



Model robust hybrid likelihood

Ingrid Dæhlen ^{a,b} , Nils Lid Hjort ^a

^a Department of Mathematics, University of Oslo, Moltke Moes vei 35, Oslo, 0851, Norway

^b Norwegian Computing Center, Post box 114 Blindern, Oslo, 0314, Norway

ARTICLE INFO

Keywords:

Empirical likelihood
Focused model selection
Large-sample theory
Parametrics and nonparametrics
Robustness

ABSTRACT

This 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 hybrid likelihood function is sorted out, also outside of the parametric model conditions. We prove a profiling result as well as limiting behaviour of the maximizer of the hybrid likelihood function. 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.

1. Introduction and summary

Hjort et al. (2018) introduced hybrid likelihood, 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

$$HL_n(\theta) = L_n(\theta)^{1-a} EL_n[\mu(\theta)]^a, \quad (1)$$

where $EL_n[\mu(\theta)]$ 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 in $[0, 1)$ 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. Classical likelihood theory corresponds to the special case $a = 0$.

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) or Qin (2000). 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.

* Corresponding author at: Department of Mathematics, University of Oslo, Moltke Moes vei 35, Oslo, 0851, Norway.

E-mail address: ingrdae@math.uio.no (I. Dæhlen).

<https://doi.org/10.1016/j.jspi.2025.106327>

Received 26 September 2024; Received in revised form 2 July 2025; Accepted 10 July 2025

Available online 22 July 2025

0378-3758/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Robustness is in this article taken to mean robust against possible model misspecification, in the sense that even when the parametric model used in construction of the hybrid likelihood function is misspecified, all results should be valid and the estimator of the control parameter should be a good one. This deviates somewhat from the traditional use of the robustness concept, related to the behaviour of methods in the presence of outliers, and techniques from this highly related field can therefore be seen as an alternative to the approach studied in this paper, e.g. Basu et al. (1998) or Huber and Ronchetti (2011). Partial protection against outliers and ‘false data’ can be achieved via choices for the control parameters.

In this article, we derive new results for the hybrid and empirical likelihood function. In Section 2 and Section 3.1, we use arguments similar to those of Molanes-Lopez et al. (2009) to derive alternative characterizations of these maps. This is used to study the maximizers of the empirical and hybrid likelihood functions (the MEL and MHL estimator respectively). In particular, we show that these estimators are consistent for minimizers of well-defined divergences and that after proper scaling and centring, they converge in distribution to a normal limit. We also work directly with the empirical and hybrid likelihood functions and show a profiling result for both maps. All of our results hold under possible model misspecification. As Hjort et al. (2018) only considered the case of correctly specified models, our theorems are novel results for hybrid likelihood theory. For the MEL estimator, Qin and Lawless (1994) proved its consistency and asymptotic normality under the assumption of correct model specification, but to the best of our knowledge, no existing result for the MEL estimator cover the case of possible model misspecification. Schennach (2007) and Chib et al. (2018) study this situation to some degree, but they both work with exponentially tilted empirical likelihoods, resulting in different estimators than the MEL estimator. We also mention that some preliminary results were reached in the master thesis Dæhlen (2021).

To prove our results, we use the framework of generalized estimating equations, see e.g. Shao (2003, Ch. 5), and the theory from this field allows us to easily extend our results to allow for plug-in estimators in the empirical likelihood function, i.e. $EL_n(\mu, \hat{P})$ where $\hat{P}$ is some estimator. The behaviour of the empirical likelihood function in the presence of plug-in parameters has been studied in Hjort et al. (2009), and Molanes-Lopez et al. (2009) worked with a particular type of plug-in estimator. So far, however, no general result concerning the limiting behaviour of the MEL and MHL estimator in the presence of plug-in estimators has been derived. We believe our results fill this gap in the literature.

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 models in a local misspecified modelling framework and in Hjort et al. (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 et al. (2024). 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.

We go on to illustrate the hybrid likelihood machinery in Section 5, pertaining to the important battle death series for all major interstate wars from 1823 to the 2022; 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. This section will serve as motivation for the theory we derive for the hybrid likelihood function in Section 3, in addition to establish notation which will be used throughout this article.

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 form

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

Using the Lagrange method, one can show

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

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, see e.g. Owen (2001) or Hjort et al. (2018).

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 an expression on the form $\sum_{i=1}^n \psi(Y_i, \theta)$ for some known function ψ . Such estimators have been well studied in the literature (see e.g. van der Vaart (2000, Ch. 5)), and we will now use the relevant 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 (4) exists with probability tending to one. In addition, $\lambda_n \xrightarrow{\text{pr}} \lambda$ and $\sqrt{n}(\lambda_n - \lambda) = S^{-1}V_n + o_{\text{pr}}(1)$, where λ is the solution to the equation $0 = E[M/(1 + \lambda^T M)]$ subjected to $0 = \Pr(1 + \lambda^T M \leq 0)$, and $S = E[MM^T/(1 + \lambda^T M)^2]$ and $V_n = n^{-1/2} \sum_{i=1}^n M_i/(1 + \lambda^T M_i)$. This follows from standard results, e.g. van der Vaart (2000, Th. 5.31 and 5.42).

