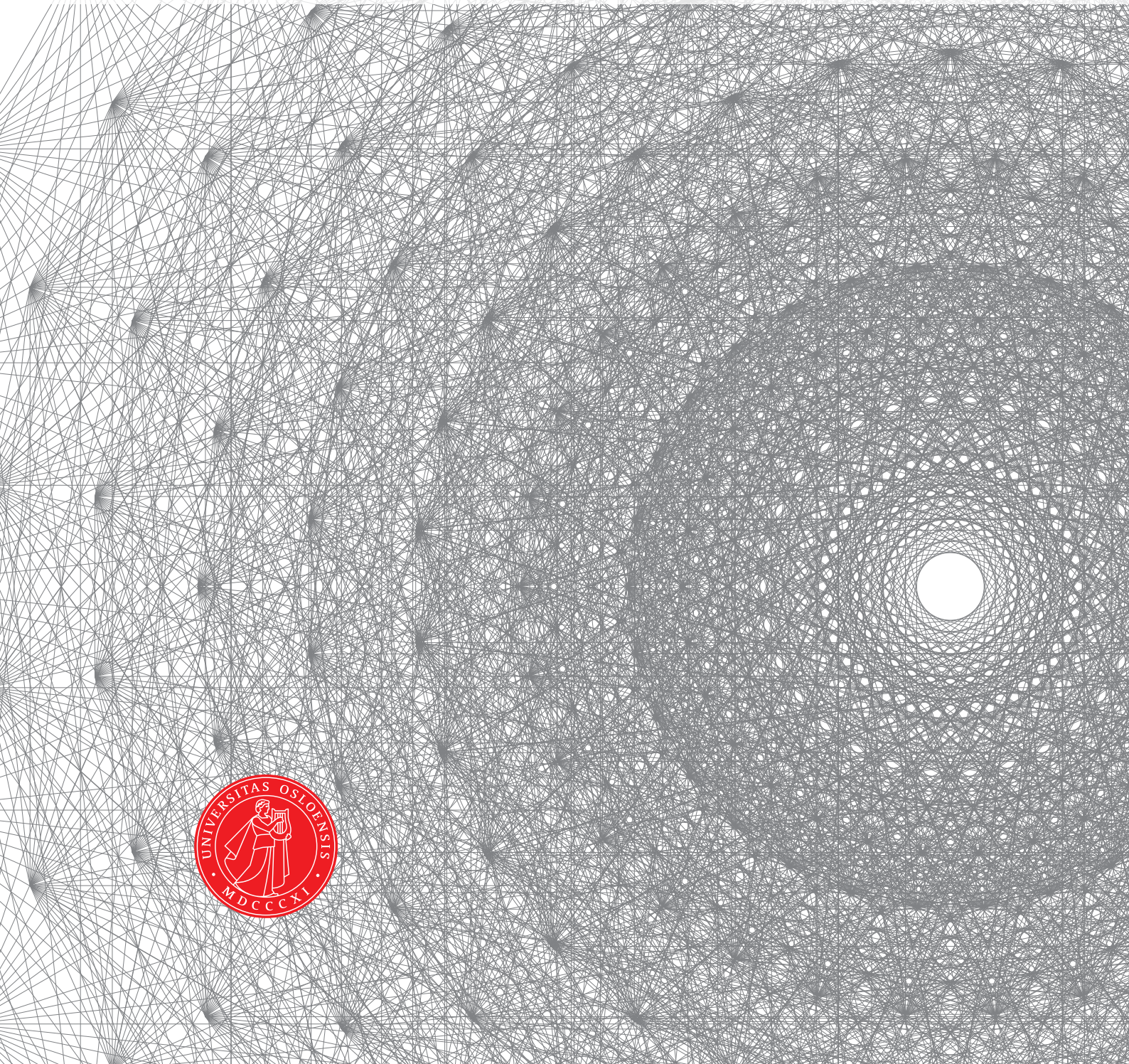
UiO : Department of Mathematics
University of Oslo

Empirical and Hybrid Likelihood

Ingrid Dæhlen

Master's Thesis, Spring 2021



This master's thesis is submitted under the master's programme *Stochastic Modelling, Statistics and Risk Analysis*, with programme option *Statistics*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

In this thesis, we work with empirical likelihood functions and hybrid combinations of these with parametric likelihoods. We state and prove an analogue to Wilks theorem for the empirical likelihood function. In addition, we derive an alternative characterization of the map, and use it to reformulate maximization of the empirical likelihood function as an M-estimation problem. We also work with the hybrid likelihood function, a combination of empirical and parametric likelihoods. We prove a profiling result for this map and investigate the case of possible model misspecification. The limit distribution of the maximizer of the hybrid likelihood function is derived in this situation. In addition, we define a focused information criterion for hybrid likelihood and use it to propose a method for selecting the tuning parameters involved in the definition of the hybrid likelihood function.

Acknowledgements

This thesis concludes a total of six years of study at the University of Oslo. I started out studying mathematics, but, after a brief detour into comic book drawing, ended up with statistics. I have wavered and doubted. My career path and study programs have changed multiple times. Through all the years, however, my love for limits has remained unchanged and unparalleled. I am glad this has been a recurring theme in my thesis, and I am proud of every epsilon.

First and foremost, I would like to thank my supervisor, Nils Lid Hjort. Both for introducing me to the topic of this thesis as well as all discussions and help along the way. Our meetings were always interesting and informative.

Secondly, I would like to thank all my friends here at the University of Oslo. Thank you to everyone in my study hall. The lunches and breaks were essential for motivation. A special thanks to Lars O. and Peder for laughing at my jokes and listing to my complaints. I am also indebted to my DnD group and to the members of the student association Realistforeningen. You provided much needed laughs and stress relief.

Thank you to my family for listing to my endless blabbering about a field you do not know. A special thanks goes out to my father for all help with the writing. I am lucky to have a dad willing to discuss the use of "neither" at 11 pm. And thank you to Lars D. for always supporting me. You are more patient and kind than I ever will be.

Last, but not least, I want to thank everyone who has proofread my emails. You know who you are and how essential you have been.

Contents

CHAPTER 1

Introduction

In the field of statistics, we want to infer information from data. For instance, we might want to estimate the mean or median pay in a country or whether a medicine has an effect or not. What questions we ask and how they are formulated are, of course, dependent on the specific problem or data set. That being said, we are usually interested in properties of the underlying distribution. A common statistical approach is to estimate this distribution. There are many ways to proceed, but a simple and popular method is to fit a parametric family to the data. This approach is both practical and theoretically convenient, but in many cases, it can be hard or even impossible to find a fitting class of distributions. Furthermore, some situations may require our results to be unaffected by prior assumptions. This motivates the study of non-parametric statistics, and in this thesis, we will be concerned with one particular strategy: empirical likelihood.

Empirical likelihood was first introduced by Art B. Owen in Owen 1988. In this article, the author defines what he calls the empirical likelihood function. The map can be created non-parametrically, and Owen showed that the quantity has a limiting chi squared distribution when evaluated at the true parameter. This result holds under very general assumptions and allows for construction of confidence intervals and regions in a non-parametric way. In addition, it justifies the use of the word "likelihood ". The limit distribution of the empirical likelihood function at the true parameter is the same as that of the likelihood ratio function used in parametric likelihood theory. Thus, the results of Owen 1988 allowed statisticians to work with something resembling the likelihood ratio test non-parametrically and under very weak assumptions.

Soon after the release of the original article, Owen published an additional text, Owen 1991. Here, he extends the original definitions and theorems to multidimensional parameters and regression settings. In the same year, Barlett correctability of empirical likelihood was shown in DiCiccio, Hall, and Romano 1991. This was an important article, as the use of Barlett corrections makes the coverage error of confidence regions constructed with empirical likelihood asymptotically smaller than that of many other nonparametric methods like bootstrap or jackknife. Three years later, Qin and Lawless 1994 extended the general theory to multidimensional estimating equations of general dimensions and proved several non-parametric versions of important theorems from parametric likelihood theory. Examples include consistency and asymptotic normality of the maximizer of the empirical likelihood function. This made the parallels between the empirical and parametric likelihood even

clearer.

Because of its elegant and general definition, empirical likelihood theory has become quite popular and generated a lot of literature. A recurring theme is the extension of the results and definitions to situations different from those described by Owen. This has been done by numerous authors, and we will not attempt to list them all, but two important examples include Mykland 1995 and Kitamura 1997 providing results for and different ways of dealing with dependent data. For a more comprehensive overview of sources concerning empirical likelihood we refer to Owen 2001. This textbook summarizes a lot of the theory and literature concerning empirical likelihood and is written by Art B. Owen himself.

Rather than focusing on a specific situation and showing how the empirical likelihood machinery can be applied, the goal of this thesis will be to expand the fundamental theory. We will do this in both a practical and a theoretical manner. First, we will provide a result extending the class of parameters we can make inference about using empirical likelihood. Afterwards, we will present theory regarding an alternative characterization of the empirical likelihood function. This will be done in part I of the thesis. In the second part, we will use what we have derived to expand on the ideas concerning hybrid combinations of parametric and empirical likelihoods from Hjort, I. McKeague, and Van Keilegom 2018. We will now describe and motivate each of these steps. This will also serve as an outline of the thesis.

The results of Qin and Lawless 1994 allow us to use empirical likelihood when constructing confidence intervals for, and conducting hypothesis tests about, parameters, $\mu \in \mathbb{R}^p$, which can be expressed as the solution to equations on the form

$$\mathrm{E}\, m(Y, \mu) = 0,$$

for some function $m\colon \mathbb{R}^{d+p} \to \mathbb{R}^q$ and $Y \in \mathbb{R}^d$ following the same distribution as the data. This is a quite general class of parameters and includes e.g. moments and the median. That being said, there are many quantities that cannot be expressed in this manner. Two simple examples are the standard deviation or the ratio of two means. In Chapter 3, we will develop a way to make inference about quantities like these. We will do this by formulating and proving a profiling result for the empirical likelihood function. Using this we can make inference about, not only solutions to estimating equations, but functions thereof. The result has clear parallels to Wilks theorem used in parametric likelihood theory. In addition, it serves as an elegant way to get rid of nuisance parameters. This is a problem that has been considered in articles like Qin and Lawless 1994, Hjort, I. W. McKeague, and Keilegom 2009 and Molanes Lopez, Van Keilegom, and Veraverbeke 2009.

In Chapter 5 we will take a second look at the definition of the empirical likelihood function and find an asymptotically equivalent expression for the quantity. This will allow us to rediscover the main result of Owen 1988. In addition, it provides some intuition and knowledge about the empirical likelihood function at values other than the true parameter. In particular, we will show that the logarithm of the empirical likelihood function, divided by the sample size, is asymptotically equivalent to a mean. To our knowledge, this material is new, and in Chapter 6 we use the results to prove consistency and asymptotic

normality of the maximizer of the empirical likelihood function. Such results have been shown before, most famously by Qin and Lawless 1994, but our method of proof will be very different from theirs and based on the alternative characterization from Chapter 5.

Empirical likelihood is a useful tool. It allows for construction of confidence intervals for very general quantities, assuming little about the distribution of the data. This is, of course, a strength, but in some situations, we might know, or at least have a good idea about, how the data is distributed. Such information can strengthen our results significantly and should be incorporated in the analysis. There already exists multiple ways to combine parametric and non-parametric methods, see e.g. Hjort and Glad 1995 or Olkin and Spiegelman 1987, but in this thesis we will work with combinations of parametric and empirical likelihoods. Some amount of research on this topic exits. Qin 1994 considered a situation with two data sets where a parametric model is available for one sample only. In Qin 2000 multiplication of a conditional with an empirical likelihood is investigated, and Qin and Wong 1996 provides results for a situations where a parametric model exists only when the variables take certain values. In this thesis, however, we will consider a newer approach involving the hybrid likelihood function, introduced by Hjort, I. McKeague, and Van Keilegom 2018. As mentioned previously, this will be the topic for the second part of the thesis.

Hybrid likelihood theory was developed in Hjort, I. McKeague, and Van Keilegom 2018 and enables the simultaneous use of parametric and empirical likelihood to make inference. In Chapter 7 we will present the main theory from the paper and formulate and prove a profiling result for the hybrid likelihood function. This allows for construction of non-symmetric confidence intervals of focus parameters and will be based on similar ideas as those used to prove the limit result concerning the profile empirical likelihood function in Chapter 3. We will also provide some examples illustrating how the theory can be applied.

The theorems proved in Hjort, I. McKeague, and Van Keilegom 2018 are mostly concerned with the behavior of the hybrid likelihood function and its maximizer when the true distribution is a member of the parametric model used in construction of the likelihood. This is mathematically convenient and ensures natural limits for many quantities. The results do, however, not hold true when the model is specified incorrectly. In Chapter 8 we will use the results of Chapter 5 and Chapter 6 to investigate what happens in this situation. We will discuss what the maximizer of the hybrid likelihood function is aiming for and prove consistency towards this quantity. In addition, we will derive the limit distribution of the maximizer.

When combining parametric and empirical likelihood, we need to choose how much credit each should be given. If the model is specified correctly, standard theory guarantees maximum likelihood is asymptotically the most efficient way to proceed. When the true distribution is not a member of the parametric family, however, empirical likelihood should be given additional weight to ensure robustness. How to choose this balance of power is not a trivial question. In Chapter 9, we will discuss how this can be done and provide algorithms for choosing how much weight should be put on the empirical and parametric part of the hybrid likelihood function. To do this, we will introduce a model selection tool called the focused information criterion and make adjustments to the hybrid likelihood situation. In Chapter 10, we will go though some examples illustrating the results of Chapter 8 and Chapter 9 and show how they can be

used to make inference.

We have chosen not to include code in the thesis. Instead, algorithms are described or references with implementation details cited. This might conceal the extent of programming involved in production of this thesis, as most of the algorithms have been implemented from scratch by the author. Code can be obtained upon request, should it be of interest. All figures and scripts have been created with the Anaconda distribution of Python (Anaconda Inc. 2020).

1.1 Remarks on notation

Before we begin, we will introduce some notation that will be used throughout the thesis.

We will differentiate between Jacobian matrices and gradients. For a function $f\colon \mathbb{R}^p \to \mathbb{R}^q$, both $f'(y)$ and $\partial f(y)/\partial y$ will denote the $q \times p$ Jacobian matrix evaluated at y. For $g\colon \mathbb{R}^p \to \mathbb{R}$, $\nabla g(y)$ will be the gradient of g at y. This is a vector in $\mathbb{R}^p$ and the transpose of the Jacobian matrix of the function.

We will also adopt some commonly used notation for different modes of convergences of random variables. Let Y_n be a sequence of random variables. The expression $Y_n = o_{\Pr}(a_n)$ means that Y_n/a_n converges to 0 in probability. Similarly, $Y_n = O_{\Pr}(a_n)$ if Y_n/a_n is bounded in probability, i.e. for every $\epsilon > 0$ there exists $M > 0$ such that

$$\sup_{n \in \mathbb{N}} \Pr(\|Y_n/a_n\| > M) < \epsilon.$$

Furthermore, $Y_n \xrightarrow{a.s.} Y$, $Y_n \xrightarrow{\Pr} Y$ and $Y_n \xrightarrow{d} Y$ are notation for Y_n converging to the random variable Y almost surely, in probability and in distribution respectively. For definitions and explanation of these concepts, see e.g. chapter 2 of Vaart 1998. Lastly, $\overset{d}{\approx}$ will be short-hand for "approximately distributed as". For instance, $Y \overset{d}{\approx} \chi^2_1$ means that Y_n is approximately chi squared distributed with one degree of freedom.

Unless stated otherwise, $\|\cdot\|$ will always denote the euclidean norm. This is also true for matrices. If $A = (a_{i,j})_{i,j}$ is a $p \times q$-matrix,

$$\|A\| = \left(\sum_{i,j} |a_{i,j}|^2 \right)^{1/2}.$$

In particular, a sequence of random matrices A_n converges to a limit A if the convergence happens with respect to the euclidean norm.

PART I

Empirical Likelihood

CHAPTER 2

The definition and main result

Empirical likelihood is a non-parametric way of both estimating and constructing confidence intervals for certain quantities. The machinery can be applied in very general settings without making strict assumptions about the distribution of the data. The definition and main theorems were first stated and proved by Art B. Owen in Owen 1988. Since then, the framework has been both extended and improved upon by several authors. In this chapter we will introduce the main concepts and theorems concerning empirical likelihood. The two first sections will follow Owen 2001 closely, while we in Chapter 3 formulate and prove a result concerning profiling of the empirical likelihood function. In Chapter 4 we will provide some examples illustrating the theory.

2.1 The definition

We start by defining empirical likelihood. Let $Y_1, \ldots, Y_n \in \mathbb{R}^d$ be independent and identically distributed random variables, following some distribution with cumulative distribution function, F. Assume we are interested in some parameter, $\theta \in \mathbb{R}^p$, of this distribution. In maximum likelihood theory one assumes that F is a member of a parametric family indexed by some parameter. When using empirical likelihood, on the other hand, we do not make such presumptions about F. Instead we assume that there exists a function $m\colon \mathbb{R}^{d+p} \to \mathbb{R}^q$ such that θ can be characterized in the following way:

$$\mathrm{E}\, m(Y, \theta) = 0, \tag{2.1}$$

where $Y \sim F$. (2.1) is called the estimating equation and m is referred to as the estimating function.

