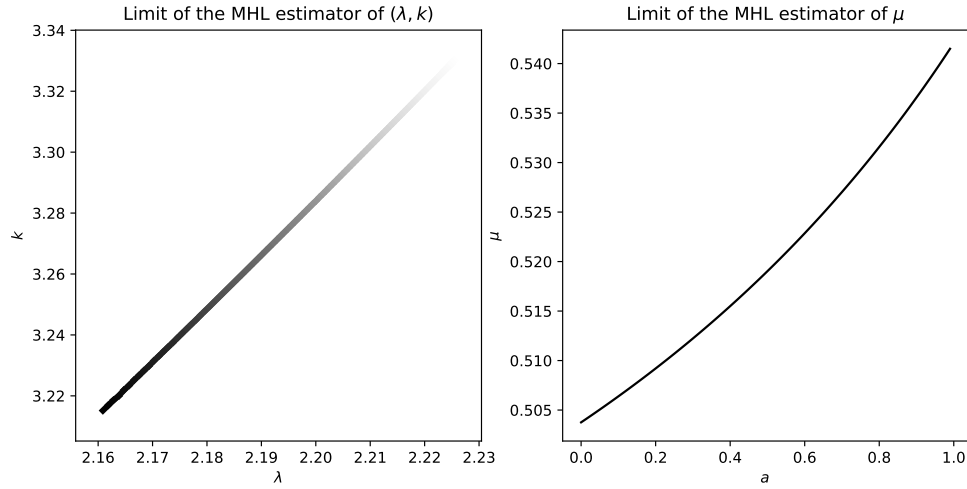


Figure 1: The figure to the left shows the limit of the MHL estimator of the parameters (λ, k) for different values of a . The balance parameter is indicated by the color of the line. Light colors corresponds to small values of a , while darker colors indicate a close to one. In the plot to the right we have displayed the limit of the MHL estimator of the control parameter μ as a function of the balance parameter a .

for the balance parameter. Usually, however, how much weight should be put on the parametric and nonparametric part of the hybrid likelihood function is dependent on the quality of the model fit and estimator of the control parameter. If either is particularly trustworthy, the balance parameter should be set to emphasize the corresponding part in the hybrid likelihood function. If, on the other hand, the model fit or estimator of the control parameter is lacking, the balance parameter should make sure that the corresponding part will not dominate the hybrid likelihood function and put more weight on the remaining part. In this section, we will propose a way to choose a in a guided fashion.

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 et al. (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 et al. (2018), where the authors derive formulas for a focused information criterion for models in shrinking neighborhoods of the true underlying density. More precisely, Hjort et al. (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 an 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 & Hjort (2017) with important modifications in Dæhlen et al. (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 the results of Jullum & Hjort (2017) and Dæhlen et al. (2023) valid in the hybrid likelihood context.

Theorem 4.1. *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*

$$\hat{\xi} = \xi + n^{-1} \sum_{i=1}^n \phi(Y_i, \xi) + \delta_n \quad (10)$$

for some function $\phi: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\delta_n = O_{\text{pr}}(1/n)$. Assume further that all conditions of Theorem 3.1 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 $\hat{b} = g(\hat{\theta}) - \hat{\xi}$, $c = \lim_{n \rightarrow \infty} nE\delta_n$, $\kappa = \text{Var}[\nabla g(\theta_0)^T (J^*)^{-1} \psi_\theta(Y, \theta_0) - \phi(Y, \xi)]$ and $\tau = (J^*)^{-1} K^* (J^*)^{-1}$ provided the assumptions outlined in Jullum & Hjort (2017) and Dæhlen et al. (2023) hold true.

The crucial part of Theorem 4.1 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 3.1 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.

Our goal is to find a way to tune the balance parameter used in the construction of the hybrid likelihood function, by minimizing the mean squared error of the MHL estimator of the focus parameter. This can be achieved by computing the approximation $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$ for a selection of a -values and choosing the balance parameter with the smallest approximated MSE. In practice, however, this is rarely possible, as the quantities c , κ and τ are usually unknown. We therefore need to define estimators for these parameters and approximations to the expression $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$. This is provided in the following definition.

Definition 4.2. Let all quantities be as defined in Theorem 4.1. We define the following approximation to the MSE of $g(\hat{\theta})$,

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

where $\hat{c}$ is either of the estimators given in Dæhlen et al. (2023), $\hat{\tau} = \nabla g(\hat{\theta})^T (\hat{J}^*)^{-1} \hat{K}^* (\hat{J}^*)^{-1} \nabla g(\hat{\theta})$ and $\hat{\kappa} = \hat{\tau} + n^{-1} \sum_{i=1}^n \phi(Y_i, \hat{\xi})^2 - 2\nabla g(\hat{\theta})^T (\hat{J}^*)^{-1} \hat{V}$, with $\hat{V} = n^{-1} \sum_{i=1}^n \psi_{\theta}(Y_i, \hat{\theta}) \phi(Y_i, \hat{\xi})$, $\hat{K}^* = n^{-1} \sum_{i=1}^n \psi_{\theta}(Y_i, \hat{\theta}) \psi_{\theta}(Y_i, \hat{\theta})^T$ and $\hat{J}^* = n^{-1} H h_n(\hat{\theta})$.