Since λ_n 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 $\log \text{EL}_n(\mu) + \sum_{i=1}^n \log(1 + \lambda^T M_i) = -\sqrt{n}(\lambda_n - \lambda)^T V_n + \frac{1}{2} \sqrt{n}(\lambda_n - \lambda)^T S_n \sqrt{n}(\lambda_n - \lambda) + \epsilon_n$, where $S_n = n^{-1} \sum_{i=1}^n M_i M_i^T / (1 + \lambda^T M_i)^2$. Hence, $\log \text{EL}_n(\mu) + \sum_{i=1}^n \log(1 + \lambda^T M_i) = -\frac{1}{2} V_n S_n^{-1} V_n + \epsilon_n + o_{\text{pr}}(1)$. Furthermore, since open balls are convex, Taylor's theorem and the triangle inequality ensures $|\epsilon_n| \leq \|\lambda_n - \lambda\|^3 \sum_{i=1}^n \|M_i\|^3 / 3L^3$, which is sufficient for

$$\log \text{EL}_n(\mu) = - \sum_{i=1}^n \log[1 + \lambda(\mu)^T m(Y_i, \mu)] - \frac{1}{2} V_n(\mu)^T S(\mu)^{-1} V_n(\mu) + o_{\text{pr}}(1), \quad (5)$$

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 (5) 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. This result was proven by Owen (1988) (albeit in a very different way) and is often referred to as the main theorem of empirical likelihood theory. Furthermore, after dividing by n we notice that the right-hand side of the equation is asymptotically equivalent to $-n^{-1} \sum_{i=1}^n \log[1 + \lambda(\mu)^T m(Y_i, \mu)]$. In the following section, we will use this fact to derive model robust theory in the context of hybrid likelihood.

3. Hybrid likelihood

Hybrid likelihood inference was introduced in Hjort et al. (2018) and is based on the study of the associated 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 called the control parameter and denoted by μ . We will assume that μ can be identified by the estimating equation $E[m(Y, \mu)] = 0$. 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 $\text{HL}_n(\theta) = L_n(\theta)^{1-a} \text{EL}_n[\mu(\theta)]^a$, where $\text{EL}[\mu(\theta)]$ 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 (parametrized 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 with $m(y, \mu) = y - \mu$ and $\mu(\lambda, k) = \lambda \Gamma(1 + 1/k)$ where Γ is the Gamma function.

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 . Such a construction could be beneficial in situations where the fit in certain regions of the domain of y are more important than others. More generally, for bins $C_1, \dots, C_k$, associated with a histogram, the hybrid likelihood framework can control for robust estimation of the probabilities $\mu_j = \Pr(Y \in C_j)$ for $j = 1, \dots, k$, by using the estimating equation $m_j(u, \mu_j) = I(y \in C_j) - \mu_j$ for $j = 1, \dots, k - 1$.

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 range 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. The fit obtained using maximum likelihood (ML) is displayed in Fig. 1(a), together with the fits resulting from maximizing the hybrid likelihood function for different values of the balance parameter. From the figure, notice that as a increases in value, the fitted density moves from the ML estimate towards a curve which agrees somewhat more with the histogram of the data. As the Weibull-model fits the data quite well in this case, the fits are quite similar for different values of a .

In other situations, however, the difference may be more substantial than for the data with Egyptian life lengths. Consider for instance Simon Newcomb's 66 measurements of the speed of light (published in e.g. Stigler, 1977). Each observation is measured in nanoseconds and recorded as the deviation from 24.8. The true value is 33.2. A histogram is shown in Fig. 1(b) together with normal densities fitted using histogram controlled maximum hybrid likelihood for different values of a . From the figure we see that we get very different fits with different choices of balance parameter. This is a consequence of the outliers in the data set dominating the maximum likelihood fit and increased robustness as $a \rightarrow 1$.

We also note that the idea of histogram correction can be extended to the multivariate case by defining suitable multidimensional regions, e.g. rectangles in $\mathbb{R}^2$ and boxes in $\mathbb{R}^3$.