(2.1) is a very general equation. Because of this, many interesting quantities can be characterized using such expressions. One immediate example is $\mathrm{E}(Y - \theta) = 0$. This equation is solved by $\theta = \mathrm{E}\, Y$ and is a special case of (2.1) with $m(y, \theta) = y - \theta$. In Section 4.1 we will make inference about the mean yearly income in Oslo using empirical likelihood with this estimating function. Another, perhaps less obvious, example is obtained with $m(y, \theta) = I(y \leq \theta) - 0.5$. Entering this function into (2.1), results in the equation $\Pr(Y \leq \theta) = 0.5$, which is solved by the median when F is a continuous distribution. In Section 4.3 empirical likelihood with this estimating equation will be used to investigate whether the world has become more peaceful or not.

Assume Y follows a continuous distribution whose density belongs to some parametric family f_θ with $\theta \in \Theta \subseteq \mathbb{R}^p$. For many such families the expected

value of the score function evaluated at the true parameter is zero. Hence, with

$$m(y, \theta) = \frac{\partial}{\partial \theta} \log\, f_\theta(y),$$

(2.1) is solved by θ_0 such that the true density of Y is given by f_{θ_0}. An example illustrating this with simulated data can be found in Section 4.2.

In the examples above θ was the unique solution to the estimating equations and q, the dimension of $m(y, \theta)$, was equal to p, the dimension of θ. This is the typical situation but by no means a requirement. The map

$$m(y, \theta) = (y - \theta, (y\theta)^2 - \theta)^T$$

is a perfectly valid estimation function, with

$$\mathrm{E}\, m(Y, \theta_0) = 0$$

when Y is Poisson-distributed with rate parameter θ_0. Furthermore,

$$m(y, \mu, \sigma) = (y - \mu, (y - \mu)^2 - \sigma^2)^T$$

will result in both $(\mu_0, \sigma_0)^T$ and $(\mu_0, -\sigma_0)^T$ being solutions to (2.1). Here μ_0 is the true mean and σ_0 the true standard deviation in the distribution of Y.

We are now ready to define the empirical likelihood function as given in Owen 2001.

Definition 2.1.1 (Owen 2001, p. 41)**.** Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables from some distribution, F, and suppose $\theta \in \mathbb{R}^p$ is a parameter such that

$$\mathrm{E}\, m(Y, \theta) = 0,$$

for some $m \colon \mathbb{R}^{d+p} \to \mathbb{R}^q$ and $Y \sim F$. The empirical likelihood function for θ is defined as

$$\mathrm{EL}_n(\theta) = \max \left\{ \prod_{i=1}^n n w_i \;\middle|\; \sum_{i=1}^n w_i m(Y_i, \theta) = 0,\ \sum_{i=1}^n w_i = 1 \text{ and } w_i \geq 0 \right\}, \tag{2.2}$$

where the last conditions should hold for $i = 1, \ldots, n$.

In the above we assumed that $Y_1, \ldots, Y_n$ were i.i.d. data points. This is not strictly necessary and generalizations of empirical likelihood to non-i.i.d. situations exists, see e.g. chapter 4 and 8 of Owen 2001. In the remainder of this chapter we will however be satisfied with the above definition as it both simplifies presentation and is sufficient for many interesting applications.

2.2 Alternative characterizations

Definition 2.1.1 is compact and elegant, but at first glance it might look slightly incomprehensible. Before we move on to the main result of empirical likelihood theory, we will therefore present two alternative characterizations of $\mathrm{EL}_n(\theta)$.

The empirical likelihood ratio function

The empirical likelihood function is sometimes called the empirical likelihood ratio function. There is a good reason for this, providing insight and additional intuition about Definition 2.1.1.

The last two conditions on the w_i-s in (2.2) are equivalent to requiring the weights to be probabilities in a discrete distribution on $\{ Y_1, \ldots, Y_n \}$. The first condition ensures that the discrete distribution with weights $w_1, \ldots, w_n$ satisfies the estimating equation (2.1) with the chosen m. In addition, the product of the w_i is the non-parametric likelihood of the distribution. (2.2) can therefore be reformulated as

$$\mathrm{EL}_n(\theta) = \max_F \left\{ n^n L(F) \,\middle|\, \mathrm{E}_F\, m(Y, \theta) = 0,\ F \in \mathcal{F} \right\},$$

where $\mathcal{F}$ is the set of discrete distributions on $\{ Y_1, \ldots, Y_n \}$, $L(F)$ is the likelihood, i.e. $\prod_{i=1}^n \Pr(Y = Y_i)$ for $Y \sim F$ and $\mathrm{E}_F\, m(Y, \theta)$ denotes the expected value of $m(Y, \theta)]$ when $Y \sim F$.

If we do not require that the estimating equation is satisfied, $L(F)$ is maximized by the empirical distribution function. A proof of this can be found in Owen 2001, p. 8. The empirical distribution his has likelihood $(1/n)^n$. Hence,

$$\mathrm{EL}_n(\theta) = \max_F \{ R(F) \mid \mathrm{E}_F\, m(Y, \theta) = 0,\ F \in \mathcal{F} \}, \tag{2.3}$$

where

$$R(F) = \frac{L(F)}{\max_G L(G)}.$$

$R(F)$ is the non-parametric likelihood ratio. So (2.3) shows that the empirical likelihood (ratio) function can be seen as the result of a profiling of the non-parametric empirical likelihood ratio.

An implicit function

The formula in (2.2) gives little to no information about how EL_n behaves as a function of θ or can be computed in practice. We will therefore explain briefly how the empirical likelihood is typically calculated. Incidentally, this also leads to a very different formula for $\mathrm{EL}_n(\theta)$. This alternative characterization will be used frequently in this thesis.

Firstly, notice that when 0 is not in the interior of the convex hull of $m(Y_1, \theta), \ldots, m(Y_n, \theta)$, $\mathrm{EL}_n(\theta) = 0$ by definition. If, on the other hand, there exists weights $w_1, \ldots, w_n > 0$ satisfying the conditions in (2.2), the maximizer of

$$\prod_{i=1}^n n w_i \quad \text{such that} \quad \sum_{i=1}^n w_i m(Y_i, \theta) = 0$$

is also the maximizer of $\sum_{i=1}^n \log w_i$ under the same condition. This optimization problem can be solved using the method of Lagrange multipliers. After some algebraic efforts, one finds that the solution is given by

$$\mathrm{EL}_n(\theta) = \prod_{i=1}^n \Big(1 + \lambda_n(\theta)^T m(Y_i, \theta) \Big)^{-1}, \tag{2.4}$$

for some $\lambda_n(\theta) \in \mathbb{R}^q$ satisfying

$$0 = \sum_{i=1}^{n} \frac{m(Y_i, \theta)}{1 + \lambda_n(\theta)^T m(Y_i, \theta)}. \tag{2.5}$$

A proof of this will not be given here, but can be found in Owen 2001, pp. 21–21.

Let $\mathcal{C}_n$ denote the interior of the convex hull of $m(Y_1, \theta), \ldots, m(Y_n, \theta)$. Because of the above, we can express the empirical likelihood function in the following way:

$$\mathrm{EL}_n(\theta) = \begin{cases} 0, & 0 \notin \mathcal{C}_n \\ \prod_{i=1}^{n} \Big(1 + \lambda_n(\theta)^T m(Y_i, \theta)\Big)^{-1}, & 0 \in \mathcal{C}_n \end{cases},$$

with $\lambda_n(\theta)$ defined implicitly as a solution to (2.5).

The above representation can be used to compute the value of the empirical likelihood function. To find $\mathrm{EL}_n(\theta)$, we typically solve (2.5) for $\lambda_n(\theta)$. If there is a solution, $\lambda_n(\theta)$, with $1 + \lambda_n(\theta)^T m(Y_i, \theta) > 0$ for $i = 1, \ldots n$, we use (2.4) to compute $\mathrm{EL}_n(\theta)$ otherwise $\mathrm{EL}_n(\theta)$ is set to 0.

In practice, solving (2.4) requires a numerical optimization algorithm. Since the function given in this equation is smooth in λ, there are multiple algorithms to choose from. That being said, checking whether 0 is in the interior of the convex hull of $m(Y_1, \theta), \ldots, m(Y_n, \theta)$ and finding a zero of (2.5) resulting in positive weights summing to one, can be quite complicated. This is particularly difficult when the estimating equation is multidimensional. Luckily, there are methods dealing with these computational issues. We will not go into detail about the specific algorithms in this thesis, but information regarding implementation of the empirical likelihood function can be found in chapter 3.14 in Owen 2001. There a numerical strategy for computation of $\log \mathrm{EL}_n(\theta)$ is discussed in detail.

2.3 The main result

As shown in the previous section, the empirical likelihood function can be seen as a profiled non-parametric empirical likelihood ratio. The deviance function based on profile likelihood in parametric models has a chi-square distributed limit distribution. It is therefore expected that something similar should hold for $\mathrm{EL}_n(\theta)$. This is indeed the case, as can be seen from the following theorem.

Theorem 2.3.1 (Owen 2001, p. 41)**.** *Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables from a distribution, F, and suppose $\theta_0 \in \mathbb{R}^p$ is a parameter such that*

$$\mathrm{E}\, m(Y, \theta_0) = 0,$$

for some $m \colon \mathbb{R}^{d+p} \to \mathbb{R}^q$ and $Y \sim F$. Assume further that $\operatorname{Var} m(Y, \theta_0)$ is finite and has full rank. Then the empirical likelihood function, EL_n, constructed with $Y_1, Y_2, \ldots, Y_n$ and estimating function, m, satisfies

$$-2 \log \mathrm{EL}_n(\theta_0) \xrightarrow{d} \chi^2_q.$$

This was first shown by Art B. Owen, and a proof can be found in Owen 2001, pp. 219–222.

As remarked previously, the i.i.d. assumption on the data in the definition of the empirical likelihood function can be relaxed somewhat. The same is true for Theorem 2.3.1. There are numerous articles dealing with versions of the above theorem for non i.i.d. cases. Among the most famous are Owen 1991, for linear regression settings, and Kitamura 1997, providing limits for the empirical likelihood function with certain types of dependent data. For a general overview, we again refer to chapter 4 and 8 in Owen 2001.

Theorem 2.3.1 can be seen as the main result of empirical likelihood theory, and can be used to construct non-parametric approximate confidence intervals and regions for almost all θ-s that can be characterized as solutions to estimating equations. To give an indication about what Theorem 2.3.1 can be used to make inference about, we will now go through some examples.

(i) We can construct confidence intervals for the expected value, μ, in a distribution with the estimating function $m(y, \mu) = y - \mu$. If we in addition want to make inference about the standard deviation, σ, and skewness, γ,

$$m(y, \mu, \sigma, \gamma) = \left(y - \mu, (y - \mu)^2 - \sigma^2, \left(\frac{y - \mu}{\sigma} \right)^3 - \gamma \right)^T$$

can be used.

(ii) Since $\mathrm{E}[\mathrm{I}(Y \leq \theta)] = \Pr(Y \leq \theta)$, the function $m(Y, \theta) = \mathrm{I}(y \leq \theta) - q$ identifies $F^{-1}(q)$ in a continuous distribution, F. Using this estimating function Theorem 2.3.1 allows us to easily construct completely non-parametric confidence intervals for quantiles. When using maximum likelihood theory or large sample distributions of quantiles (see for example Ferguson 1996, pp. 87-92) density estimation is needed. This is not the case for empirical likelihood, and shows how general and easily applicable this non-parametric estimation technique really is.

(iii) Consider a linear regression setting where we have data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the joint distribution $F_{X,Y}$, and want to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, minimizing expected squared loss,

$$\mathrm{E}\,(Y - \beta_0 - X\beta_1)^2.$$

The minimizer of this expression is the solution to

$$\frac{\partial}{\partial \beta} \mathrm{E}\,(Y - \beta_0 - X\beta_1)^2 = 0.$$

Assuming we can apply Leibniz integral theorem, see e.g. Lindstrøm 2017, p. 276, to pass the derivative under the integral sign, this is equivalent to requiring

$$\mathrm{E} \begin{pmatrix} -2(Y - \beta_0 - X\beta_1) \\ -2X(Y - \beta_0 - X\beta_1) \end{pmatrix} = 0.$$

So any $\boldsymbol{\beta}$ minimizing squared loss satisfies the equations

$$\mathrm{E}\,(Y - \beta_0 - X\beta_1) = 0 \quad \text{and} \quad \mathrm{E}\,[X\,(Y - \beta_0 - X\beta_1)] = 0.$$

With this we can use empirical likelihood with the estimating function

$$m(y, x, \beta_0, \beta_1) = (y - \beta_0 - x\beta_1, x\,(y - \beta_0 - x\beta_1))^T$$

to make inference about the regression coefficients.

This example warrants some additional comments. In regression settings, we usually consider the x_i-s as non-stochastic covariates and the $Y_i \mid X_i = x_i$-s as independent random variables. This is not the case in this example. Here we assume the vectors (X_i, Y_i) for $i = 1, \ldots, n$ are independently drawn and identically distributed. In other words, we consider both X_i and Y_i as random variables. For methods and theory regarding linear regression with non-random predictors, see Owen 1991 or chapter 4 of Owen 2001. For extensions of empirical likelihood to generalized linear models, see Kolaczyk 1994.

Looking at examples (i) and (iii) above, an obvious question comes to mind. Is it possible to consider only one parameter at a time? Can we make inference about the variance or skewness in a distribution without computing or caring about the mean? And if we in (iii) are interested in β_1 only, it seems unnecessary to construct a two dimensional confidence region for the full parameter vector, β. The profile empirical likelihood function is an elegant solution to these problems and will be the topic of the upcoming chapter.

CHAPTER 3

Profile empirical likelihood

As argued at the end of the previous section, we are sometimes interested in a subset of, rather than the full, parameter vector, $\theta \in \mathbb{R}^p$. With π denoting the projection onto the relevant coordinates, we can express this as being interested in $\pi(\theta)$. This is an example of a more general idea. Rather than wanting to make inference about the full parameter vector, we want to study a focus parameter which can be expressed as a function of θ.

Focus parameters are used frequently in parametric likelihood theory. One way to make inference about $\psi = g(\theta)$ in this case is to use the profile likelihood function. This map is computed by maximizing the likelihood function over all θ such that $g(\theta) = \psi$. It can be shown that, after proper scaling and centering, the profile log-likelihood has a limiting chi-square distribution. This is called Wilks theorem, see e.g. section 2.4 in Schweder and Hjort 2016 for a reference, and in this section we will define a similar quantity for the empirical likelihood function. In addition, we will prove and state a version of Wilks theorem for this map.

We start by defining the profile empirical likelihood function. This is done analogously as in parametric likelihood theory. For more details about the parametric case see for example Schweder and Hjort 2016, pp. 32–40.

Definition 3.0.1 (Schweder and Hjort 2016, p. 329)**.** Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables. For $\theta \in \mathbb{R}^p$ and $g\colon \mathbb{R}^p \to \mathbb{R}$, the profile empirical likelihood function for the focus parameter $\psi = g(\theta)$ is defined as

$$\mathrm{PEL}_n(\psi) = \sup_{\theta} \{ \, \mathrm{EL}_n(\theta) \mid g(\theta) = \psi \, \}.$$

As mentioned at the beginning of this chapter, the profile log-likelihood converges to a chi square limit, after proper scaling and centering. We would therefore expect similar behavior from the logarithm of $\mathrm{PEL}_n(\psi)$. Results like these are often called Wilks theorems, and in this section we will indeed show a Wilks theorem for the profile empirical likelihood. Before we can do this, however, we will prove a lemma which will be useful in the arguments to come.

Lemma 3.0.2. *Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables following some distribution, F. Assume $\theta_0 \in \mathbb{R}^p$ can be characterized as the unique solution to the estimating equation*

$$\mathrm{E}\, m(Y, \theta) = 0$$

for some $m\colon \mathbb{R}^{d+p} \to \mathbb{R}^q$ *and* $Y \sim F$*. Define the following stochastic process on* $\mathbb{R}^p$*:*

$$A_n(s) = -2 \log \mathrm{EL}_n\left(\theta_0 + \frac{s}{\sqrt{n}}\right), \tag{3.1}$$

where EL_n *is the empirical likelihood function constructed with* m *and* $Y_1, \ldots, Y_n$*. For a compact set,* $K \subseteq \mathbb{R}^p$*, assume the conditions of Lemma 1 in Hjort, I. McKeague, and Van Keilegom 2018 hold true and*

$$\sup_{s \in K} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m\left(y_i, \theta_0 + \frac{s}{\sqrt{n}}\right) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(y_i, \theta_0) - \xi_n s \right\| = o_{\mathrm{Pr}}(1), \tag{3.2}$$

for some $q \times p$ *matrix* ξ_n *tending to* ξ_0 *in probability. Then*

$$A_n \xrightarrow{d} A$$

in $\ell^\infty(K)$*.*

Remark 3.0.3. Here $\ell^\infty(K)$ denotes the vector space of bounded functions from K into $\mathbb{R}$ equipped with the uniform norm, i.e.

$$\|f\|_\infty = \sup_{x \in K} |f(x)|.$$

Remark 3.0.4. Assumption (i) from Lemma 1 in Hjort, I. McKeague, and Van Keilegom 2018 follows from a combination of (3.2) and the central limit theorem. Hence, this condition need not be checked. This is proved in the proof of Theorem 2 in Hjort, I. McKeague, and Van Keilegom 2018.