Remark. 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 $\hat{J}^*$, $\hat{K}^*$, $\hat{V}$ and $n^{-1} \sum_{i=1}^n \phi(Y_i, \hat{\xi})^2$ are consistent for J^* , K^* , $\text{Var}\phi(Y, \xi)$ and $E \psi_{\theta}(Y, \theta_0) \phi(Y)$ respectively under weak conditions. See e.g. Lemma 16 of Dæhlen et al. (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 3.1, 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. In Section 6, we will illustrate how this is done in practice.

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 et al. (2018).

Let $Y \in \mathbb{R}$ be the response 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 we wish to fit a linear regression model to these data. The ordinary least square estimator is the maximum likelihood estimator of β in the model $X \sim N(\mu, \Sigma)$ and $Y | X = x \sim N(\beta_0 + \beta^T x, \sigma^2)$. For simplicity we will assume that μ , Σ , and σ are known. In this case, the model has log-likelihood function $\ell_n(\beta) = \sum_{i=1}^n [\log \phi_d(X_i | \mu, \Sigma) + \log \phi_1(Y_i | \beta_0 + \beta^T x, \sigma)]$, where ϕ_d is the density of a multivariate normal distribution on $\mathbb{R}^d$ for $d > 1$ and ϕ_1 is the normal density on $\mathbb{R}$.

Suppose we are unsure if a subset of the covariates is relevant for predicting the outcome of Y . It could then 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 $\beta_{\mathcal{S}}(\beta) = (E X_{\mathcal{S}} X_{\mathcal{S}}^T)^{-1} E X_{\mathcal{S}} X = (\Sigma_{\mathcal{S}, \mathcal{S}} + \mu \mu^T)^{-1} (\Sigma_{\mathcal{S}, :} + \mu \mu^T) \beta$, where $\Sigma_{\mathcal{S}, \mathcal{S}}$ is the covariance matrix of $X_{\mathcal{S}}$ and $\Sigma_{\mathcal{S}, :}$ is the covariance between $X_{\mathcal{S}}$ and the full component vector X . Hence, for fixed a we can define the hybrid log-likelihood function as $h_n(\beta) = (1 - a)\ell_n(\beta) + a \log \text{EL}_n[\beta_{\mathcal{S}}(\beta)]$, 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 $X_{\mathcal{S}}$. If overall predictive power is the main goal, the balance parameter can be selected to minimize mean squared error estimated with cross validation. If there is a certain focus parameter we wish to tailor the fit to, HFIC can be used instead. An example of the latter situation can be in personalized prediction where $\beta^T x_0$ for some fixed x_0 can be used as focus parameter. In either case, once a is chosen, inference of β can be made using Theorem 3.1.

Remark. 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 & Taleb (2016*a,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 & Wayman (2010)). Cunen, Hjort & 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 et al. (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 & Nygård (2020). Because of this, we will use shifted log-normal models for both data sets and fit them using maximum hybrid likelihood.

We start by choosing a control parameter. Since $\mu \mapsto I(Y \leq \mu) - 1/2$ is discontinuous in μ and Theorem 3.1 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.

We chose separate balance parameters for each model. In each case, this was done by

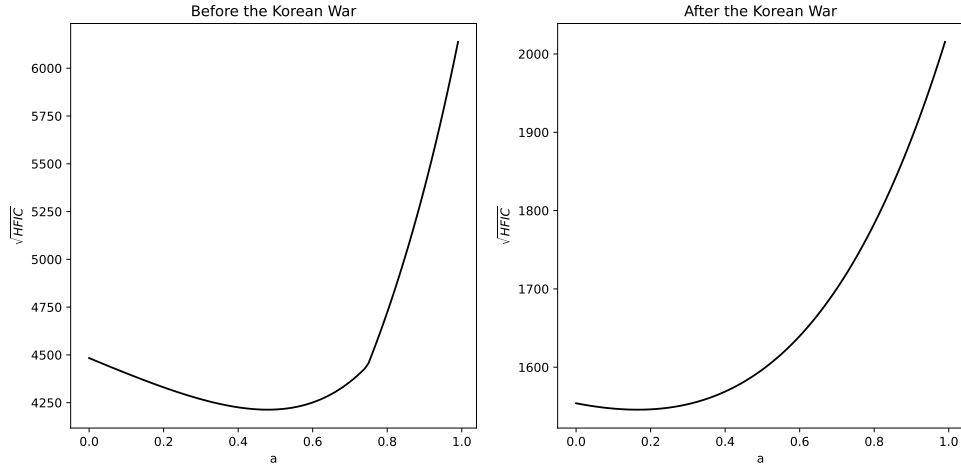


Figure 2: The figure displays root-HFIC scores as a function of the balance parameter (a) for older and newer wars, in the left and right plots respectively.