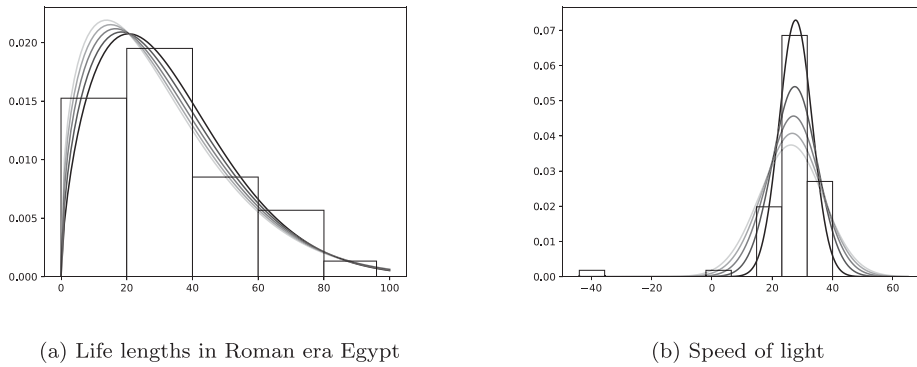


Fig. 1. Histograms and densities of distributions fitted using maximum hybrid likelihood for two data examples. The values used for a are $a = 0, 0.25, 0.5, 0.75, 0.99$, indicated by the lightness of the lines (from light to dark).

3.1. The hybrid likelihood function and its maximizer

Now that we have introduced the framework of hybrid likelihood, we will study the asymptotic behaviour of the maximizer of the hybrid likelihood function. We will refer to this as the maximum hybrid likelihood estimator (the MHL estimator).

Similarly as we did for the empirical likelihood function, we start by deriving a pointwise asymptotically equivalent expression for $n^{-1} \log \text{HL}_n(\theta)$, so $a \in [0, 1)$ 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 (5) shows $h_n(\theta) = \sum_{i=1}^n \psi[Y_i, \mu(\theta)] + O_{\text{pr}}(1)$, where $\psi(y, \theta) = a \log f_\theta(y) - (1-a) \log \{1 + \lambda[\mu(\theta)]^T m[y, \mu(\theta)]\}$, showing that $n^{-1} h_n(\theta)$ is asymptotically equivalent to a mean. For this reason, we would expect h_n to behave “like a likelihood”. In this section, we will show that this is indeed the case and that many famous results in parametric likelihood theory have counterparts for the hybrid likelihood.

When the parametric likelihood is misspecified, the maximum likelihood estimator is famously consistent for the minimizer of the Kullback–Leibler divergence and asymptotically normal after proper scaling and centring (White, 1982). By the previous arguments, we would expect a similar property to hold true for the MHL estimator. The following theorem confirms this intuition. The proof is given in the supplementary material.

Theorem 1. Assume the conditions given in the supplementary material hold true and let $\psi(y, \theta) = (1-a) \log f_\theta(y) - a \log [1 + \lambda(\theta)^T m(y, \theta)]$. Then the MHL estimator $\hat{\theta}_a$ is consistent for θ_a^* , the minimizer of $E\psi(Y, \theta)$, and

$$\hat{\theta}_a = \theta_a^* + n^{-1} (J_a^*)^{-1} \sum_{i=1}^n \nabla_{\theta_a^*} \psi(Y_i, \theta_a^*) + o_{\text{pr}}(1/\sqrt{n}), \quad (6)$$

where $J_a^* = -H_{\theta_a^*} E\psi(Y, \theta)$ and $\nabla_{\theta_a^*}$ and $H_{\theta_a^*}$ denotes the gradient and hessian matrix evaluated at θ_a^* respectively. In particular,

$$\sqrt{n}(\hat{\theta}_a - \theta_a^*) \xrightarrow{d} N[0, (J_a^*)^{-1} K_a^* (J_a^*)^{-1}],$$

where $K_a^* = \text{Var} \nabla_{\theta_a^*} \psi(Y, \theta)$.

Remark 1. Most of the assumptions of Theorem 1 are more or less standard conditions ensuring sufficient regularity. There is, however, one assumption which warrants some discussion. For Theorem 1 to go through, we need there to be a neighbourhood of $((\theta_a^*)^T, \lambda(\theta_a^*)^T)^T$ on which $1 + \lambda^T M(\theta) > L$ with probability 1 for some $L > 0$. This condition is very limiting, as it excludes situations where $m(Y, \mu)$ has unbounded support. The assumption is, however, fully necessary by Schennach (2007, Th. 1). In practice, this problem can be avoided by assuming the $m(Y_i, \theta)$ always lie within some arbitrary large and fixed set.

By definition, θ_a^* is equal to the minimizer of the Kullback–Leibler divergence from f_θ to F when $a = 0$. Furthermore, notice that J_0^* and K_0^* are equal to the Fisher information matrix and the variance of the score function in the model f_θ , respectively. Hence, Theorem 1 reduces to the results of White (1982) when $a = 0$. Additionally, even though the hybrid likelihood function is defined only for $a \in [0, 1)$, the theorem above can be applied when $a = 1$ and $h_n(\theta) = \log \text{EL}_n[\mu(\theta)]$. In this case, it is most natural to let $\mu(\theta) = \theta$, for which Theorem 1 simplifies to a version of Qin and Lawless (1994, Th. 1).

Notice that Theorem 1 requires differentiability of m with respect to μ and not y . For this reason, histogram correction can be achieved through the maximum likelihood framework without requiring nonparametric density estimation. This is very useful, especially in higher dimensions where kernel methods work poorly.