Proof. Fix a compact subset, K, of $\mathbb{R}^p$. We start by defining some quantities which will be useful in the arguments that follow. For each $s \in K$, let

$$U_n(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m\left(Y_i, \theta_0 + \frac{s}{\sqrt{n}}\right) \quad \text{and} \tag{3.3}$$

$$W_n(s) = \frac{1}{n} \sum_{i=1}^{n} m\left(Y_i, \theta_0 + \frac{s}{\sqrt{n}}\right) m\left(Y_i, \theta_0 + \frac{s}{\sqrt{n}}\right)^T. \tag{3.4}$$

As $\mathrm{E}\|m(Y, \theta_0)\|^2 < \infty$ by the assumptions of Lemma 1 in Hjort, I. McKeague, and Van Keilegom 2018, $U_n(0)$ converges to a normal limit, U, by the central limit theorem. Furthermore, $W_n(0)$ goes in probability to $W = \mathrm{Var}\, U$ by the weak law of large numbers.

From Lemma 1 in Hjort, I. McKeague, and Van Keilegom 2018, we have

$$A_n(s) = U_n(s)^T W_n(s)^{-1} U_n(s) + o_{\mathrm{Pr}}(1)$$

uniformly in K. Furthermore, $\|W_n(s) - W\| = o_{\mathrm{Pr}}(1)$ uniformly over compact sets by the assumptions of Lemma 1 in Hjort, I. McKeague, and Van Keilegom 2018. Hence,

$$A_n(s) = U_n(s)^T W^{-1} U_n(s) + o_{\mathrm{Pr}}(1) \tag{3.5}$$

uniformly in K. Here we have used that

$$\sup_{s\in K}\|U_n(s)\| = O_{\Pr}(1)$$

which is one of the assumptions made in Lemma 1 in Hjort, I. McKeague, and Van Keilegom 2018.

By assumption (3.2), there exists ξ_n tending to ξ_0 in probability such that

$$\sup_{s\in K}\|U_n(s) - U_n(0) - \xi_n s\| \xrightarrow{\Pr} 0.$$

Combining this with (3.5), shows

$$A_n(s) = [U_n(0) + \xi_n s]^T W^{-1}[U_n(0) + \xi_n s] + r_n(s) \tag{3.6}$$

where

$$\sup_{s\in K}|r_n(s)| \xrightarrow{\Pr} 0. \tag{3.7}$$

So $A_n = Z_n + r_n$ where Z_n is the process defined as

$$Z_n(s) = [U_n(0) + \xi_n s]^T W^{-1}[U_n(0) + \xi_n s]. \tag{3.8}$$

By definition, (3.7) is equivalent to r_n converging in probability to 0 as a random sequence in $\ell^\infty(K)$. So by Slutsky's theorem, A_n and Z_n converge to the same limit process if Z_n converges in distribution to some element of $\ell^\infty(K)$.

We claim that $Z_n \xrightarrow{d} A$ as processes in $\ell^\infty(K)$, where A is defined as

$$A(s) = (U + \xi_0 s)^T W^{-1} (U + \xi_0 s).$$

Since Z_n is convex for each $n \in \mathbb{N}$, Theorem 1 in Arcones 1998 ensures that Z_n converges as a process to A if and only if

$$(Z_n(t_1), \ldots, Z_n(t_k))^T \xrightarrow{d} (A(t_1), \ldots, A(t_k))^T$$

for every finite choice of $t_1, \ldots, t_k \in K$. So fix such such a choice of $t_1, \ldots, t_k \in K$. Since $(Z_n(t_1), \ldots, Z_n(t_k))^T$ is a smooth function of $(U_n(0), \xi_n)^T$ converging in distribution to $(U, \xi_0)^T$,

$$\begin{pmatrix} Z_n(t_1) \\ Z_n(t_2) \\ \cdots \\ Z_n(t_k) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} (U + \xi_0 t_1)^T W^{-1} (U + \xi_0 t_1) \\ (U + \xi_0 t_2)^T W^{-1} (U + \xi_0 t_2) \\ \cdots \\ (U + \xi_0 t_k)^T W^{-1} (U + \xi_0 t_k) \end{pmatrix} = \begin{pmatrix} A(t_1) \\ A(t_2) \\ \cdots \\ A(t_k) \end{pmatrix}.$$

Hence,

$$Z_n \xrightarrow{d} A$$

as elements of $\ell^\infty(K)$. Combining this with the arguments above, shows

$$A_n \xrightarrow{d} A \tag{3.9}$$

as a process on K. ■

Lemma 3.0.2 can be seen as a special case of Theorem 1 in Hjort, I. McKeague, and Van Keilegom 2018. Our proof is, however, more detailed and formulated slightly different. It is therefore included for the sake of completion.

With Lemma 3.0.2, we are ready to show a Wilks theorem for the profile empirical likelihood function.

Theorem 3.0.5. *Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables following a distribution, F. Assume $\theta_0 \in \mathbb{R}^p$ can be characterized as the unique solution to the estimating equation*

$$\mathrm{E}\, m(Y, \theta) = 0$$

for some $m \colon \mathbb{R}^{d+p} \to \mathbb{R}^q$ and $Y \sim F$. Suppose the focus parameter, ψ, is equal to $g(\theta)$ for some function $g \colon \mathbb{R}^p \to \mathbb{R}$, whose second order partial derivatives are all continuous. We then have

$$-2 \log \mathrm{PEL}_n(\psi_0) \xrightarrow{d} \chi^2_1,$$

where $\psi_0 = g(\theta_0)$, provided the conditions of Lemma 3.0.2 hold for all compact subsets of $\mathbb{R}^p$, the matrices

$$\xi_0 \xi_0^T \quad \text{and} \quad \xi_0^T W^{-1} \xi_0$$

are invertible and

$$\sqrt{n}\Big(\widehat{\theta} - \theta_0\Big) = O_{\mathrm{Pr}}(1), \tag{3.10}$$

where $\widehat{\theta}$ is a maximizer of $\mathrm{EL}_n(\theta)$ over the set of θ-values such that $g(\theta) = \psi_0$.

A result similar to Theorem 3.0.5 is stated in Schweder and Hjort 2016 without proof or a full set of conditions. We will therefore take the time to show Theorem 3.0.5. The proof we present follows the arguments in Remark 2.5 of Schweder and Hjort 2016, p. 36–37, but includes modifications and additions.

Proof. First, fix

$$K = \{\, s \in \mathbb{R}^p \mid \|s\| \leq M \,\},$$

for some $M > 0$. Define the following sets:

$$S_n = \left\{ s \in \mathbb{R}^p \;\middle|\; g\left(\theta_0 + \frac{s}{\sqrt{n}}\right) = \psi_0,\ ||s|| \leq M \right\}$$

for each $n \in \mathbb{N}$ and

$$T = \left\{\, s \in \mathbb{R}^p \;\middle|\; b^T s = 0,\ ||s|| \leq M \,\right\}.$$

Our first goal will be to show

$$\inf_{s \in S_n} A_n(s) = \inf_{s \in T} A_n(s) + o_{\mathrm{Pr}}(1), \tag{3.11}$$

with A_n defined as in (3.1).

Let b denote the gradient of g at θ_0. Using a first order Taylor expansion around θ_0, we notice

$$g\left(\theta_0 + \frac{s}{\sqrt{n}}\right) = g(\theta_0) + b^T \frac{s}{\sqrt{n}} + \epsilon_n(s),$$

where $\epsilon_n(s)$ is a remainder term, bounded in norm by

$$\frac{C(|s_1| + \ldots + |s_p|)^2}{n},$$

where $s = (s_1, \ldots, s_p)^T$ and C is a fixed positive number. This follows from continuity of the second order partial derivatives of g and a version of Taylors theorem for vector-valued functions, see e.g. Corollary 6.5.8 in Lindstrøm 2017, p. 199. Since K is compact, $|s_1| + \ldots + |s_p|$ is bounded. Hence,

$$g\left(\theta_0 + \frac{s}{\sqrt{n}}\right) = \psi_0 + b^T \frac{s}{\sqrt{n}} + \epsilon_n(s)$$

where $\epsilon_n(s)$ tends uniformly to 0 over K at speed $O(1/n)$. Because of this,

$$S_n = \left\{ s \in \mathbb{R}^p \;\middle|\; b^T s + \sqrt{n}\epsilon_n(s) = 0,\ ||s|| \leq M \right\},$$

where, again, $\epsilon_n(s)$ tends uniformly to 0 over K at speed $O(1/n)$. In particular,

$$\sup_{s \in S_n} \left| b^T s \right| = \left| \sqrt{n}\epsilon_n(s) \right| = O\left(1/\sqrt{n}\right). \tag{3.12}$$

To show (3.11) we will first prove

$$\inf_{s \in S_n} A_n(s) = \inf_{s \in S_n} A_n[\mathrm{Proj}_T(s)] + o_{\mathrm{Pr}}(1) \tag{3.13}$$

where $\mathrm{Proj}_T(s)$ denotes the projection of s onto T. This is given by

$$\mathrm{Proj}_T(s) = s - \frac{b^T s}{\|b\|^2} b.$$

In the proof of Lemma 3.0.2, we showed that A_n and Z_n, given in (3.8), were asymptotically equivalent as stochastic processes on K. Hence, for (3.13), it suffices to show

$$\inf_{s \in S_n} Z_n(s) = \inf_{s \in S_n} Z_n[\mathrm{Proj}_T(s)] + o_{\mathrm{Pr}}(1).$$

We will now do this.

Some algebraic efforts lead to the following identity

$$Z_n(s) - Z_n[\mathrm{Proj}_T(s)] = \tag{3.14}$$

$$\frac{b^T s}{\|b\|^2} \cdot 2b^T \xi_n^T W^{-1} U_n - \frac{(b^T s)^2}{\|b\|^4} b^T \xi_n^T W^{-1} \xi_n b + \frac{b^T s}{\|b\|^2} \cdot 2b^T \xi_n^T W^{-1} \xi_n s. \tag{3.15}$$

Where U_n is $U_n(0)$ defined in (3.3) and W is the limit of $W_n(0)$. We know that U_n tends in distribution to a normally distributed variable and that ξ_n

goes in probability to ξ_0. Because of this, and Slutsky's theorem, $2b^T\xi_n^T W^{-1}U_n$ converges in distribution to some random variable. Furthermore,

$$\sup_{s\in S_n}\left|b^T s\right| = O(1/\sqrt{n}).$$

by (3.12). Hence,

$$\sup_{s\in S_n}\left|\frac{b^T s}{\|b\|^2}\cdot 2b^T\xi_n^T W^{-1}U_n\right| \xrightarrow{\Pr} 0.$$

Arguing similarly, we can show that also

$$\sup_{s\in S_n}\left|\frac{(b^T s)^2}{\|b\|^4}b^T\xi_n^T W^{-1}\xi_n b\right|$$

tends to 0 in probability. For the last term of (3.15), notice that

$$\left|2b^T\xi_n^T W^{-1}\xi_n s\right| \le 2\left\|\xi_n^T W^{-1}\xi_n b\right\|\cdot\|s\| \le 2M\left\|\xi_n^T W^{-1}\xi_n b\right\|,$$

for all $s\in K$. Arguing as before, we get

$$\sup_{s\in S_n}\left|\frac{b^T s}{\|b\|^2}\cdot 2b^T\xi_n^T W^{-1}\xi_n s\right| \le \sup_{s\in S_n}\left|\sqrt{n}\epsilon_n(s)\right|\cdot O_{\Pr}(1)\xrightarrow{\Pr} 0.$$

Combining all our results, with (3.15) and applying the triangle inequality, shows

$$\sup_{s\in S_n}|Z_n(s) - Z_n[\mathrm{Proj}_T(s)]| = o_{\Pr}(1). \tag{3.16}$$

Fix $\epsilon > 0$ and assume

$$\sup_{s\in S_n}|Z_n(s) - Z_n[\mathrm{Proj}_T(s)]| < \frac{\epsilon}{2}. \tag{3.17}$$

Then

$$\inf_{s\in S_n} Z_n(s) \le Z_n[\mathrm{Proj}_T(u)] + \frac{\epsilon}{2} \tag{3.18}$$

for all $u\in S_n$. To see this, notice that otherwise there exists $u\in S_n$ such that

$$\inf_{s\in S_n} Z_n(s) > Z_n[\mathrm{Proj}_T(u)] + \frac{\epsilon}{2},$$

and then (3.17) ensures

$$Z_n(u) \le Z_n[\mathrm{Proj}_T(u)] + \frac{\epsilon}{2}.$$

Hence,

$$Z_n(u) \le Z_n[\mathrm{Proj}_T(u)] + \frac{\epsilon}{2} < \inf_{s\in S_n} Z_n(s),$$

which is a contradiction. Because of this, (3.18) holds for all $u \in S_n$, and we must have

$$\inf_{s\in S_n} Z_n(s) \leq \inf_{s\in S_n} Z_n[\mathrm{Proj}_T(s)] + \frac{\epsilon}{2}.$$

Hence,

$$\sup_{s\in S_n} |Z_n(s) - Z_n[\mathrm{Proj}_T(s)]| < \epsilon \implies \inf_{s\in S_n} Z_n(s) \leq Z_n[\mathrm{Proj}_T(s)] + \frac{\epsilon}{2}.$$

Similarly, one can show

$$\sup_{s\in S_n} |Z_n(s) - Z_n[\mathrm{Proj}_T(s)]| < \epsilon \implies \inf_{s\in S_n} Z_n[\mathrm{Proj}_T(s)] \leq \inf_{s\in S_n} Z_n(s) + \frac{\epsilon}{2}.$$

Combining these two implications, guarantees

$$\sup_{s\in S_n} |Z_n(s) - Z_n[\mathrm{Proj}_T(s)]| < \frac{\epsilon}{2} \implies \left| \inf_{s\in S_n} Z_n(s) - \inf_{s\in S_n} Z_n \mathrm{Proj}_T(s) \right| \leq \epsilon.$$

So,

$$\begin{aligned} &\Pr\left(\left| \inf_{s\in S_n} Z_n(s) - \inf_{s\in S_n} Z_n \mathrm{Proj}_T(s) \right| \leq \epsilon \right) \geq \\ &\Pr\left(\sup_{s\in S_n} |Z_n(s) - Z_n[\mathrm{Proj}_T(s)]| < \frac{\epsilon}{2} \right). \end{aligned}$$

The right-hand side of this equation goes to 1 as $n \to \infty$ by (3.16). Hence

$$\inf_{s\in S_n} Z_n(s) = \inf_{s\in S_n} Z_n[\mathrm{Proj}_T(s)] + o_{\Pr}(1),$$

showing (3.13).

We have now shown (3.13). If the image of S_n under Proj_T is T,

$$\inf_{s\in S_n} A_n[\mathrm{Proj}_T(s)] = \inf_{s\in T} A_n(s),$$

and (3.13) is sufficient for (3.11). We will now show that, eventually, this holds true.

Notice that,

$$g\left(\theta_0 + \frac{s}{\sqrt{n}}\right) = \psi_0 + \frac{b^T s}{\sqrt{n}} + \frac{\sqrt{n}\epsilon_n(s)}{\sqrt{n}},$$

with $\sqrt{n}\epsilon_n(s)$ tending to 0 uniformly in K. Because of this, we can choose $N \in \mathbb{N}$ such that for all $n \geq N$, $\sup_{s\in K} |\sqrt{n}\epsilon_n(s)| < 1$. Fix $t \in T$ and consider the functions $f_n \colon \mathbb{R} \to \mathbb{R}$ defined as

$$\begin{aligned} f_n(x) &= g\left(\theta_0 + \frac{1}{\sqrt{n}}(t + x \cdot b)\right) \\ &= \psi_0 + \frac{b^T}{\sqrt{n}}(t + x \cdot b) + \frac{\sqrt{n}\epsilon_n(t + x \cdot b)}{\sqrt{n}} \end{aligned}$$

$$= \psi_0 + \frac{x\|b\|^2}{\sqrt{n}} + \frac{\sqrt{n}\epsilon_n(t + x \cdot b)}{\sqrt{n}}.$$

for each $n \geq N$. These are continuous functions with

$$f_n\left(-\frac{2}{\|b\|^2}\right) \leq \psi_0 - \frac{2}{\sqrt{n}} + \frac{1}{\sqrt{n}} < \psi_0$$

and

$$f_n\left(\frac{2}{\|b\|^2}\right) \geq \psi_0 + \frac{2}{\sqrt{n}} - \frac{1}{\sqrt{n}} > \psi_0.$$

Hence, by the intermediate value theorem, there exists $x_n \in [-2/\|b\|^2, 2/\|b\|^2]$ such that $f_n(x_n) = \psi_0$ for all $n \geq N$. Define the following vector,

$$u_n = t + x_n \cdot b.$$

By construction $g\left(\theta_0 + \frac{u_n}{\sqrt{n}}\right) = f_n(x_n) = \psi_0$, so $u_n \in S_n$. Furthermore, $\text{Proj}_T(u_n) = t$ as $b \perp T$ and $t \in T$. Hence, the image of S_n under Proj_T is T, for $n \geq N$. By the previous arguments, this ensures (3.11).