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 4.2. Since the median satisfies (10) 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 4.2, with $f(\mu)$ replaced by non-parametric kernel density estimates. For more details see Dæhlen et al. (2023), in particular their Appendix A. The results are displayed in Figure 2. We see that minima are attained neither at 0 nor 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 3.1 in conjunction with the delta method and the estimators defined in Definition 4.2. 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 estimated standard deviation of 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 3.1. 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 was rejected. Hence, we can conclude that the larger wars have indeed become less deadly, which is consistent with the findings of Cunen, Hjort & 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 with and significantly extended the the hybrid likelihood framework of Hjort et al. (2018). Firstly, we derived alternative characterizations of the log-empirical and hybrid likelihood function, giving insight into how the maximizer of these functions

behave. Furthermore, we have shown that the maximizer of the empirical as well as the hybrid likelihood estimator converge in probability to the minimizer of well-defined limit functions. We also studied the divergence minimized by the best hybrid likelihood fit and how it behaves for different value of the balance parameter. Furthermore, we proved that the maximizers of the empirical and hybrid likelihood functions converge in distribution to normal distributions after properly scaling and centering the quantities. This allows for the construction of approximate confidence intervals and hypothesis tests. We also defined estimators of the asymptotic variances and used these in conjunction with the aforementioned limit results to define a focused information criterion. This allowed us to propose a way of choosing a value for the balance parameter, which minimizes the estimated mean squared error of some prespecified quantity of interest. Lastly, we showed how the hybrid likelihood framework could be applied to regression settings and illustrated the developed theory in an application.

The perhaps most pressing issue with the theory developed in this article, is the assumption (2) of Theorem 2.1. If m is the estimating function used in the construction of the empirical likelihood function, this condition requires $m(y, \mu)$ to have bounded support for each fixed μ in the parameter space. This excludes many very important cases, e.g. when estimating the mean using $m(y, \mu) = y - \mu$ and data is normally distributed. In fact, none of the theorems can be applied if $m(y, \mu) = h(y) - \mu$ and $h(Y)$ has unbounded support. This is of course quite problematic, but the assumption is crucial for the theorems to go through. In practice, however, this is less of a problem as approximations using truncated distributions often give very good results. Take for instance, the situation considered above 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 . We believe investigating the limiting behavior of the MHL estimator as the smoothed approximations get better and better can lead to results covering the case of quantiles, but further investigation is required to obtain formal proofs.

In Section 5, we discussed how the hybrid likelihood framework could be applied to regression settings. For this case we suggested using cross validation to find a fitting value for the balance parameter a rather than the information criterion developed in this article. The reason for this is mainly that there is no obvious choice of focus parameter in regression settings. A possibility is to use $\beta^T x_0$ for some fixed x_0 , but in order to compute HFIC, we need a consistent estimator of the true value of $\beta^T x_0$ and it is not immediately obvious what such an estimator should be. There has been a couple of attempts to solve this problem in the literature concerning the focused information criterion (see e.g. Cunen, Walløe & Hjort (2020)), but a general solution which can be applied in most situations has yet to be found.

Lastly, we would like to mention that the hybrid likelihood framework can be extended by replacing the empirical likelihood by a generalized empirical likelihood, see e.g. Newey & Smith (2004). This is perhaps most apparent when looking at the distance d_a defined in (9). In Section 3.2 we showed that the maximum hybrid likelihood estimator aims for θ_0 such that f_{θ_0} minimized d_a in the class f_θ for $\theta \in \Theta$, but a natural question to ask is why this particular divergence is beneficial to minimize? In principle, we could replace the Kullback-

Leibler divergence in (9) by any other divergence function. Doing this for the rightmost appearance of KL only, corresponds to replacing the empirical likelihood function by the generalized empirical likelihood function. Although such ideas are outside the scope of this article, further investigation into potential benefits with using the generalized empirical likelihood function is certainly worth investigating.