Since the right hand side of (6) is a mean, Theorem 1 can be used to approximate the limiting distribution of $\hat{\theta}_a$ by a normal distribution. This can in turn be used to create approximate confidence intervals for θ_a^* . By applying the delta method, inference can be made of functions of θ_a^* in a similar manner. Since the limiting distribution of $\sqrt{n}(\hat{\theta}_a - \theta_a^*)$ is normal, however, confidence intervals based on Theorem 1 will always be symmetric. A way to create skewed confidence intervals in parametric likelihood

theory, is to use Wilks' theorem (see e.g. [Schweder and Hjort \(2016, Ch. 2\)](#)), stating that the profile log-likelihood function has a limiting chi-squared distribution. The following theorem shows that h_n has a similar behaviour.

Theorem 2. Let $g: \mathbb{R}^p \rightarrow \mathbb{R}^k$ be a function and define the profile hybrid log-likelihood function as $h_{n,\text{prof}}(\zeta) = \max\{h_n(\theta) : g(\theta) = \zeta\}$. Then, under the conditions stated in the supplementary material,

$$D_{n,a} = -2\{h_{n,\text{prof}}[g(\hat{\theta}_a)] - h_{n,\text{prof}}(\zeta^*)\} \xrightarrow{d} \text{Tr}\{MUU^T\}$$

where $\hat{\zeta} = g(\hat{\theta}_a)$, $\zeta^* = g(\theta_a^*)$, $U \sim N(0, I_p)$ and $M = [g'(\theta_a^*)(J^*)^{-1}g'(\theta_a^*)^T]^{-1}g'(\theta_a^*)(J_a^*)^{-1}K_a^*(J_a^*)^{-1}$. In particular, when $\zeta \in \mathbb{R}$

$$D_{n,a} \rightarrow \frac{\nabla g(\theta_a^*)^T (J_a^*)^{-1} K_a^* (J_a^*)^{-1} \nabla g(\theta_a^*)}{\nabla g(\theta_a^*)^T (J_a^*)^{-1} \nabla g(\theta_a^*)} \chi_1^2. \quad (7)$$

By previous arguments $J_0^* = J$ and $K_0^* = K$, where J is the Fisher matrix and $K = \text{Var} \nabla \log f_{\theta_a^*}(Y)$ when $a = 0$. Hence, [Theorem 2](#) reduces to a version of Wilks' theorem (possibly) misspecified models when $a = 0$. By letting $g(\theta) = \theta$, we also notice that $M = K_a^*(J_a^*)^{-1}$, which can be shown to be equal to I_p when $a = 1$. Hence, [Theorem 2](#) can also be seen as an extension of the classic result of empirical likelihood theory allowing for more complex profiling schemes.

We note that the scaled χ_1^2 result of (7) may be used to form confidence intervals, and also full confidence curves, that better reflect degrees of asymmetry in the distribution of the MHL estimator for small and moderate n .

In certain situations, it can be beneficial to use a plug-in estimator inside the empirical likelihood part of h_n . Say that we, for instance, want to control for robust estimation of the covariance matrix Σ in a distribution. This can be achieved by using the hybrid likelihood framework with estimating function $m(y, \mu, \Sigma) = (y - \mu, (y - \mu)(y - \mu)^T - \Sigma)$ this method however introduces the mean as a second control parameter. A perhaps more intuitive way could be to replace μ with the empirical mean of the observations $\hat{\mu}$ and use $m(y, \sigma^2) = (y - \hat{\mu})(y - \hat{\mu})^T - \Sigma$. The following result allows for such constructions. The proof is given in the supplementary material.

Theorem 3. Assume the conditions given in the supplementary material hold true and that $\hat{P} \in \mathbb{R}^r$ is used in construction of the empirical likelihood function and satisfies $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_p(1/\sqrt{n})$ for some function $h: \mathbb{R}^{d+r} \rightarrow \mathbb{R}^r$. Then the MHL estimator $\hat{\theta}_a(\hat{P})$ is consistent for $\theta_a^* = \theta(P_0)$, the minimizer of $E\psi(Y, \theta, P_0)$, and

$$\hat{\theta}_a(\hat{P}) = \hat{\theta}_a(P_0) + \theta'(P_0)(\hat{P} - P_0) + o_{\text{pr}}(1/\sqrt{n}). \quad (8)$$

Remark 2. Although $\theta'(P_0)$ is well-defined, it can be hard to compute in practice as $\theta(P)$ is only implicitly defined. The implicit function theorem can, however, be used to find an explicit expression for the quantity.