We have now shown

$$\inf_{s \in S_n} A_n(s) = \inf_{s \in T} A_n(s) + o_{\text{Pr}}(1), \tag{3.19}$$

and from Lemma 3.0.2 we know $A_n \xrightarrow{d} A$ as a process in $\ell^\infty(K)$. Since,

$$f \mapsto \inf_{b^T s = 0} f(s)$$

is a continuous as a function from $\ell^\infty(K)$ into $\mathbb{R}$, the continuous mapping theorem guarantees

$$\inf_{s \in T} A_n(s) \xrightarrow{d} \inf_{s \in T} A(s).$$

Combining this with (3.19) and Slutsky's theorem, reveals

$$\inf_{s \in S_n} A_n(s) \xrightarrow{d} \inf_{s \in T} A(s)$$

The compact set K used above, was arbitrary. What we have, in fact, shown is

$$\inf_{g_n(s) = \psi_0,\, \|s\| \leq M} A_n(s) \xrightarrow{d} \inf_{b^T s = 0,\, \|s\| \leq M} A(s) \tag{3.20}$$

for all $M > 0$. Here $g_n(s)$ is short-hand for

$$g\left(\theta_0 + \frac{s}{\sqrt{n}}\right).$$

By definition

$$-2 \log \text{PEL}_n(\psi_0) = \min_{g_n(s) = \psi_0} A_n(s).$$

Furthermore, the sequence of maximizers

$$s_n = \sqrt{n}\Big(\widehat{\theta} - \theta_0\Big)$$

from (3.10) is stochastically bounded. Hence,

$$\lim_{n\to\infty} \Pr(\|s_n\| \leq M) \geq 1 - \epsilon_M$$

with ϵ_M tending to 0 as $M \to \infty$. In particular, this ensures

$$\lim_{M\to\infty} \lim_{n\to\infty} \Pr(\|s_n\| \leq M) \geq 1,$$

which implies

$$\lim_{M\to\infty} \lim_{n\to\infty} \Pr(\|s_n\| \leq M) = 1.$$

Now,

$$\begin{aligned}
&\Pr(-2\log \mathrm{PEL}_n(\psi_0) \leq x) = \\
&\Pr\Big(\inf_{g_n(s)=\psi_0,\, \|s\|\leq M} A_n(s) \leq x\Big) \Pr(\|s_n\| \leq M) + O[\Pr(\|s_n\| > M)].
\end{aligned}$$

So,

$$\begin{aligned}
&\lim_{n\to\infty} \Pr(-2\log \mathrm{PEL}_n(\psi_0) \leq x) = \\
&\lim_{M\to\infty} \lim_{n\to\infty} \Pr[-2\log \mathrm{PEL}_n(\psi_0) \leq x] = \\
&\lim_{M\to\infty} \lim_{n\to\infty} \Pr\Big(\inf_{g_n(s)=\psi_0,\, \|s\|\leq M} A_n(s) \leq x\Big) \cdot 1 + 0.
\end{aligned}$$

By (3.20),

$$\lim_{n\to\infty} \Pr\Big(\inf_{g_n(s)=\psi_0,\, \|s\|\leq M} A_n(s) \leq x\Big) = \Pr\Big(\inf_{b^T s=0,\, \|s\|\leq M} A(s) \leq x\Big),$$

provided x is a continuity point in the distribution of the limit. And hence,

$$\lim_{n\to\infty} \Pr(-2\log \mathrm{PEL}_n(\psi_0) \leq x) = \lim_{M\to\infty} \Pr\Big(\inf_{b^T s=0,\, \|s\|\leq M} A(s) \leq x\Big).$$

Let $\hat{s}$ denote the minimizer of A in $\mathbb{R}^p$. Then, for every $\epsilon > 0$ and $M \in \mathbb{R}$,

$$\Pr\Big(\Big|\min_{b^T s=0, \|s\|\leq M} A(s) - \min_{b^T s=0} A(s)\Big| \geq \epsilon\Big) \leq \Pr(\|\hat{s}\| > M).$$

The right hand side of the above equation goes to 0 as $M \to \infty$. Convergence in probability implies convergence in distribution, so this guarantees that

$$\lim_{M\to\infty} \Pr\Big(\min_{b^T s=0, \|s\|\leq M} A(s) \leq x\Big) = \Pr\Big(\min_{b^T s=0} A(s) \leq x\Big). \tag{3.21}$$

All of this shows that

$$-2\log \mathrm{PEL}_n(\psi_0) \xrightarrow{d} \min_{b^T s=0} A(s) \tag{3.22}$$

We will now find the distribution of this limit.

To compute the limit distribution, we use the method of Lagrange multipliers. Some algebra reveals that the minimum of A over the set of $s \in \mathbb{R}^p$ with $b^T s = 0$ is given by

$$U^T W^{-1} U - U^T W^{-1} \xi_0 \left(\xi_0^T W^{-1} \xi_0\right)^{-1} \xi_0^T W^{-1} U + \frac{X^2}{b^T \left(\xi_0^T W^{-1} \xi_0\right)^{-1} b}, \quad (3.23)$$

where

$$X = b^T \left(\xi_0^T W^{-1} \xi_0\right)^{-1} \xi_0^T W^{-1} U.$$

By assumption $\xi_0 \xi_0^T$ is an invertible $q \times q$-matrix, so

$$\begin{aligned} W^{-1} \xi_0 \left(\xi_0^T W^{-1} \xi_0\right)^{-1} \xi_0^T W^{-1} &= \left(\xi_0 \xi_0^T\right)^{-1} \xi_0 \xi_0^T W^{-1} \xi_0 \left(\xi_0^T W^{-1} \xi_0\right)^{-1} \xi_0^T W^{-1} \\ &= \left(\xi_0 \xi_0^T\right)^{-1} \xi_0 \xi_0^T W^{-1} \\ &= W^{-1}. \end{aligned}$$

Hence, the two first terms in (3.23) cancels, and

$$\min_{b^T s = 0} A(s) = \frac{X^2}{b^T \left(\xi_0^T W^{-1} \xi_0\right)^{-1} b}.$$

U is central normal distributed variable with variance matrix W. Because of this, the variance of X is

$$b^T \left(\xi_0^T W^{-1} \xi_0\right)^{-1} \xi_0^T W^{-1} W W^{-1} \xi_0 \left(\xi_0^T W^{-1} \xi_0\right)^{-1} b = b^T \left(\xi_0^T W^{-1} \xi_0\right)^{-1} b.$$

Therefore,

$$X \sim \mathrm{N}\left(0, b^T \left(\xi_0^T W^{-1} \xi_0\right)^{-1} b\right).$$

This means that

$$\min_{b^T s = 0} A(s) = \frac{X^2}{b^T \left(\xi_0^T W^{-1} \xi_0\right)^{-1} b} \sim \chi_1^2.$$

Combining the above with (3.22), shows

$$-2 \log \mathrm{PEL}_n(\psi_0) \xrightarrow{d} \chi_1^2$$

and concludes the proof. ■

Let π_j be the projection onto the j-th coordinate. As explained above, we can use Theorem 3.0.5 with $h = \pi_3$ to make inference about the skewness using Theorem 3.0.5 and the estimating function given in example (i) in the previous section. Similarly, profiling with $h = \pi_2$ can be used in example (iii) to construct confidence intervals for the slope without the second dimension introduced by the intercept. However, the use of Theorem 3.0.5 is not limited to getting rid of nuisance parameters. For two-dimensional data, we could, for instance, use θ equal to the mean and $\psi = g(\theta) = \theta_1/\theta_2$. In this case Theorem 3.0.5 gives a

way to make inference about ψ, allowing us to compare the expectations in the two distributions. In Chapter 4 we will use the empirical likelihood machinery in multiple examples. We therefore refer to that section for illustrations of how Theorem 3.0.5 can be used.

Versions of Theorem 3.0.5 have been established before. In Qin and Lawless 1994 a profiling result for focus parameters on the form $g(\theta_1, \theta_2) = \theta_1$ is stated and proved for smooth estimating functions, m. This is applicable in many cases, but the theorem cannot be used with, for instance, $m(y, \theta) = I(y \leq \theta) - q$. Molanes Lopez, Van Keilegom, and Veraverbeke 2009, on the other hand, arrive at a profiling result for non-smooth estimating functions. Their theorem does, however, need m to be bounded. This excludes multiple interesting quantities like means and variances in distributions with unbounded support. We would therefore claim that the theorem proved in this section is slightly more general that these two results. We do require some sort of smoothness in terms of (3.2), but this is not as strict as demanding differentiability of the estimating function. As an example, take $m(y, \theta) = I(y \leq \theta) - q$. This map is not differentiable in θ, but by Stute 1982 or arguments similar to those of example 19.29 in Vaart 1998, p. 283,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[I\left(Y_i \leq \theta_0 + s/\sqrt{n}\right) - I(Y_i \leq \theta_0) \right] = f(\theta_0) + o_{\Pr}(1)$$

uniformly over compact sets when the distribution of the data is continuous with density function, f. In addition, boundedness of m is not required for Theorem 3.0.5, making the result applicable in situations where the main theorem of Molanes Lopez, Van Keilegom, and Veraverbeke 2009 is not.

Theorem 3.0.5 can also be used for more complicated functions than projections. We can use the theorem to make inference about every focus parameter that can be expressed as $g(\theta)$ as long as g is "smooth enough". This is neither possible with the results in Qin and Lawless 1994 nor Molanes Lopez, Van Keilegom, and Veraverbeke 2009. So Theorem 3.0.5 generalizes these propositions in that regard as well.

We have also been made aware that a result similar to Theorem 3.0.5 is shown in Guolo and Adimari 2010. Sadly, this was not discovered until the proof, illustration and discussion in this thesis were already in place. This is, of course, unfortunate, but as the method of proof presented here is both very different and more detailed than the one found in Guolo and Adimari 2010, we believe this section still serves a purpose and deserves its place in the thesis. We do, for example, prove convergence of the sequence of processes, A_n, to the limit, A. This is not done in Guolo and Adimari 2010, and might be of use in other applications than the proof of Theorem 3.0.5.

It is also worth noting that Theorem 3.0.5 can be significantly strengthened. Closer inspection of the proof of Lemma 1 in Hjort, I. McKeague, and Van Keilegom 2018 and the arguments above, reveals that, rather than needing the data to be i.i.d., we need only

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Y_i, \theta_0) \xrightarrow{d} U$$

for some random vector, U, with nonsingular variance matrix, W, such that

$$\frac{1}{n}\sum_{i=1}^{n} m(\,Y_i,\theta_0)m(\,Y_i,\theta_0)^T \xrightarrow{\Pr} W.$$

This allows Theorem 3.0.5 to be applied in regression settings, as well as in certain situations with dependent data. The same conclusion is reached by Guolo and Adimari 2010. We will therefore refrain from giving a full proof here, and only assure the reader that the calculations go through after small modifications.

Lastly, we want to comment on another approach to making inference about focus parameters with empirical likelihood. Assume a parameter vector $\left(\theta_1^0,\theta_2^0\right)$ can be expressed as the solution to

$$\mathrm{E}\, m(\,Y,\theta_1,\theta_2) = 0,$$

and that we have a good estimate, $\widehat{\theta}_2$, of θ_2. In such situations, one might wish to use this estimate of θ_2 rather than profiling out the parameter. This is possible, and a limit distribution for $\mathrm{EL}_n(\theta_1^0,\hat{\theta}_2)$ does exist. That being said, the limit is not always on the simple χ_1^2 form, but can be expressed as a weighted sum of chi-square distributed variables. See Hjort, I. W. McKeague, and Keilegom 2009 for illustrations, proofs and a full statement of such a result.

CHAPTER 4

Examples

Empirical likelihood methods is a very general way of making inference about quantities that can be expressed as solutions to estimating equations. In addition, Theorem 3.0.5 allows us to easily construct confidence intervals for functions of such parameters as well. In this chapter we will present some examples illustrating how the results from the previous chapters can be used.

4.1 Income in Oslo

In this section we will use empirical likelihood to make inference about the mean yearly income in the capital of Norway. We will also investigate how much the earnings vary between sub-districts. Lastly we will use Theorem 3.0.5 to make inference about the coefficient of variation for yearly income in the sub-districts of Oslo. All data is extracted from Statistikkbanken 2020.

There are 98 sub-districts in Oslo. We will treat the mean yearly income from 2019, given in NOK, as i.i.d. random variables $Y_1, \ldots, Y_n$ for $n = 98$. Furthermore, we will work with the numbers divided by one million for numerical stability. For this data set the observed values range from 0.293 to 1.149 between regions of the city. Our first goal will be to estimate and construct confidence intervals for the mean and variance of yearly income in the sub-districts of Oslo. The empirical likelihood theory of Chapter 2 will be used to do this.

To construct the empirical likelihood function, we need to decide on an estimating function. We will use $m\colon \mathbb{R}^3 \to \mathbb{R}^2$ defined as

$$m\big(y, \mu, \sigma^2\big) = \begin{pmatrix} x - \mu \\ (y-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Let Y be a random variable following the same distribution as $Y_1, \ldots, Y_n$. For the true mean, $\mathrm{E}Y = \mu_0$, and variance, $\mathrm{Var}Y = \sigma_0^2$, we then have

$$\mathrm{E}\, m\big(Y, \mu_0, \sigma_0^2\big) = \begin{pmatrix} \mathrm{E}\, Y - \mu_0 \\ \mathrm{Var}\, Y - \sigma_0^2 \end{pmatrix} = 0.$$

Because of this, the estimating equation $\mathrm{E}[m(Y, \mu, \sigma^2)] = 0$ is solved by (μ_0, σ_0^2). Hence, the empirical likelihood function of (μ, σ^2) can be constructed with this m and be computed as described in Section 2.2. We implemented this for our data set, and the resulting plot of the graph of the empirical likelihood function can be found in Figure 4.1.

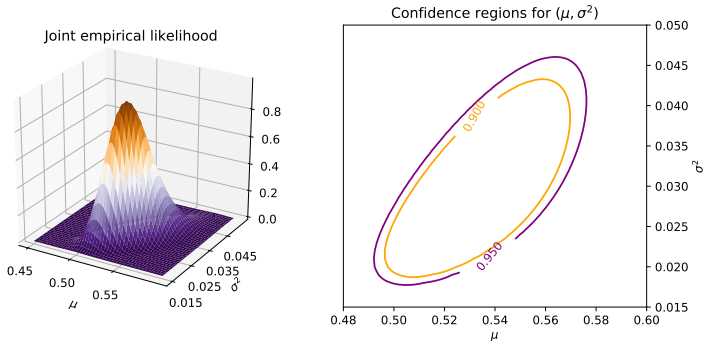


Figure 4.1: A plot of the joint empirical likelihood function of (μ, σ^2) together with contours of 90%- and 95%-confidence regions for the parameter vector.

We can now use Theorem 2.3.1 to construct confidence regions for the parameter vector (μ, σ^2). By this result, $-2\log \mathrm{EL}_n\big(\mu_0, \sigma_0^2\big)$ is approximately chi square distributed with two degrees of freedom. Hence,

$$\begin{aligned}\Pr\big[-2\log \mathrm{EL}_n\big(\mu_0, \sigma_0^2\big) \le \Gamma_2^{-1}(1-\alpha)\big] &= \Pr\big\{\Gamma_2\big[-2\log \mathrm{EL}_n\big(\mu_0, \sigma_0^2\big)\big] \le 1-\alpha\big\} \\ &\approx 1-\alpha,\end{aligned}$$

where Γ_2 is the cumulative distribution function in the χ_2^2-distribution. Hence, the following is an approximate $1-\alpha$ confidence region for (μ_0, σ_0^2):

$$\left\{ \big(\mu, \sigma^2\big) \;\middle|\; \Gamma_2\big[-2\log \mathrm{EL}_n\big(\mu, \sigma^2\big)\big] \le 1-\alpha \right\}.$$

Contours of 90%- and 95%-confidence regions are shown in Figure 4.1 together with the empirical likelihood function.

Although the preceding calculations allow us to make educated guesses about the parameter vector (μ, σ^2), they do not give a way to construct confidence intervals for μ or σ alone. One solution is to use Theorem 3.0.5, and we will now apply the result to make inference about each of the parameters separately.

We start with μ. To compute $\mathrm{PEL}_n(\mu)$ for a fixed value of μ, we maximize $\mathrm{EL}_n(\mu, \sigma^2)$ over all values of σ^2. This can be done numerically by maximizing the function $x \mapsto \mathrm{EL}_n(\mu, x)$ for each fixed μ, and a plot of the function can be found in Figure 4.2. After computing the profile empirical likelihood function, we can construct approximate confidence intervals for the mean yearly income in Oslo. By Theorem 3.0.5, $-2\log \mathrm{PEL}_n(\mu_0)$ is approximately chi square distributed with one degree of freedom. So, with Γ_1 denoting the cumulative distribution function in the χ_1^2-distribution,

$$\begin{aligned}1-\alpha &= \Pr\big(-2\log \mathrm{PEL}_n(\mu_0) \le \Gamma_1^{-1}(1-\alpha)\big) \\ &= \Pr\{\Gamma_1[-2\log \mathrm{PEL}_n(\mu_0)] \le 1-\alpha\}.\end{aligned}$$

Hence,

$$\left\{ \mu \;\middle|\; \Gamma_1[-2\log \mathrm{PEL}_n(\mu_0)] \le 1-\alpha \right\}$$

is a $1-\alpha$ confidence set for the mean yearly income. To illustrate, we computed approximate 90% and 95% confidence intervals for μ, and obtained [0.504, 0.559] and [0.499, 0.565] respectively.