References

- Adimari, G. (1997), ‘Empirical likelihood type confidence intervals under random censorship’, *Annals of the Institute of Statistical Mathematics* **49**, 447–466.
- Basu, A., Harris, I. R., Hjort, N. L. & Jones, M. (1998), ‘Robust and efficient estimation by minimising a density power divergence’, *Biometrika* **85**(3), 549–559.
- Chen, J., Peng, L. & Zhao, Y. (2009), ‘Empirical likelihood based confidence intervals for copulas’, *Journal of Multivariate Analysis* **100**(1), 137–151.
- Chen, S. X. (1994), ‘Empirical likelihood confidence intervals for linear regression coefficients’, *Journal of Multivariate Analysis* **49**(1), 24–40.
- Cirillo, P. & Taleb, N. N. (2016a), ‘The decline of violent conflicts: What do the data really say?’, *Nobel Foundation Symposium 161: The Causes of Peace* pp. 1–26.
- Cirillo, P. & Taleb, N. N. (2016b), ‘On the statistical properties and tail risk of violent conflicts’, *Physica A: Statistical Mechanics and its Applications* **452**, 29–45.
- Claeskens, G. & Hjort, N. L. (2003), ‘The focused information criterion’, *Journal of the American Statistical Association* **98**(464), 900–916.

- Claeskens, G., Hjort, N. L. et al. (2008), *Model selection and model averaging*, Cambridge University Press, Cambridge, England.
- Clauset, A. (2018), ‘Trends and fluctuations in the severity of interstate wars’, *Science Advances* **4**(2), eaao3580.
- Cunen, C., Hjort, N. L. & Nygård, H. M. (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.
- Cunen, C., Walløe, L. & Hjort, N. L. (2020), ‘Focused model selection for linear mixed models with an application to whale ecology’, *The Annals of Applied Statistics* **14**, 872–904.
- Dæhlen, I., Hjort, N. L. & Hobæk Haff, I. (2023), ‘Accurate bias estimation with applications to focused model selection’, *Scandinavian Journal of Statistics*. Online publication. DOI: 10.1111/sjos.12696.
- 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>.
- Hjort, N. L. & Glad, I. K. (1995), ‘Nonparametric density estimation with a parametric start’, *The Annals of Statistics* **23**, 882–904.
- Hjort, N. L., McKeague, I. W. & Van Keilegom, I. (2009), ‘Extending the scope of empirical likelihood’, *The Annals of Statistics* **37**, 1079–1111.
- Hjort, N. L., McKeague, I. W. & Van Keilegom, I. (2018), ‘Hybrid combinations of parametric and empirical likelihoods’, *Statistica Sinica* **28**(4), 2389–2407.

- Imbens, G. W., Spady, R. H. & Johnson, P. (1998), ‘Information theoretic approaches to inference in moment condition models’.
- Jullum, M. & Hjort, N. L. (2017), ‘Parametric or nonparametric: The fic approach’, *Statistica Sinica* **27**, 951–981.
- Kitamura, Y. (1997), ‘Empirical likelihood methods with weakly dependent processes’, *The Annals of Statistics* **25**(5), 2084–2102.
- Kolaczyk, E. D. (1994), ‘Empirical likelihood for generalized linear models’, *Statistica Sinica* **4**, 199–218.
- Lazar, N. A. (2003), ‘Bayesian empirical likelihood’, *Biometrika* **90**(2), 319–326.
- Li, G. (1995), ‘On nonparametric likelihood ratio estimation of survival probabilities for censored data’, *Statistics & Probability Letters* **25**(2), 95–104.
- Molanes Lopez, E. M., Van Keilegom, I. & Veraverbeke, N. (2009), ‘Empirical likelihood for non-smooth criterion functions’, *Scandinavian Journal of Statistics* **36**(3), 413–432.
- Mykland, P. A. (1995), ‘Dual likelihood’, *The Annals of Statistics* **23**, 396–421.
- Newey, W. K. & Smith, R. J. (2004), ‘Higher order properties of gmm and generalized empirical likelihood estimators’, *Econometrica* **72**(1), 219–255.
- 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*, CRC press, Boca Raton, FL.

- Pinker, S. (2011), *The better angels of our nature: Why violence has declined*, Viking, Toronto.
- 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. & Lawless, J. (1994), ‘Empirical likelihood and general estimating equations’, *the Annals of Statistics* **22**(1), 300–325.
- Sarkees, M. R. & Wayman, F. (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.
- Van der Vaart, A. W. (2000), *Asymptotic statistics*, Cambridge University Press, Cambridge, England.
- White, H. (1982), ‘Maximum likelihood estimation of misspecified models’, *Econometrica: Journal of the econometric society* **50**, 1–25.
- Zou, C., Liu, Y., Qin, P. & Wang, Z. (2007), ‘Empirical likelihood ratio test for the change-point problem’, *Statistics & probability letters* **77**(4), 374–382.

SUPPLEMENTARY MATERIAL

Title: Proofs of Theorem 2.1 and Theorem 3.1. (pdf)