Let V be the Jacobian matrix of $\theta \mapsto -Eh(Y, P)$ at P_0 . Then 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 \(2000, Th. 5.21\)](#). Furthermore, $\hat{\theta}_a(P_0)$ satisfies (6) with $m(y, \mu)$ replaced by $(y, \mu) \mapsto m(y, \mu, P_0)$ by [Theorem 1](#). Hence, the central limit theorem guarantees that $\sqrt{n}[\hat{\theta}_a(P_0) - \theta(P_0)]$ and $\sqrt{n}(\hat{P} - P_0)$ converge jointly in distribution to a central normal distribution with covariance matrix

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

where $V = E[\partial_{P_0} h(Y, P)]$ and C_θ is the covariance between $\partial_{\theta_a^*} \psi_P(Y, \theta, P_0)$ and $h(Y, P_0)$ where $Y \sim F$. This together with (8) ensures that $\sqrt{n}[\hat{\theta}_a(\hat{P}) - \theta(P_0)]$ converges to a normal distribution with mean zero and variance equal to $(I_p, \mu'(P_0)^T) \Sigma_\theta (I_p, \mu'(P_0)^T)^T$.

3.2. The hybrid divergence

By [Theorem 1](#), the maximum hybrid likelihood estimator converges in probability to the maximizer of $E\psi(Y, \theta) = (1 - a)E[\log f_\theta(Y)] - aE\{\log[1 + \lambda(\theta)^T M(\theta)]\}$. 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 $E\psi(Y, \theta)$ 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 $-E\{\log[1 + \lambda(\theta)^T M(\theta)]\}$. Using the Lagrange method of multipliers, one can show that for each fixed μ , the expression is the solution to the following optimization problem:

$$\max_{w: \mathbb{R}^d \rightarrow \mathbb{R}} \{E[\log w(Y)] : E[w(Y)M(\mu)] = 0 \text{ and } E[w(Y)] = 1\}. \quad (9)$$

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 (9) is equivalent to minimizing $\text{KL}(g, h)$ with respect to h under the condition that h is a pmf/pdf and $E_h[M(\mu)] = 0$. Here $\text{KL}(g, h)$ 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 $E\psi(Y, \theta)$, i.e. the limit of the MHL estimator, is also the minimizer of $\theta \mapsto d_a(g, f_\theta)$ where

$$d_a(g, h) = (1 - a)\text{KL}(g, h) + a \min_h \{\text{KL}(g, h) : E_h\{M[\mu(h)]\} = 0\}, \quad (10)$$

with $\mu(h)$ being the solution to the estimating equation $E_h[M(\mu)] = 0$. The value of $KL(g, h) > 0$ is zero when $g \neq h$. Since also 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 (10) is small when the overall model fit is good, i.e. when h is close to g overall. The second term, on the other hand, does not emphasize 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 1, the MHL estimator converges to θ_a^* such that $f_{\theta_a^*}$ 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 balance parameter.

The expression in (10) makes it clear in what way the hybrid likelihood framework achieves robustness. From the expression we see that what the method aims to do, is to fit models in a way ensuring that the estimated control parameter is close to its true value. This is in line with the paper’s definition of robustness, i.e. related to behaviour outside of model conditions rather than sensitivity to outliers in the data set.

The divergence d_a relates the hybrid likelihood procedure to other robust divergence based methods like Basu et al. (1998), Ferrari and Yang (2010), Ferrari and La Vecchia (2012). Notice also that d_a ensures robustness through the second term in (10), which can be estimated without nonparametric density estimation.

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, and should not be chosen in an automatic way. 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 based on the focused information criterion.

4.1. FIC for models fitted with hybrid likelihood

The focused information criterion or the FIC for short, ranks models by the quality of their estimate of the focus parameter. See Claeskens and Hjort (2008) for an introduction. We propose to use the information criterion to choose the optimal value for the balance parameter a . This idea was briefly visited in Hjort et al. (2018), where the authors derive formulas for an FIC for models in shrinking neighbourhoods of the true underlying density. Their version of the information criterion therefore requires the model to be asymptotically correctly specified. We will instead create a fully model-agnostic FIC, based upon the formulas and results of Jullum and Hjort (2017) with important modifications in Dæhlen et al. (2024). These articles find asymptotically unbiased approximations to the mean squared error of estimators of the focus parameter. We state a version of the results of Jullum and Hjort (2017) and Dæhlen et al. (2024) valid in the hybrid likelihood context.

Theorem 4. Fix a balance parameter a and let $\hat{\theta}_a$ be the MHL estimator. Let $\xi \in \mathbb{R}$ be some focus parameter and $\hat{\xi}$ a consistent estimator of ξ satisfying

$$\hat{\xi} = \xi + n^{-1} \sum_{i=1}^n \phi(Y_i, \xi) + \epsilon_n \quad (11)$$

for some function $\phi: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\epsilon_n = O_{pr}(1/n)$. Assume further that all conditions of Theorem 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 $g(\hat{\theta}_a)$ in expected value where $\hat{b} = g(\hat{\theta}_a) - \hat{\xi}$, $c = \lim_{n \rightarrow \infty} nE(\epsilon_n)$,

$$\kappa = \text{Var}[\nabla g(\theta_a^*)^T (J_a^*)^{-1} \psi_\theta(Y, \theta_a^*) - \phi(Y, \xi)] \text{ and } \tau = (J_a^*)^{-1} K_a^* (J_a^*)^{-1},$$

provided the assumptions of Jullum and Hjort (2017) and Dæhlen et al. (2024) hold.