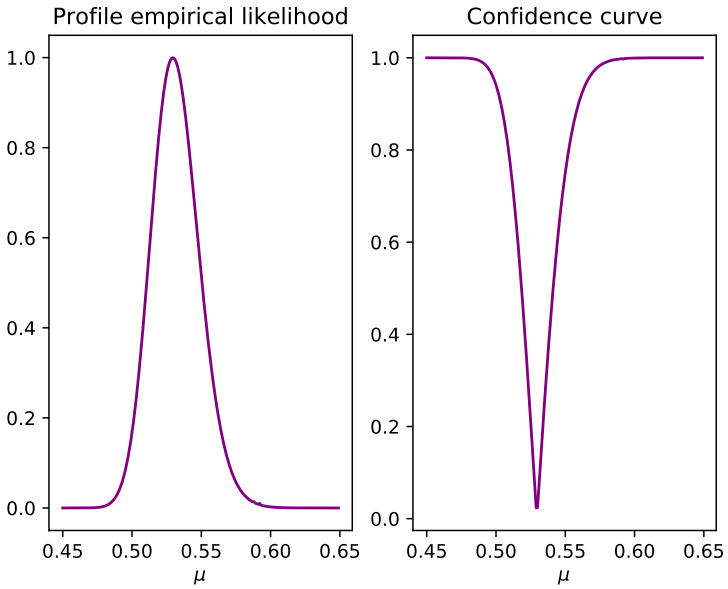


Figure 4.2: The profile empirical likelihood function of μ together with a confidence curve for the parameter.

The above can also be used to construct a confidence curve for μ. A confidence curve is a graphical summary visualizing the result of a statistical analysis. In particular, confidence intervals of all levels for a parameter can be read of the corresponding confidence curve. This is done by tracing a horizontal line from the desired confidence level and setting the two intersections with the curve as bounds for the interval. A full introduction to confidence curves can be found in Schweder and Hjort 2016, but for our purposes it suffices to use these as visual representations of our certainty about parameters from which confidence intervals of all levels can be read of. For the mean yearly income in Oslo, μ, the confidence curve is given by

$$\Gamma_1[-2\log \mathrm{PEL}_n(\mu)],$$

as the solutions to

$$\Gamma_1[-2\log \mathrm{PEL}_n(\mu)] = 1-\alpha$$

are the bounds for an approximate $(1-\alpha)$-confidence interval for μ as shown in the previous paragraph. We computed this curve for a selection of μ-values, and the resulting plot can be found in Figure 4.2.

We can repeat the above process to make inference about σ. The procedure is more or less identical to the situation with μ. The only difference is that we now use

$$\mathrm{PEL}_n(\sigma) = \max\left\{ \mathrm{EL}_n\left(\mu, \sigma^2\right) \;\middle|\; \mu \in \mathbb{R} \right\}$$

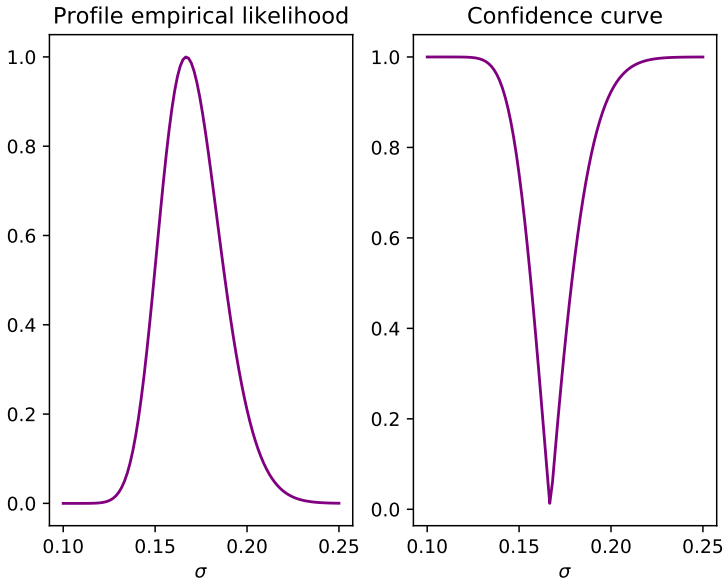


Figure 4.3: The profile empirical likelihood function of σ together with a confidence curve for the parameter.

rather than $\mathrm{PEL}_n(\mu)$. This function can be computed by maximizing

$$x \mapsto \mathrm{EL}_n\left(x, \sigma^2\right)$$

for each fixed σ. As the remaining arguments are more or less identical as those presented above, we will not go through all the details. However, a plot of the empirical likelihood function, together with a confidence curve can be found in Figure 4.3. Furthermore, 90% and 95% confidence intervals were computed to be [0.144, 0.197] and [0.140, 0.204] respectively.

We now turn our attention to the focus parameter $\psi = \sigma/\mu$. This is called the coefficient of variation and is a normalized measure of variability in a distribution. High values of ψ means that the income between sub-districts is very variable, while low values indicate the opposite. It can, in fact, be shown that this quantity can be used as a measure of inequality, see e.g. Campano 2006.

To make inference about ψ, we will again use Theorem 3.0.5. We start by constructing the profile empirical likelihood function,

$$\mathrm{PEL}_n(\psi) = \max\left\{ \mathrm{EL}_n\left(\mu, \sigma^2\right) \;\middle|\; \frac{\sqrt{\sigma^2}}{\mu} = \psi \right\}.$$

For each fixed ψ, $\mathrm{PEL}_n(\psi)$ can be computed by numerically maximizing $x \mapsto \mathrm{EL}_n(x, (\psi x)^2)$. We did this for a selection of ψ-values, and a plot of the profile empirical likelihood function can be found in Figure 4.4.

After computing the profile empirical likelihood function, we can constructed confidence curves and intervals for the coefficient of variation. This is done

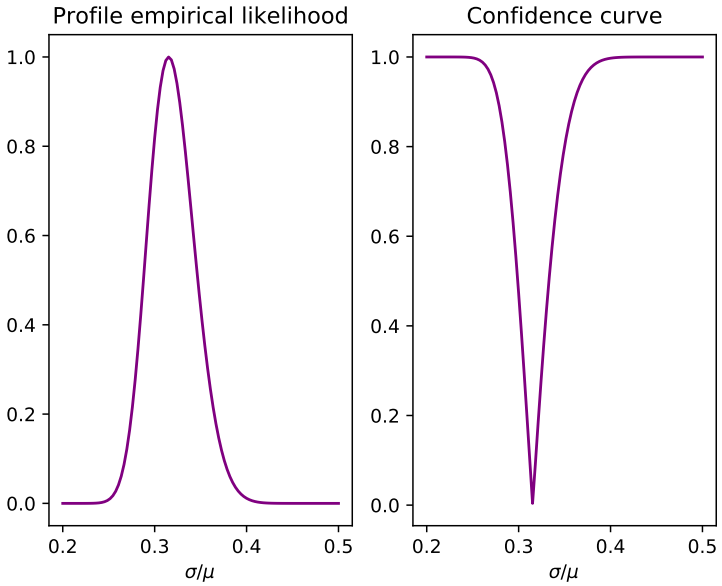


Figure 4.4: The profile empirical likelihood function of $\psi = \sigma/\mu$ together with a confidence curve for the parameter.

similarly as for μ and σ, and details will therefore be left out. With our data set we find that [0.278, 0.360] is an approximate 90%-confidence interval, while [0.272, 0.369] is a 95% one. We also computed a confidence curve. This is displayed in Figure 4.4 together with $\mathrm{PEL}_n(\psi)$ for a range of ψ values.

The analysis done in this section can easily be modified to other situations and data sets. This is particularly interesting when one wishes to estimate the coefficient of variation or quantities related to this value. An example of where this might be useful is in archeology. In this field the coefficient of variation is used to assess whether a group of artifacts are made by standardized production or not, see Eerkens and Bettinger 2001.

4.2 Score functions as estimating functions

In this section we will illustrate the theorems and definitions on simulated data. We will generate $n = 100$ i.i.d. data points from a Gamma-distribution with shape 2 and rate 3 and attempt to estimate the parameters in the distribution using empirical likelihood. We will also compare these estimates with what we get using maximum likelihood.

In a Gamma(α, β)-distribution the density function takes the form

$$f_{\alpha,\beta}(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y), \quad \text{for } y > 0.$$

This results in the following log-density:

$$\log f_{\alpha,\beta}(y) = \alpha \log \beta - \log \Gamma(\alpha) + (1-\alpha) \log y - \beta y, \quad \text{for } y > 0,$$

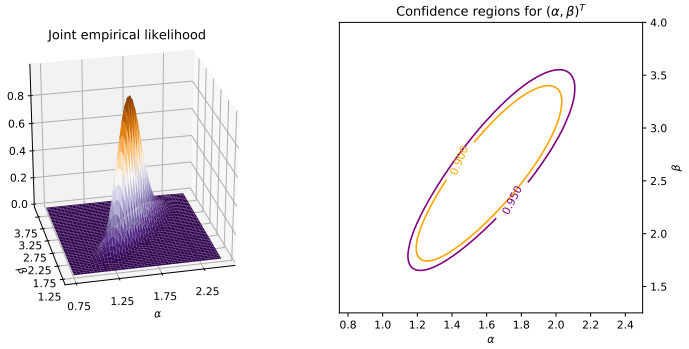


Figure 4.5: To the left we see the joint empirical likelihood function of $(\alpha, \beta)^T$. On the right-hand side contours of confidence regions of level 90% and 95% are displayed.

with gradient

$$\left(\log\beta + \psi(\alpha) - \log y, \frac{\alpha}{\beta} - y\right)^T, \quad \text{for } y > 0.$$

Here ψ denotes the digamma function, the derivative of the logarithm of Γ. The expected value of the score function is 0 at the true parameter vector. Therefore

$$\mathrm{E}\, m(Y, \alpha, \beta) = 0,$$

when $Y \sim \text{Gamma}(\alpha, \beta)$ and

$$m(y, \alpha, \beta) = \left(\log\beta + \psi(\alpha) - \log y, \frac{\alpha}{\beta} - y\right)^T.$$

Since our simulated data follows a Gamma, distribution, we can use this m as our estimating function in the construction of the empirical likelihood function.

Now that we have decided on an estimating function, we can compute the empirical likelihood function as described in Section 2.2. The graph of the function for a selection of values of αs and βs can be found in Figure 4.5 together with confidence regions of two levels. These were constructed as explained in the previous example, and we refer to this section for further explanation.

We will now use Theorem 3.0.5 to construct approximate confidence intervals and curves for two quantities, the mean and standard deviation in the distribution.

We start with the mean. In a Gamma distribution with shape parameter α and rate parameter β the expectation is given by α/β. For this focus parameter the profile empirical likelihood function takes the form

$$\mathrm{PEL}_n(\psi) = \max\left\{ \mathrm{EL}_n(\alpha, \beta) \;\middle|\; \frac{\alpha}{\beta} = \psi \right\}.$$

This can be computed numerically by maximizing $x \mapsto \mathrm{EL}_n(\psi x, x)$ for each fixed ψ. Using Theorem 3.0.5 we can construct confidence intervals and curves

for the expected value of the data with this function. The procedure is more or less identical to that of the previous example. Because of this, the arguments will not be repeated here, but a confidence curve can be found in Figure 4.6. Confidence intervals of all levels can be read off this.

The analysis can be repeated for the standard deviation. We then use

$$\mathrm{PEL}_n(\psi) = \max\left\{ \mathrm{EL}_n(\alpha,\beta) \;\middle|\; \frac{\sqrt{\alpha}}{\beta} = \psi \right\},$$

which can be computed by numerically maximizing $x \mapsto \mathrm{EL}_n[(x\psi)^2, x]$. Otherwise the procedure is the same as for α/β. Again details will be omitted, but relevant plots can be found in Figure 4.7.

The empirical likelihood function is not the only way to make inference about focus parameters. In this example, another obvious choice is to use maximum likelihood theory, as we do, in fact, know what distribution the data follows. To get an idea of how accurate the confidence intervals and curves obtained with empirical likelihood theory are, we will use the standard parametric version of Wilks theorem to make inference about the mean and standard deviation in the distribution of $Y_1, \ldots, Y_n$, and compare the results obtained with the two methods.

Let ℓ_n denote the parametric log-likelihood of the data,

$$\ell_n(\alpha,\beta) = \sum_{i=1}^{n} \log f_{\alpha,\beta}(Y_i),$$

and define the profile likelihood function for a focus parameter $\psi = g(\alpha,\beta)$ as

$$\ell_{n,prof}(\psi) = \max\{\, \ell_n(\alpha,\beta) \mid g(\alpha,\beta) = \psi \,\}.$$

Furthermore, let D_n denote the profile deviance function, i.e.

$$D_n(\psi) = \ell_{n,prof}\left(\widehat{\psi}_{ml}\right) - \ell_{n,prof}(\psi),$$

where $\widehat{\psi}_{ml}$ is the maximum likelihood estimate of ψ. By Wilks theorem, see e.g. Schweder and Hjort 2016, p. 35,

$$2 \cdot D_n(\psi_0) \xrightarrow{d} \chi^2_1$$

for the true value of ψ, ψ_0. We can use this result to construct approximate parametric confidence curves and intervals for focus parameters. This procedure is similar to those involving the profile empirical likelihood function. Because of this, arguments will not be repeated, but the resulting confidence curves have been added to the plots in Figure 4.6 and Figure 4.7.

From the figures we see that the results obtained using empirical and maximum likelihood are very similar. That being said, the confidence curves constructed with the parametric approach are slightly narrower than the ones we get using empirical likelihood. This is not too surprising as the likelihood ratio test is asymptotically optimal, i.e. uniformly most powerful, for testing simple hypothesis like the ones we have considered here, see e.g. Vaart 1998, p. 236 for more information and discussion around this. In this example, however, we are in an ideal situation where we know what parametric family

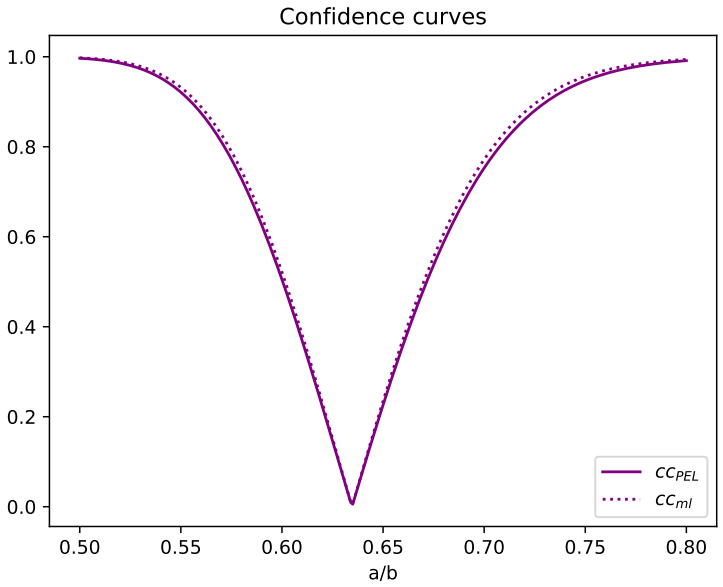


Figure 4.6: Confidence curves for the mean, α/β, based on empirical and maximum likelihood. The full drawn line is constructed with PEL_n and the dotted one maximum likelihood theory.

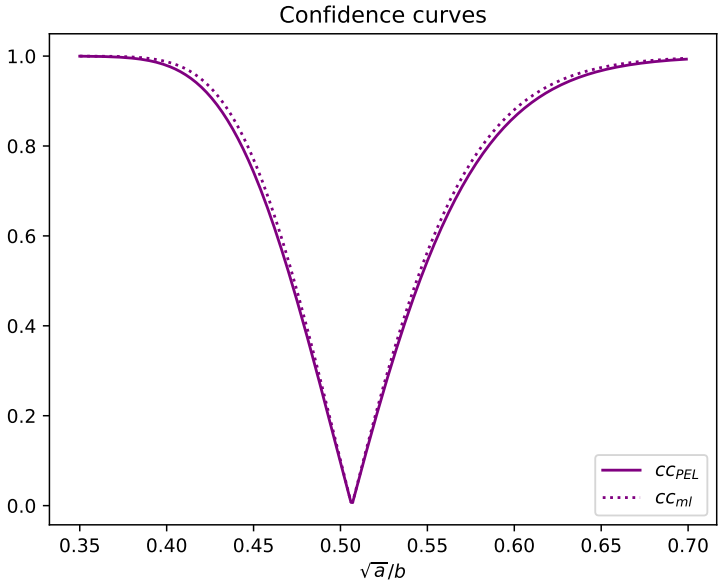


Figure 4.7: Confidence curves for the standard deviation, $\sqrt{\alpha}/\beta$ based on empirical and maximum likelihood. The full drawn curve is constructed with PEL_n and the dotted one maximum likelihood theory.

the distribution of the data belongs to. This might not always be the case, and under misspecification of the model the non-parametric approach might outperform maximum likelihood. In such cases α and β can no longer be interpreted as parameters in the true distribution, but the focus parameter α/β is still an interesting quantity. The reason for this is the second entry in the score function, $\alpha/\beta - y$. Since the expectation of this expression is 0 if and only if $\alpha/\beta = \mathrm{E}\, Y$, the empirical likelihood confidence intervals for α/β will, indeed, be confidence intervals for the mean in the distribution. The same is true when using maximum likelihood. Even when the model is specified incorrectly, maximum likelihood estimates are consistent for the minimizer of the Kullback-Leibler distance, (α_0, β_0), and this is at its smallest when

$$\begin{aligned} 0 &= \frac{\partial}{\partial(\alpha,\beta)^T}\Big|_{\alpha_0,\beta_0} \mathrm{E}\, \log f_{\alpha,\beta}(Y) \\ &= \mathrm{E}\left(\frac{\partial}{\partial(\alpha,\beta)^T}\Big|_{\alpha_0,\beta_0} \log f_{\alpha,\beta}(Y) \right) \\ &= \mathrm{E} \begin{pmatrix} \log\beta_0 + \psi(\alpha_0) - \log Y \\ \alpha_0/\beta_0 - Y \end{pmatrix}. \end{aligned}$$

In particular, this means that the maximum likelihood estimate is consistent for $(\alpha_0, \beta_0)^T$ such that $\alpha_0/\beta_0 = \mathrm{E}\, Y$. It is therefore possible to use maximum likelihood theory to make inference about the true mean in the distribution, even when the true underlying distribution is not really a Gamma-distribution. To do this one needs to use a version of Wilks theorem for misspecified models, see e.g. Schweder and Hjort 2016, pp. 43–44. Construction of confidence intervals and curves is still doable, but the procedure changes and requires additional effort. When using empirical likelihood, on the other hand, the analysis carries through unmodified.

4.3 A deadly example

In Pinker 2011, the writer argues that violence in the world has declined. Numerous authors have since attempted to prove or disprove this statement, see Cunen, Hjort, and Nygård 2020 for a general overview, and in this section we will use empirical likelihood to give an answer of our own. The data we will use is part of the Correlates of War data set (Sarkees and Wayman 2010). In particular, we will use the number of battle deaths in inter-state wars between 1816 and 2007 to determine whether the world has become more peaceful or not.

To talk about violence we need to specify what we mean when we use this word. We are working with a data set containing casualties in wars. Number of battle deaths will therefore be our measure of violence. This is, of course, somewhat limited as neither civilian casualties nor wounded soldiers are included in this number. Furthermore the data set only considers inter-state wars. Struggles within nations and conflicts that are not formally wars are not included in the data set. Nevertheless, we will use this definition and investigate whether the number of battle deaths in inter-state wars has declined.

We are interested whether newer conflicts are more deadly than older ones. To test this we will compare parameters in the distribution of older and newer

wars. We will first investigate whether the "typical" number of battle deaths have declined. We will use the median as measure of this "typical" number of casualties. At the end of the example we will also analyze the third quartile. With this we are investigating something slightly different: whether the larger conflicts have become more deadly or not.

We want to investigate whether the number of battle deaths has changed over time. It is therefore not sufficient to simply estimate the median number of casualties in wars. Firstly, we need to separate wars into "older" and "newer". Secondly, we have to compare the medians in the two groups. To differentiate between newer and older conflicts, we use the results of Cunen, Hjort, and Nygård 2020. In this article the authors find that the Korean War is the maximum likelihood estimate for the change point in the number of battle deaths. Because of this, we will take the 60 conflicts before, and including, the Korean War as "older" wars. The remaining 35 will be the newer conflicts.

We will treat the data points as observations of 95 independent random variables, and assume that the first sixty follow a continuous distribution, F_1. The remaining 35 variables will be assumed to follow another distribution, F_2. Provided the distribution of a stochastic variable, Y, is continuous,

$$\mathrm{E}[I(Y \leq \theta) - 0.5] = \Pr(Y \leq \theta) - 0.5 = 0,$$

when θ equals the median in the distribution of Y. We will therefore use the estimating function $m\colon \mathbb{R}^2 \to \mathbb{R}$ defined as

$$m(y, \theta) = I(y \leq \theta) - 0.5$$

to construct two empirical likelihood functions. One for the median number of battle deaths in conflicts before the Korean War, and one for the median number of casualties in newer wars.

The empirical likelihood functions can be computed as explained in Section 2.2. We will not go through the implementation details, but a plot of the graphs of the functions can be found in Figure 4.8. Looking at this figure, we notice that the empirical likelihood function for the median number of battle deaths in newer wars does not reach 1. This happens because $\mathrm{EL}_n(\theta) = 1$ for some θ if and only if

$$\frac{1}{n}\sum_{i=1} m(Y_i, \theta) = 0 \tag{4.1}$$

for this θ. The estimating function $m(y, \theta) = I(y \leq \theta) - 0.5$ takes two values only: -0.5 and 0.5. Hence, (4.1) can only be solved if n is even. We have used the 35 most recent conflicts as "newer" wars and 35 is an odd number. Because of this, (4.1) cannot be solved in this case. A common way of correcting for this is to use half-correction. This corresponds to replacing $m(y, \theta)$ with the estimating function

$$m(y, \theta) = I(y < \theta) - \frac{1}{2} I(y = \theta),$$

as explained in Owen 2001, pp. 45–48. This sets the value of EL_n at the empirical median to 1. Such a correction is needed to identify the median uniquely in discrete distributions, but as we have assumed the true underlying

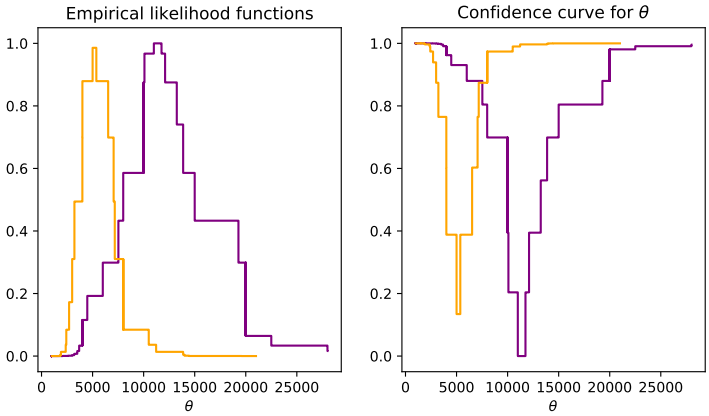


Figure 4.8: The plot on the left is $\mathrm{EL}_n(\theta)$ for a selection of θ-values. To the right we see the corresponding confidence curves based on Theorem 2.3.1. We have used purple for the values corresponding to wars before the Korean war and orange for the ones corresponding to the newer conflicts.

distribution to be continuous, we decided not to use this technique. Our reason being that the original choice of m is slightly easier to work with. Furthermore, although resulting in more aesthetic pleasing plots, there is no real reason to need EL_n to be 1 at the empirical median.

Since,

$$\operatorname{Var} m(Y,\theta) = \operatorname{Var}[I(Y \leq \theta) - 0.5] = \Pr(Y \leq \theta)[1 - \Pr(Y \leq \theta)] < \infty$$

for both $Y \sim F_1$ and $Y \sim F_2$, we can apply Theorem 2.3.1 to make inference about the medians in the two distributions. For $j = 1, 2$,

$$-2 \log \mathrm{EL}_{n_j,j}(\theta_j) \overset{d}{\approx} \chi^2_1,$$

where θ_j is the true median in the j-th population, by this result. With this we can construct confidence curves for the parameters with the following expression:

$$\mathrm{cc}_j(\theta) = \Gamma_1[-2 \log \mathrm{EL}_{n_j,j}(\theta)] \qquad \text{for } j = 1, 2.$$

Plots of the curves can be found in Figure 4.8 together with the empirical likelihood functions. Confidence intervals of all levels can be read off this figure.

From Figure 4.8 it looks like the median number of casualties has declined. To test this hypothesis formally, we will use the profiling result from Chapter 3. So, let X_i denote the number of deaths in the i-th conflict before the Korean war and Y_i the same for after this conflicts. Since the X_i-s and the Y_i-s are not assumed to come from the same distribution, we cannot use Definition 2.1.1 directly to form the empirical likelihood function for the parameter vector $\theta = (\theta_1, \theta_2)$. Inspired by the classical parametric likelihood, we can, however, define

$$\mathrm{EL}(\theta) = \mathrm{EL}_{n_1}(\theta_1) \cdot \mathrm{EL}_{n_2}(\theta_2).$$

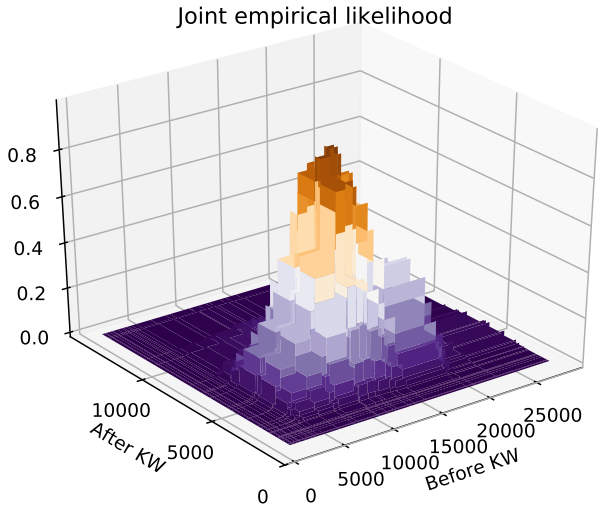


Figure 4.9: A plot of the joint empirical likelihood function of (θ_1, θ_2) where θ_j is the median in data set j. The left-hand axis in the xy-plane corresponds to θ_2 values and the right-hand side to values of θ_1.

Using this definition we get

$$-2\log \mathrm{EL}(\theta) = -2\log \mathrm{EL}_{n_1}(\theta_1) - 2\log \mathrm{EL}_{n_2}(\theta_2),$$

and as $-2\log \mathrm{EL}_{n_j}(\theta_j) \xrightarrow{d} \chi^2_1$ for $j = 1, 2$ and the X_i and Y_i are independent,

$$-2\log \mathrm{EL}(\theta) \xrightarrow{d} \chi^2_2,$$

showing that a version of Theorem 2.3.1 holds for $\mathrm{EL}(\theta)$ as well. With this we can compute approximate confidence regions and curves for θ based on $\mathrm{EL}(\theta) \overset{d}{\approx} \chi^2_2$. A plot of the joint empirical likelihood function can be found in Figure 4.9. Contours of approximate confidence regions of level $90\%, 95\%$ and 99% are plotted in Figure 4.10.

We are now ready to compare the median number of casualties before and after the Korean war. We will do this by making inference about the focus parameter

$$\psi = g(\theta_1, \theta_2) = \frac{\theta_1}{\theta_2}.$$

We would like to apply Theorem 3.0.5 to this situation. We cannot do this directly as we are working with two different data sets: $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$. That being said, X_i and Y_j are assumed to be independent for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Hence, the process $(s_1, s_2) \mapsto B_{n_1}(s_1) + C_{n_2}(s_2)$ converges to $(s_1, s_2) \mapsto B(s_1) + C(s_2)$, where B_{n_1} and B are the processes A_n

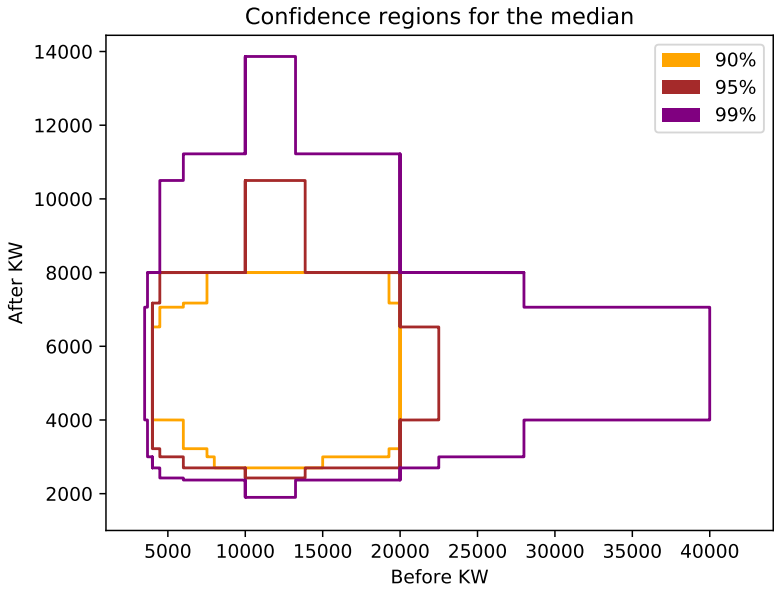


Figure 4.10: Boundaries of confidence regions for (θ_1, θ_2) where θ_j is the median in data set j. The x-axis corresponds to θ_1 and the y-axis values of θ_2.

and A constructed in the proof of Theorem 3.0.5 for data points before the Korean War, and C_{n_2} and C are the corresponding processes for the second data set. Using these process in place of A_n and A, the rest of the proof of Theorem 3.0.5 goes through without notable changes, and we arrive at the conclusion

$$-2\log \mathrm{PEL}(\psi) = -2 \cdot \max_{\theta_1,\theta_2} \left\{ \log \mathrm{EL}(\theta_1,\theta_2) \,\middle|\, \frac{\theta_1}{\theta_2} = \psi \right\} \xrightarrow{d} \chi^2_1,$$

as n_1 and n_2 goes to infinity.

Using the approximation $-2\log \mathrm{PEL}(\psi) \overset{d}{\approx} \chi^2_1$, we can compute approximate confidence intervals and curves based on $\mathrm{PEL}(\psi)$. The approximate 95% confidence interval for ψ was computed to be [0.69, 6.20]. So on a 5% level, we are not able to reject the null hypothesis $H_0 \colon \psi = 1$ against the alternative $H_1 \colon \psi > 1$. Furthermore computation of the 90% confidence interval gives us [0.92, 4.99]. So nor on a 10% level are we able to conclude that the median number of battle deaths has decreased since the Korean war. Consult Figure 4.11 for a plot of the confidence curve for ψ.

The p-value for testing the hypothesis $H_0 \colon \psi_0 = 1$ vs $H_1 \colon \psi_0 > 1$ can be read of Figure 4.11. The value is 10.2%, which is too high to reject H_0 on most relevant significance levels. This is similar to the results obtained in Cunen, Hjort, and Nygård 2020. The authors are not able to conclude that the median number of battle deaths has declined on all relevant significance levels. They do, however, find that for higher quantiles, the ratio of quantiles is significantly larger than 1. We will therefore repeat the above analysis for the

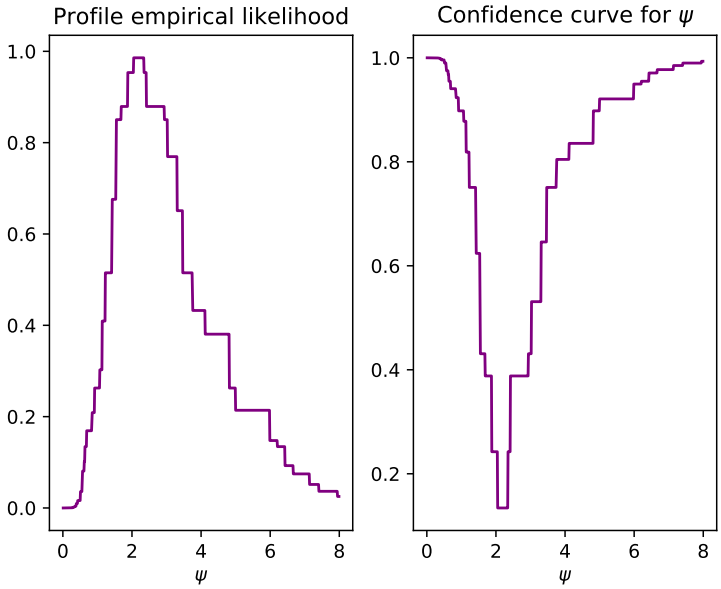


Figure 4.11: The plot on the left-hand side is PEL(ψ) for different values of $\psi = \theta_1/\theta_2$, where θ_j is the median in F_j. To the right we see the corresponding confidence curve.

third quartile instead of the median. After changing the estimating equation to $m(x, \theta) = I(x \leq \theta) - 0.75$ the approach is identical to the one we just used. We will therefore not repeat the arguments here, but a plot of the empirical likelihood function for the ratio of the quantiles before and after the Korean war can be found in Figure 4.12, together with the corresponding confidence curve. The new 95% confidence interval is [1.2, 18.9], and the p-value for testing $H_0 : \psi = 1$ against $H_1 : \psi > 1$ is 0.028. So, significance levels of both 5 and 10% leads us to reject the null hypothesis. Hence, using the third quantile as the measure of violence, we can conclude that the world has, indeed, become more peaceful.