The crucial part of Theorem 4 is the fact that $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$ is only $o(1/n)$ away from the true MSE of $g(\hat{\theta}_a)$ in expected value. By Theorem 1 the variance of $\hat{\theta}_a$ 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 unknown. To approximate the MSE of the MHL estimator using Theorem 4, 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 4](#). We define the following approximation to the MSE of $g(\hat{\theta}_a)$,

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

where $\hat{c}$ is either of the estimators given in [Dæhlen et al. \(2024\)](#), $\hat{\tau} = \nabla g(\hat{\theta}_a)^T (\hat{J}_a^*)^{-1} \hat{K}_a^* (\hat{J}_a^*)^{-1} \nabla g(\hat{\theta}_a)$ and $\hat{\kappa}$ is given by $\hat{\tau} + n^{-1} \sum_{i=1}^n \phi(Y_i, \hat{\xi})^2 - 2\nabla g(\hat{\theta}_a)^T (\hat{J}_a^*)^{-1} \hat{V}$, with $\hat{J}_a^* = n^{-1} H h_n(\hat{\theta}_a)$, $\hat{V} = n^{-1} \sum_{i=1}^n \nabla_{\hat{\theta}_a} \hat{\psi}_i(\theta) \phi(Y_i, \hat{\xi})^T$ and $\hat{K}_a^* = n^{-1} \sum_{i=1}^n \nabla_{\hat{\theta}_a} \hat{\psi}_i(\theta) \nabla_{\hat{\theta}_a} \hat{\psi}_i(\hat{\theta}_a)^T$. Here $\hat{\psi}_i(\theta)$ is equal to $\hat{\psi}(Y_i, \theta)$ with $\lambda(\theta)$ replaced by $\hat{\lambda}$, the root of $V_n(\lambda, \theta) = \sum_{i=1}^n M_i(\theta) / [1 + \lambda^T M_i(\theta)]$, and $\lambda'(\theta)$ by $-(S_n^1)^{-1} S_n^2$ where $S_n^1 = \partial_{\lambda} V_n(\lambda, \hat{\theta}_a)$ and $S_n^2 = \partial_{\theta} V_n(\theta, \hat{\lambda})$.

Remark 3. 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 5](#).

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 $E[\psi_{\theta}(Y, \theta_a^*) \phi(Y)]$ respectively under weak conditions. See e.g. Lemma 16 of [Dæhlen et al. \(2024\)](#) 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 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.

The above HFIC is only valid when plug-in estimators are not present. Thanks to [Theorem 3](#), 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 3](#) 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 (11) for some ϕ . Define $\psi_F(y, \theta, \lambda, P) = (m(y, \mu, P) / [1 + \lambda^T m(y, \mu, P)], \nabla_{\theta} \psi_P(y, \theta, P)^T, h(y, P)^T)^T$, and let $J\psi_F(y, \theta, \lambda, P)$ denote its Jacobian matrix with respect to all parameters but y . Let $\hat{A} = n^{-1} \sum_{i=1}^n J\psi_F(Y_i, \hat{\theta}_a, \hat{\lambda}, \hat{P})$ and $\hat{B} = n^{-1} \sum_{i=1}^n \psi_F(Y_i, \hat{\theta}_a, \hat{\lambda}, \hat{P}) \psi_F(Y_i, \hat{\theta}_a, \hat{\lambda}, \hat{P})^T$ and $\hat{C} = \hat{A}^{-1} \hat{B} (\hat{A}^{-1})^T$. With this we define the HFIC as in [Definition 1](#) with $\hat{b} = g[\hat{\theta}_a(\hat{P})] - \hat{\xi}$, $\hat{c}$ equal to either of the estimators given in [Dæhlen et al. \(2024\)](#) and $\hat{\tau} = (\nabla g[\hat{\theta}_a(\hat{P})], 0)^T \hat{C} (\nabla g[\hat{\theta}_a(\hat{P})], 0)^T$ and $\hat{\kappa}$ equal to

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

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

5. 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 the academic community, see e.g. [Taleb and Cirillo \(2019\)](#), [Cirillo and Taleb \(2016\)](#). 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) inter-state wars from the Correlates of War data set ([Sarkees and Wayman, 2010](#)). [Cunen et al. \(2020\)](#) argue that the Korean War marks a change in the number of battle deaths, and for this reason 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. [Dæhlen et al. \(2024\)](#) find that a shifted log-normal model outperforms other options for answering this question. 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 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. 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](#). 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 et al. \(2024\)](#), in particular their Appendix A. The results are shown in [Fig. 2](#). We see that neither minima are attained at 0 or 1, i.e. a genuine hybrid method is chosen for both data sets, with the optimal balance parameters being $a = 0.48$ for the older wars and $a = 0.17$ for the newer conflicts.