In conclusion, we notice that, while it is not clear whether the median number of casualties has declined, we are quite certain that the more deadly conflicts have become less lethal. This agrees with the conclusion of Cunen, Hjort, and Nygård 2020. In this article the authors conclude that the distribution of battle deaths has changed, but that the changes are greater for the upper than lower parts of the distribution. In particular, they find that the ratio of medians is significantly greater than 1 only on a level slightly larger than 5%. They are therefore unable to certainly conclude that the median number of battle deaths has declined. This agrees with our analysis as we were unable to reject the hypothesis that the median has remained unchanged on most relevant significance levels. In addition, their and our conclusions regarding the upper quartile match. They find that the ratio of these values before and after the Korean War is significantly larger than 1 on a 5%-significance level. This is the same conclusion we reached in this section.

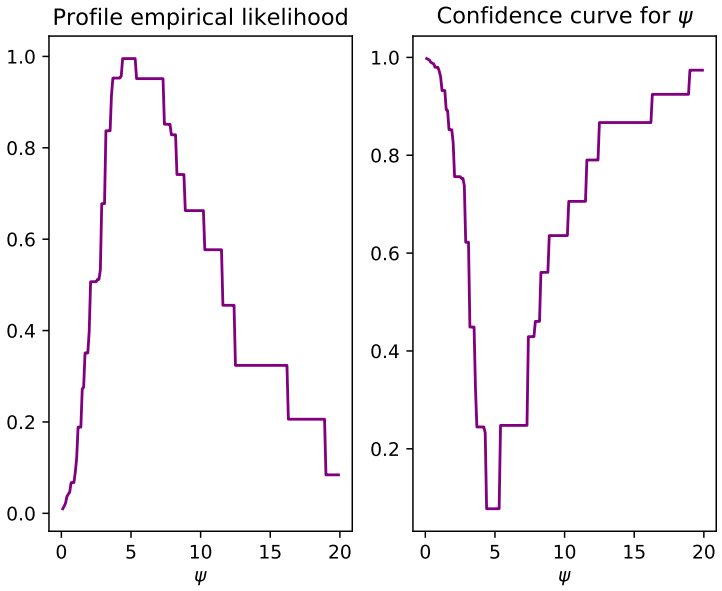


Figure 4.12: The plot on the left-hand side is PEL(ψ) for different values of $\psi = \theta_1/\theta_2$, where θ_j is the third quartile in F_j. To the right we see the corresponding confidence curve.

CHAPTER 5

A mean in disguise

In Section 2.2 we derived an alternative characterization of the empirical likelihood function. This was accomplished by introduction of a parameter, λ_n, given implicitly as the solution to a certain function. We will now build on the ideas from this section to derive asymptotic equivalent expressions to the empirical likelihood function. In particular we will show that $n^{-1} \log \mathrm{EL}_n(\theta)$ is close to the mean of a certain function. This result will be used in Chapter 6 to show consistency and asymptotic normality of the maximizer of the empirical likelihood function and again in Chapter 8 in a hybrid setting.

The proofs and theory in this chapter is, to our knowledge, new, but Molanes Lopez, Van Keilegom, and Veraverbeke 2009 used similar ideas to show a profiling result in their article.

Throughout this chapter we will assume that $Y_1, \ldots, Y_n \in \mathbb{R}^d$ are i.i.d. variables following some unknown distribution, F, such that, for a certain function $m \colon \mathbb{R}^{d+p} \to \mathbb{R}^q$,

$$\mathrm{E}\, m(Y, \mu_0) = 0$$

for $Y \sim F$ and some $\mu_0 \in \mathbb{R}^p$. We will also let $M_i(\mu)$ denote $m(Y_i, \mu)$ and $M(\mu)$ be shorthand for $m(Y, \mu)$ for a general $Y \sim F$.

5.1 An overview

In this, and the following chapter, we will present multiple theorems and proofs. Before we start with the mathematics, we will therefore attempt to explain the general ideas in more informal terms. The goal of this section is to provide the reader with some intuition, while also serving as an overview of the theory that is to come.

We start by repeating some of the results and ideas from Section 2.2. By definition of the empirical likelihood function,

$$\mathrm{EL}_n(\mu) = \left\{ \prod_{i=1}^n n w_i \;\middle|\; \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i M_i(\mu) = 0 \text{ and } w_i \geq 0 \right\},$$

where the last condition should hold for $i = 1, \ldots, n$. Hence $\mathrm{EL}_n(\mu)$ is the solution to the following optimization problem.

Optimize

$$f(w) = \prod_{i=1}^{n} nw_i$$

subjected to the equality constraints

$$h(w) = \sum_{i=1}^{n} w_i M_i(\mu) = 0 \quad \text{and} \quad g(w) = \sum_{i=1}^{n} w_i - 1 = 0. \tag{5.1}$$

and inequality constraints

$$w_i \geq 0 \text{ for } i = 1, \ldots, n. \tag{5.2}$$

If there are no set of strictly positive weights, w, satisfying (5.1), the empirical likelihood function of μ is zero. Otherwise the maximum of f subjected to the constraints $h(w) = 0$ and $g(w) = 0$ is also the maximum of $\log f$ over the same set. Using the method of Lagrange multipliers, one can show that in this case

$$\log \mathrm{EL}_n(\mu) = -\sum_{i=1}^{n} \log\Big(1 + \lambda_n(\mu)^T M_i(\mu)\Big) \tag{5.3}$$

for some $\lambda_n(\mu)$, such that

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i(\mu)}{1 + \lambda_n(\mu)^T M_i(\mu)}. \tag{5.4}$$

This is shown in Owen 2001, pp. 21–22. The argument will not be repeated here.

Reformulating the above slightly, reveals

$$\mathrm{EL}_n(\mu) = \prod_{i=1}^{n} \Big(1 + \lambda_n(\mu)^T M_i(\mu)\Big)^{-1} \tag{5.5}$$

for some $\lambda_n(\mu)$ solving

$$\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} M_i(\mu)/[1 + \lambda^T M_i(\mu)] \\ I[1 + \lambda^T M_i(\mu) \leq 0] \end{pmatrix} = 0 \tag{5.6}$$

and 0 if no solution exists. By the law of large numbers (5.6) goes almost surely to

$$\begin{pmatrix} \mathrm{E}\{M(\mu)/[1 + \lambda^T M(\mu)]\} \\ \Pr[1 + \lambda^T M(\mu) \leq 0] \end{pmatrix} = 0 \tag{5.7}$$

for every λ such that the expectations in (5.7) exists. For large sample sizes we would therefore expect (5.6) to have a zero if (5.7) has one. This would imply that $\mathrm{EL}_n(\mu)$ can be expressed as (5.5) asymptotically provided (5.7) has a solution. In the ensuing section we will show that this is indeed the case, and that, under sufficient conditions, the probability of solving (5.6) goes to 1 if (5.7) has a solution.

The solution to (5.6), $\lambda_n(\mu)$, is what we call a Z-estimator, see e.g. Vaart 1998, p. 41 for a definition of Z-estimators. Such estimators are popular in statistics and enjoy properties like consistency and asymptotic normality under weak conditions. We would therefore expect $\lambda_n(\mu)$ to converge in probability to the solution of (5.7) and to have a normal limit distribution after proper scaling and centering. In the next section will show that such properties do, indeed, hold.

If $\lambda_n(\mu)$ goes sufficiently fast to $\lambda(\mu)$, $n^{-1}\log \mathrm{EL}_n(\mu)$ will be very close to

$$-\frac{1}{n}\sum_{i=1}^{n}\log\Big(1+\lambda(\mu)^T M_i(\mu)\Big).$$

This is a mean, and means converge to expected values by the law of large numbers. Because of this, some sort of convergence of $n^{-1}\log \mathrm{EL}_n(\mu)$ towards

$$-\mathrm{E}\,\log\Big(1+\lambda(\mu)^T M(\mu)\Big)$$

is to be expected. This will be the topic of Section 5.3.

If the above results hold true, standard theory concerning M-estimators, see e.g. Vaart 1998, p. 41, can be applied to prove consistency and asymptotic normality of maximizers of the empirical likelihood function. This will be the topic of Chapter 6, and in Chapter 8 the results will be applied in a hybrid setting combining parametric and empirical likelihood.

Although the ideas presented in this section are intuitively easy to grasp, providing rigorous proofs of the statements is far from straightforward. We therefore recommend keeping the ideas presented here in mind when reading the following sections.

The expression in (5.7) will be used frequently in this thesis. Because of this, having a name for the equation will be useful. As (5.7) is derived as the limit of an expression used in the method of Lagrange multipliers, we will refer to it as the Lagrange equation from now on.

5.2 The solution to the Lagrange equation

We start by showing some limit properties of $\lambda_n(\mu)$, the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} M_i(\mu)/[1+\lambda^T M_i(\mu)] \\ I[1+\lambda^T M_i(\mu)\le 0]\end{pmatrix}=0. \tag{5.8}$$

As this vector is a Z-estimator, standard theory concerning such estimators can be applied. The following result uses theorem 5.41 and 5.42 in Vaart 1998, p. 68 to guarantee that (5.8) has a zero with probability tending to 1. Furthermore, consistency of λ_n towards the solution of the population version of (5.8) is proved, as well as limit normality after proper scaling and centring.

Theorem 5.2.1. *Fix $\mu\in\mathbb{R}^p$ and define the following set*

$$\Lambda_\mu=\big\{\,\lambda\in\mathbb{R}^p \;\big|\; \Pr\big(1+\lambda^T M(\mu)\le 0\big)=0\,\big\}$$

and let $\lambda(\mu)$ be a vector such that

$$0=\mathrm{E}\left(\frac{M(\mu)}{1+\lambda(\mu)^T M(\mu)}\right). \tag{5.9}$$

Assume further that $\mathrm{E}\|M(\mu)\|^3$ *is finite and that*

$$\mathrm{E}\left(\frac{M(\mu)M(\mu)^T}{[1+\lambda(\mu)^T M(\mu)]^2}\right)$$

is non-singular. Lastly, let there be a neighborhood, N*, of* $\lambda(\mu)$ *and a positive real number,* L*, such that* $\Pr[1+\lambda^T M(\mu) > L] = 1$ *for all* $\lambda \in N$*. Under these conditions*

$$\mathrm{EL}_n(\mu) = \prod_{i=1}^{n}\Big(1+\lambda_n(\mu)^T M_i(\mu)\Big)^{-1},$$

for some $\lambda_n(\mu)$ *solving*

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} M_i(\mu)/[1+\lambda^T M_i(\mu)] \\ I[1+\lambda^T M_i(\mu)\le 0]\end{pmatrix} = 0 \tag{5.10}$$

happens with probability tending to 1. Furthermore, if (5.10) *has at most one solution for each* n*, any sequence of* $\lambda_n(\mu)$ *solving* (5.10) *converges in probability to* $\lambda(\mu)$ *and has the property*

$$\sqrt{n}[\lambda_n(\mu)-\lambda(\mu)] = S(\mu)^{-1}V_n(\mu) + o_{\Pr}(1) \tag{5.11}$$

where

$$S(\mu) = \mathrm{E}\left(\frac{M(\mu)M(\mu)^T}{[1+\lambda(\mu)^T M(\mu)]^2}\right) \quad \textit{and} \quad V_n(\mu) = \frac{1}{\sqrt{n}}\sum_{i=1}\frac{M_i(\mu)}{1+\lambda(\mu)^T M_i(\mu)}.$$

Proof. For every $\lambda \in \Lambda_\mu$, $\Pr[1+\lambda^T M(\mu)\le 0] = 0$. Because of this

$$\frac{1}{n}\sum_{i=1}^{n} I\big(1+\lambda^T M_i(\mu)\le 0\big) = 0$$

is satisfied with probability 1 for every $\lambda \in \Lambda_\mu$. Hence, it suffices to show that the probability of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{M_i(\mu)}{1+\lambda^T M_i(\mu)} = 0 \tag{5.12}$$

having a solution in Λ_μ tends to 1 to have $\Pr[\mathrm{EL}_n(\mu)=0]\to 0$. In addition, it suffices to work only with roots of (5.12) in Λ_μ to find limits of solutions to (5.10).

We would like to apply theorem 5.41 and 5.42 in Vaart 1998, p. 68 and will therefore start by showing the conditions of these result. In the following we will assume Λ_μ is an open set. If this is not the case, we can replace it with an open subset containing $\lambda(\mu)$. By assumption there exists a neighborhood of $\lambda(\mu)$, N, with $\Pr[1+\lambda^T M(\mu) > L] = 1$ for some $L > 0$. Because of this $\lambda(\mu)$ lies in the interior of Λ_μ. Replacing Λ_μ with an open subset containing $\lambda(\mu)$ is therefore always possible.

Since $\Pr[1 + \lambda^T M(\mu) \leq 0] = 0$ for all $\lambda \in \Lambda_\mu$,

$$\psi(y, \lambda) = \frac{m(y, \mu)}{1 + \lambda^T m(y, \mu)}$$

is smooth in λ for every fixed y in the support of $Y_1, \ldots, Y_n$. In particular the function is twice differentiable. Furthermore,

$$\mathrm{E}\,\psi[\,Y, \lambda(\mu)] = \mathrm{E}\left(\frac{M(\mu)}{1 + \lambda(\mu)^T M(\mu)}\right) = 0,$$

by definition of $\lambda(\mu)$, and

$$\mathrm{E}\left\|\frac{M(\mu)}{1 + \lambda(\mu)^T M(\mu)}\right\|^2 \leq \frac{\mathrm{E}\|M(\mu)\|^2}{L^2} < \infty$$

as $\|M(\mu)\|^2$ has finite mean and $\Pr[1 + \lambda(\mu)^T M(\mu) > L] = 1$ for some $L > 0$ by assumption. The matrix

$$\mathrm{E}\left(\frac{\partial}{\partial\lambda}\Big|_{\lambda(\mu)} \psi(\,Y, \lambda)\right) = -S$$

is assumed to be non-singular, so the only condition left to check is that there is a neighborhood of $\lambda(\mu)$ on which the second order partial derivatives of $\lambda \mapsto \psi(y, \lambda)$ are dominated by a function, h, such that $h(\,Y)$ has finite expectation.

Let m_j denote j^{th} estimating function. Then

$$\frac{\partial^2 \psi}{\partial\lambda_j \partial\lambda_k}(y, \lambda) = 2 \cdot \frac{m_h(y, \mu) m_j(y, \mu) m_k(y, \mu)}{[1 + \lambda^T m(y, \mu)]^3}.$$

By assumption there is a neighborhood, N, of $\lambda(\mu)$ and a strictly positive number, L, such that $\Pr[1 + \lambda^T m(Y, \mu) > L] = 1$. On this set, we have

$$\left|\frac{\partial^2}{\partial\lambda_j \partial\lambda_k} \frac{m_h(y, \mu)}{1 + \lambda^T m(y, \mu)}\right| \leq 2 \cdot \left|\frac{m_h(y, \mu) m_j(y, \mu) m_k(y, \mu)}{L^3}\right| \leq \frac{2\|m(y, \mu)\|^3}{L^3}.$$

And since

$$\mathrm{E}\|M(\mu)\|^3 < \infty$$

by the conditions in Theorem 5.2.1, this proves that the second order partial derivatives of $\lambda \mapsto \psi(y, \lambda)$ are bounded by a function, h, for which $\mathrm{E}\,h(\,Y) < \infty$. We have now shown that the conditions of theorem 5.41 and 5.42 in Vaart 1998, p. 68 hold.

Theorem 5.2.1 now follows more or less directly from theorem 5.41 and 5.42 in Vaart 1998, p. 68. By theorem 5.42 the probability of (5.10) having a solution tends to 1. Hence

$$\mathrm{EL}_n(\mu) = \prod_{i=1}^{n} \Big(1 + \lambda_n(\mu)^T M(\mu)\Big)^{-1}$$

for some $\lambda_n(\mu)$ solving (5.10) happens with probability tending to 1. Furthermore, the same theorem ensures that there exists a sequence of roots tending to $\lambda(\mu)$ in probability. Since the roots are unique by assumption, any sequence of solutions, $\lambda_n(\mu)$, must be this consistent sequence. In conclusion $\lambda_n(\mu) \xrightarrow{\Pr} \lambda(\mu)$. Theorem 5.41 in Vaart 1998, p. 68 then guarantees

$$\sqrt{n}[\lambda_n(\mu) - \lambda(\mu)] = S(\mu)^{-1}V_n(\mu) + o_{\Pr}(1).$$

This concludes the proof. ■

One drawback with the above result is that, we need to know if

$$\begin{pmatrix} \mathrm{E}\{M(\mu)/[1+\lambda^T M(\mu)]\} \\ \Pr[1+\lambda^T M(\mu) \leq 0] \end{pmatrix} = 0$$

can be solved. In Section 5.4 we provide some discussion and illustration of when a solution exists and it behaves as a function of μ. These arguments mostly involves mathematical analysis of functions and serve no further purpose in the proofs. Because of this, we will postpone the discussion to the end of this chapter.

Before we move on to deriving limits of quantities related to the empirical likelihood function, we will take the time to discuss one of the conditions in Theorem 5.2.1.