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) 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, by applying [Theorem 1](#) in conjunction 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

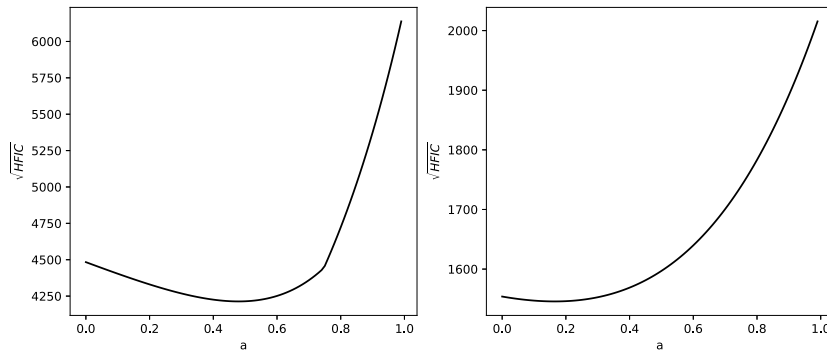


Fig. 2. The figure displays root-HFIC scores as functions of a (the balance parameter) for older and newer wars, in the left and right plots respectively. The minimum was attained at $a = 0.48$ for the older wars and $a = 0.17$ for the newer conflicts.

more or less equal to the that 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 the number of battle deaths between newer and older wars. The MHL estimator was found to be 3441, which, relatively speaking, is a lot larger than the corresponding ML estimator of 2066. Standard deviations for these quantities were computed using the delta method and Theorem 1, and were computed to 3815 and 3155 respectively. With these numbers we obtained the following 95% confidence intervals: $[-4036, 10919]$ and $[-4135, 8232]$ for hybrid and maximum likelihood respectively. Similarly, we can also tested whether the number of battle deaths has stayed the same against the alternative that violence has declined, and obtained the p -values 0.18 and 0.26 when using hybrid and maximum likelihood respectively. 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, we can conclude that the larger wars have indeed become less deadly, which is consistent with the findings of Cunen et al. (2020).

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 .

6. Concluding remarks

We have worked and significantly extended the hybrid likelihood framework of Hjort et al. (2018), proving that many famous theorems for the ML estimator (consistency, asymptotic normality, Wilks' theorem) can be extended to the more general hybrid likelihood theory. Our theory also allowed for the presence of plug-in parameters. We also proposed a way of choosing the balance parameter based on the FIC. We will now briefly go through some limitations of our work and five possible further research topics.

(1): The perhaps most pressing issue with the theory developed in this article was discussed in Remark 1 following Theorem 1. For each fixed μ in a neighbourhood of μ^* , we require the estimating function to have bounded support. As is shown in Schennach (2007, Th. 1), such a condition is necessary for the EL-framework to function outside of model conditions. An interesting further research topic could be to investigate whether it would be possible to allow the support of Y to slowly increase with n , easing the strict condition of bounded estimating functions.

(2): 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 . Further investigation of such settings, perhaps based on theory in van der Vaart and Wellner (1997) for maximizers of non-smooth functions could be promising, especially since such estimators often exhibit slower rates of convergences than $O_p(1/\sqrt{n})$.

(3): In the beginning of Section 3.1, we proved that .

$$h_n(\theta) = \sum_{i=1}^n \psi(Y_i, \theta) + V_n[\mu(\theta)]^T S([\mu(\theta)])^{-1} V[\mu(\theta)] + o_p(1), \quad (12)$$

where $\psi(y, \theta) = (1 - a) \log f_\theta(y) - a \log\{1 + \lambda[\mu(\theta)]^T m(y, \mu(\theta))\}$. We used this result to give intuition about the hybrid likelihood function, but the alternative characterization could be used for other purposes. Using the expansion to formulate estimation of $\hat{\theta}_a$ as an M-estimation problem instead of a Z-estimation one is one idea. Furthermore, investigating whether one can develop tests for checking if the parametric model holds true based on (12) and arguments in White (1982) seems promising.

(4): A further topic for further research could be to explore using hybrid likelihoods to incorporate auxiliary information as in Qin (2000). For data $(Y_1, X_1), \dots, (Y_1, X_1)$ and $X_{m+1}, \dots, X_n$ where a parametric likelihood is available for $Y \mid X$ only, but where we know $Em(X, \mu) = 0$ for some μ , this could be achieved by looking at

$$h_n = (1 - a)\ell_m(\theta) + a \log \text{EL}_n[\mu(\theta)]$$

where $\ell_m(\theta) = \sum_{i=1}^m \log f_\theta(Y_i \mid X_i)$ and EL_n is the empirical likelihood for μ based on the estimating function m and data $X_1, \dots, X_n$. With $a = 1/2$ this is equal to the model of Qin (2000), but by varying the balance parameter one can decide how much weight should be put on the parametric fit compared to the auxiliary information.

(5): Our HFIC of Section 4 was constructed to help us find the best balance parameter for a given situation and for the parametric model in use. Since HFIC has direct interpretation as the estimated mean squared error, however, the method immediately lends itself to model selection setups, with a list of parametric candidate models, as long as the focus parameter ξ has a clear statistical interpretation across models. Thus our recipes properly generalize the model selection methods of Jullum and Hjort (2017) and Dæhlen et al. (2024), which for the most part deal with maximum likelihood estimation, i.e. the case of using $a = 0$ for each model.

Acknowledgements

This research has been partially funded by the Norwegian Research Council, Norway 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.

Appendix A. Supplementary data

The supplementary material contains proofs of all theorems, full sets of regularity conditions, some simulations and an application with multivariate data, pertaining to the joint distribution of body weight and brain weight for $n = 56$ mammals.

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jspi.2025.106327>.

References

- 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.
- Chib, S., Shin, M., Simoni, A., 2018. Bayesian estimation and comparison of moment condition models. *J. Amer. Statist. Assoc.* 113 (524), 1656–1668.
- Cirillo, P., Taleb, N.N., 2016. On the statistical properties and tail risk of violent conflicts. *Phys. A* 452, 29–45.
- Claeskens, G., Hjort, N.L., 2003. The focused information criterion. *J. Amer. Statist. Assoc.* 98 (464), 900–916.
- Claeskens, G., Hjort, N.L., 2008. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, England.
- Clauset, A., 2018. Trends and fluctuations in the severity of interstate wars. *Sci. Adv.* 4, eaao3580.
- Cunneen, C., Hjort, N.L., Nygård, H.M., 2020. Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *J. Peace Res.* 57 (2), 221–234.
- Dæhlen, I., 2021. *Empirical and Hybrid Likelihood* (Master thesis). University of Oslo.
- Dæhlen, I., Hjort, N.L., Hobæk Haff, I., 2024. Accurate bias estimation with applications to focused model selection. *Scand. J. Stat.* 51 (2), 724–759.
- Ferrari, D., La Vecchia, D., 2012. On robust estimation via pseudo-additive information. *Biometrika* 99 (1), 238–244.
- Ferrari, D., Yang, Y., 2010. Maximum L_q -likelihood method. *Ann. Statist.* 38 (2), 753–783.
- Hjort, N.L., McKeague, I.W., Van Keilegom, I., 2009. Extending the scope of empirical likelihood. *Ann. Statist.* 37 (3), 1079–1111.
- Hjort, N.L., McKeague, I.W., Van Keilegom, I., 2018. Hybrid combinations of parametric and empirical likelihoods. *Statist. Sinica* 28 (4), 2389–2407.
- Huber, P.J., Ronchetti, E.M., 2011. *Robust Statistics*. John Wiley & Sons.
- Jullum, M., Hjort, N.L., 2017. Parametric or nonparametric: The FIC approach. *Statist. Sinica* 27 (3), 951–981.
- Molanes-Lopez, E.M., Van Keilegom, I., Veraverbeke, N., 2009. Empirical likelihood for non-smooth criterion functions. *Scand. J. Stat.* 36 (3), 413–432.
- Owen, A.B., 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75 (2), 237–249.
- Owen, A.B., 2001. *Empirical Likelihood*. CRC Press, Boca Raton, FL.
- 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*. Viking, Toronto.
- Qin, J., 1994. Semi-empirical likelihood ratio confidence intervals for the difference of two sample means. *Ann. Inst. Statist. Math.* 46 (1), 117–126.
- Qin, J., 2000. Combining parametric and empirical likelihoods. *Biometrika* 87 (2), 484–490.
- Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. *Ann. Statist.* 22 (1), 300–325.
- Sarkees, M.R., Wayman, F., 2010. *Resort to War: A Data Guide to Inter-state, Extra-state, Intra-state, and Non-state Wars, 1816–2007*. CQ Press, Washington DC.
- Schennach, S.M., 2007. Point estimation with exponentially tilted empirical likelihood. *Ann. Statist.* 35 (2), 634–672.
- Schweder, T., Hjort, N.L., 2016. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.
- Shao, J., 2003. *Mathematical Statistics*. Springer, New York.
- Stigler, S.M., 1977. Do robust estimators work with real data? *Ann. Statist.* 5 (6), 1055–1098.
- Taleb, N.N., Cirillo, P., 2019. The decline of violent conflicts: What do the data really say? In: Toje, A., Bård, N.V.S. (Eds.), *The Causes of Peace: What We Know Now*. Nobel Symposium Proceedings. pp. 57–85.
- van der Vaart, A.W., 2000. *Asymptotic Statistics*. Cambridge University Press, Cambridge, England.
- van der Vaart, A.W., Wellner, J.A., 1997. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.
- White, H., 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50 (1), 1–25.