The lower bound

In Theorem 5.2.1 we imposed the following condition on the solution to

$$\begin{pmatrix} \mathrm{E}\{M(\mu)/[1+\lambda^T M(\mu)]\} \\ \Pr[1+\lambda^T M(\mu) \leq 0] \end{pmatrix} = 0.$$

There should exist a neighborhood of $\lambda(\mu)$, N, on which

$$\Pr\big(1+\lambda^T M(\mu) > L\big) = 1$$

for some $L > 0$ and all $\lambda \in N$. This condition is convenient when proving the result, but can be hard to check in practice as $\lambda(\mu)$ is generally unknown. In this section we will propose an alternative condition that is often easier to check.

Assume there exists a continuous function, B, such that

$$1 + \lambda^T M(\mu) \geq B(\lambda) > 0$$

for all $\lambda \in \Lambda_\mu$ and y in the support of Y. By assumption $B[\lambda(\mu)] > 0$ and since the function is continuous, there exists a neighborhood, N, of $\lambda(\mu)$ such that $B(\lambda) > B[\lambda(\mu)]/2$ for all $\lambda \in N$. This neighborhood together with the bound $L = B[\lambda(\mu)]/2$ satisfies the conditions of Theorem 5.2.1.

In many cases, proving the existence of such a function is easier than the existence of N and L directly. If, for instance, $m(y,\mu) = h(y) - \mu$ where the support of $h(Y)$ is $[a,b]$,

$$1 + \lambda(h(y) - \mu) \geq B(\lambda) = \begin{cases} 1+\lambda(b-\mu), & \lambda \leq 0 \\ 1+\lambda(a-\mu), & \lambda \geq 0 \end{cases}$$

for all y in the support of Y. This is a continuous function and since the support of $h(Y)$ is $[a, b]$, $\Pr[1 + \lambda(h(Y) - \mu) \leq 0] = 0$ if and only if

$$-\frac{1}{b-\mu} < \lambda < -\frac{1}{a-\mu}. \tag{5.13}$$

As long as (5.13) holds true, $B(\lambda) > 0$. Hence, B satisfies the conditions above, guaranteeing that N and L as in Theorem 5.2.1 exist.

When m is on the form $I(y \leq \mu) - q$ for some $q \in (0, 1)$, $m(y, \mu)$ takes the values $-q$ and $(1 - q)$. Hence

$$1 + \lambda[I(y \leq \mu) - q] \geq B(\lambda) = \begin{cases} 1 + \lambda(1-q), & \lambda \leq 0 \\ 1 - \lambda q, & \lambda \geq 0 \end{cases},$$

for all y in the support of Y. Furthermore, $\Pr\{1 + \lambda[I(y \leq \mu) - q] \leq 0\} = 0$ is equivalent to

$$-\frac{1}{1-q} < \lambda < \frac{1}{q}, \tag{5.14}$$

and under this condition, $B(\lambda) > 0$. So, as long as $\lambda(\mu)$ satisfies (5.14), N and L as in Theorem 5.2.1 exist.

5.3 The emprical likelihood function

We will now use Theorem 5.2.1 to provide an alternative characterization of the logarithm of the empirical likelihood function. In this section we will work with a fixed μ satisfying the conditions of Theorem 5.2.1. To improve readability we will omit the vector from the notation and use λ, λ_n, M_i and M as short-hand for $\lambda(\mu)$, $\lambda_n(\mu)$, $M_i(\mu)$ and $M(\mu)$ respectively.

By Theorem 5.2.1, the following holds with probability tending to 1:

$$\log \mathrm{EL}_n(\mu) = -\sum_{i=1}^{n} \log\bigl(1 + \lambda_n^T M_i\bigr).$$

Adding and subtracting $\sum_{i=1}^{n} \log(1 + \lambda^T M_i)$ from this expression, shows

$$\begin{aligned} \log \mathrm{EL}_n(\mu) &= -\sum_{i=1}^{n} \log\bigl(1 + \lambda^T M_i\bigr) + \sum_{i=1}^{n} \log\left(\frac{1 + \lambda^T M_i}{1 + \lambda_n^T M_i}\right) \\ &= -\sum_{i=1}^{n} \log\bigl(1 + \lambda^T M_i\bigr) + \sum_{i=1}^{n} \log\left(1 + (\lambda - \lambda_n)^T \frac{M_i}{1 + \lambda_n^T M_i}\right), \end{aligned}$$

with probability tending to 1. We can now use a second order Taylor expansion of $\log(1 + x)$ around 0 to see,

$$\sum_{i=1}^{n} \log\left(1 + (\lambda - \lambda_n)^T \frac{M_i}{1 + \lambda_n^T M_i}\right) = \tag{5.15}$$

$$(\lambda - \lambda_n)^T \sum_{i=1}^{n} \frac{M_i}{1 + \lambda_n M_i} - \frac{1}{2} \cdot (\lambda - \lambda_n)^T \left(\sum_{i=1}^{n} \frac{M_i M_i^T}{(1 + \lambda_n^T M_i)^2}\right)(\lambda - \lambda_n) + \epsilon_n \tag{5.16}$$

where ϵ_n is the sum of remainder terms in the Taylor expansions. Because λ_n solves

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{M_i}{1+\lambda^T M_i},$$

the first term of (5.16) disappears. This leaves us with the ensuing expression:

$$\sum_{i=1}^{n} \log\left(1 + (\lambda-\lambda_n)^T \frac{M_i}{1+\lambda_n^T M_i}\right) =$$
$$-\frac{1}{2}\cdot\sqrt{n}(\lambda-\lambda_n)^T\left(\frac{1}{n}\sum_{i=1}^{n}\frac{M_i M_i^T}{(1+\lambda_n^T M_i)^2}\right)\sqrt{n}(\lambda-\lambda_n)+\epsilon_n.$$

By Theorem 5.2.1 there exists a neighborhood, N, of λ on which

$$\Pr(1+x^T M > L) = 1$$

for all $x \in N$ and some $L > 0$. We will assume, without loss of generality, that N is convex and work with the expressions as though $\lambda_n \in N$. As λ_n is consistent for λ by Theorem 5.2.1, this holds with probability tending to 1.

We can now show that the remainder term ϵ_n goes to 0 in probability. On $[1,\infty)$ the third order derivative of $\log(1+x)$ is bounded by 2. So if

$$(\lambda-\lambda_n)^T \frac{M_i}{1+\lambda_n^T M_i} \geq 0,$$

the norm of the remainder term in the Taylor expansion of

$$\log\left(1 + (\lambda-\lambda_n)^T \frac{M_i}{1+\lambda_n^T M_i}\right)$$

around 0 is bounded by

$$\frac{2}{3!}\left|(\lambda-\lambda_n)^T \frac{M_i}{1+\lambda_n^T M_i}\right|^3.$$

Since $\Pr(1+\lambda_n^T M_i > L) = 1$, this is bounded by

$$\frac{1}{3L^4}\|\lambda-\lambda_n\|^3\|M_i\|^3.$$

If, on the other hand,

$$(\lambda-\lambda_n)^T \frac{M_i}{1+\lambda_n^T M_i} < 0,$$

one can show that the remainder term is bounded by

$$\frac{1}{3}\left|\frac{x}{1+x}\right|^3$$

where

$$x = (\lambda-\lambda_n)^T \frac{M_i}{1+\lambda_n^T M_i}.$$

Some algebra reveals

$$\frac{x}{1+x} = (\lambda - \lambda_n)^T \frac{M_i}{1+\lambda^T M_i}.$$

Hence, the norm of the remainder term is bounded by

$$\frac{1}{3}\left|(\lambda - \lambda_n)^T \frac{M_i}{1+\lambda^T M_i}\right|^3 \leq \frac{1}{3L}\|\lambda - \lambda_n\|^3 \|M_i\|^3.$$

Because of this,

$$|\epsilon_n| \leq \frac{1}{3L}\|\lambda - \lambda_n\|^3 \sum_{i=1}^n \|M_i\|^3.$$

As $\lambda - \lambda_n = O_{\mathrm{Pr}}(1/\sqrt{n})$, by Theorem 5.2.1, and the third moment of $\|M_i\|$ is finite by assumption, $\epsilon_n = O_{\mathrm{Pr}}(1/\sqrt{n})$. In particular, this implies that the logarithm of the empirical likelihood function is asymptotically equivalent to

$$-\sum_{i=1}^n \log\bigl(1+\lambda^T M_i\bigr) - \frac{1}{2}\cdot\sqrt{n}(\lambda-\lambda_n)^T \left(\frac{1}{n}\sum_{i=1}^n \frac{M_i M_i^T}{(1+\lambda_n^T M_i)^2}\right)\sqrt{n}(\lambda-\lambda_n).$$

As before let

$$V_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{M_i}{1+\lambda^T M_i} \quad \text{and} \quad S = \mathrm{E}\left(\frac{MM^T}{(1+\lambda^T M)^2}\right),$$

and define

$$S_n = \frac{1}{n}\sum_{i=1}^n \frac{M_i M_i^T}{(1+\lambda_n^T M_i)^2}.$$

Using this notation and the limit (5.11) from Theorem 5.2.1, we arrive at the following expression for the logarithm of the empirical likelihood function:

$$\log \mathrm{EL}_n(\mu) = -\sum_{i=1}^n \log\bigl(1+\lambda^T M_i\bigr) - \frac{1}{2}V_n^T S^{-1} S_n S^{-1} V_n + o_{\mathrm{Pr}}(1) \tag{5.17}$$

Let $f\colon [0,1] \to \mathbb{R}^{q^2}$ be defined as

$$f(t) = \frac{1}{n}\sum_{i=1}^n \frac{M_i M_i^T}{\{1+[t\lambda+(1-t)\lambda_n]^T M_i\}^2}.$$

For each $t \in [0,1]$,

$$\|f'(t)\| \leq \frac{1}{n}\sum_{i=1}^n \frac{\|M_i\|^3\|\lambda-\lambda_n\|}{\bigl|1+[t\lambda+(1-t)\lambda_n]^T M_i\bigr|^3} \leq \frac{\|\lambda-\lambda_n\|}{L^3}\frac{1}{n}\sum_{i=1}^n \|M_i\|^3.$$

Hence, by the mean value theorem for functions of several variables, see e.g. Lindstrøm 2017, p. 187–189,

$$\left\|S_n - \frac{1}{n}\sum_{i=1}^n \frac{M_i M_i^T}{(1+\lambda^T M_i)^2}\right\| = \|f(0)-f(1)\| \leq \frac{\|\lambda-\lambda_n\|}{L^3}\frac{1}{n}\sum_{i=1}^n \|M_i\|^3. \tag{5.18}$$

Here we have used that convexity of N ensures

$$\Pr\big(1 + [t\lambda + (1-t)\lambda_n]^T M > L\big) = 1$$

for all $t \in [0,1]$ and identified $q \times q$-matrices with vectors in $\mathbb{R}^{q^2}$. Since $\mathrm{E}\,\|M\|^3 < \infty$ and $\|\lambda - \lambda_n\| = O_{\Pr}(1/\sqrt{n})$, (5.18) implies

$$\left\| S_n - \frac{1}{n}\sum_{i=1}^{n} \frac{M_i M_i^T}{(1+\lambda^T M_i)^2} \right\| = O_{\Pr}(1/\sqrt{n}).$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{M_i M_i^T}{(1+\lambda^T M_i)^2}$$

converges to S by the law of large numbers. So,

$$S_n = S + o_{\Pr}(1).$$

Entering this into (5.17), shows

$$\log \mathrm{EL}_n(\mu) = -\sum_{i=1}^{n} \log\big(1+\lambda^T M_i\big) - \frac{1}{2} V_n^T S^{-1} V_n + o_{\Pr}(1), \tag{5.19}$$

as V_n has a normal limit by the central limit theorem and hence is stochastically bounded.

$V_n^T S^{-1} V_n$ converges in distribution, and $1/n$ converges to 0 in probability. Therefore, $V_n^T S_n^{-1} V_n / n \xrightarrow{\Pr} 0$ by Slutsky's theorem. Hence,

$$\frac{1}{n}\log \mathrm{EL}_n(\mu) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(1+\lambda^T M_i\big) + o_{\Pr}(1).$$

By the Law of large numbers, the right-hand side converges to its population version, and so

$$\frac{1}{n}\log \mathrm{EL}_n(\mu) \xrightarrow{\Pr} -\mathrm{E}\,\log\big(1+\lambda^T M\big).$$

For easier reference we will summarize the above findings.

Theorem 5.3.1. *Assume the conditions of Theorem 5.2.1 hold for $\mu \in \mathbb{R}^p$ and let $\lambda(\mu)$ denote the solution to*

$$\begin{pmatrix} \mathrm{E}\{M(\mu)/[1+\lambda^T M(\mu)]\} \\ \Pr[1+\lambda^T M(\mu) \le 0] \end{pmatrix} = 0.$$

As before, let

$$V_n(\mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{M_i(\mu)}{1+\lambda(\mu)^T M_i(\mu)} \quad \textit{and} \quad S(\mu) = \mathrm{E}\left(\frac{M(\mu)M(\mu)^T}{[1+\lambda(\mu)^T M(\mu)]^2}\right).$$

Then, the following holds with probability tending to 1

$$\log \mathrm{EL}_n(\mu) = -\sum_{i=1}^{n} \log\bigl(1 + \lambda(\mu)^T M_i(\mu)\bigr) - \frac{1}{2} V_n(\mu)^T S(\mu)^{-1} V_n(\mu) + \delta_n(\mu), \tag{5.20}$$

with $\delta_n(\mu)$ tending in probability to 0. In particular,

$$\frac{1}{n} \log \mathrm{EL}_n(\mu) \xrightarrow{\Pr} -\mathrm{E}\, \log\Bigl(1 + \lambda(\mu)^T M(\mu)\Bigr). \tag{5.21}$$

Since the expressions in Theorem 5.3.1 differ somewhat from the traditional formulations of theorems concerning empirical likelihood, we will comment briefly on how they compare to corresponding results in literature. To our knowledge, limits related to the empirical likelihood function at other values than the true parameter have not been derived before. However, by the main result of empirical likelihood (see Theorem 2.3.1 or Owen 2001, p. 41),

$$-2 \log \mathrm{EL}_n(\mu_0) \xrightarrow{d} \chi^2_q,$$

at the true parameter, μ_0, such that

$$\mathrm{E}\, M(\mu_0) = 0.$$

Applying Theorem 5.3.1 to this μ_0 results in the same conclusion, as we will now show.

Since $\mathrm{E}[M(\mu_0)] = 0$,

$$\mathrm{E}\left(\frac{M(\mu_0)}{1 + 0 \cdot M(\mu_0)} \right) = \mathrm{E}\, M(\mu_0) = 0.$$

Hence $\lambda(\theta_0) = 0$. Because of this (5.20) simplifies to

$$\begin{aligned} \log \mathrm{EL}_n(\mu) &= -\sum_{i=1}^{n} \log[1 + 0 \cdot M(\mu_0)] - \frac{1}{2} V_n(\mu_0)^T S(\mu_0)^{-1} V_n(\mu_0) + o_{\Pr}(1) \\ &= 0 - \frac{1}{2} V_n(\mu_0)^T S(\mu_0)^{-1} V_n(\mu_0) + o_{\Pr}(1), \end{aligned}$$

at the true parameter, μ_0. Furthermore, with $\lambda(\mu_0) = 0$, we get

$$V_n(\mu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{M(\mu_0)}{1 + 0 \cdot M(\mu_0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} M(\mu_0) \xrightarrow{d} \mathrm{N}_q(0, \Sigma),$$

and

$$S(\mu_0) = \mathrm{E}\left(\frac{M(\mu_0) M(\mu_0)^T}{[1 + 0 \cdot M(\mu_0)]^2} \right) = \mathrm{E}\bigl[M(\mu_0) M(\mu_0)^T \bigr] = \Sigma.$$

So, at the true parameter, we do, indeed, get

$$-2 \log \mathrm{EL}_n(\mu_0) \xrightarrow{d} \chi^2_q$$

from Theorem 5.3.1, just as when using the classic result. Furthermore the expression

$$-2 \log \mathrm{EL}_n(\mu_0) = V_n(\mu_0) S(\mu_0)^{-1} V_n(\mu_0) + o_{\Pr}(1)$$

is asymptotically equivalent to the one given in, e.g. Schweder and Hjort 2016, p. 328.

5.4 Investigating the solution to the Lagrange equation

In the previous sections, we derived limits for the empirical likelihood function. In the proofs, as well as statements of theorems, a function defined implicitly as the solution to

$$\begin{pmatrix} \mathrm{E}\{M(\mu)/[1+\lambda^T M(\mu)]\} \\ \Pr[1+\lambda^T M(\mu) \leq 0] \end{pmatrix} = 0, \tag{5.22}$$

is used. In this section we will discuss the existence, and behavior, of this solution.

We will start by going through an example investigating the solution to (5.22) in a specific situation. Afterwards, we state and prove a theorem about the existence of such solutions and discuss its implications when the estimating function is on the form $m(y,\mu) = h(Y) - \mu$. Lastly, a discussion concerning necessity of the last entry in (5.22), as well as a strategy for what to do when 5.22 cannot be solved, is provided.

A first example

Assume Y follows a uniform distribution on the unit interval, and let $m(y,\mu) = y - \mu$. For $\lambda = 0$,

$$\mathrm{E}\left(\frac{Y-\mu}{1+0\cdot(Y-\mu)}\right) = \mathrm{E}\,Y - \mu = \frac{1}{2} - \mu.$$

Otherwise, standard integration techniques results in the following expression

$$\mathrm{E}\left(\frac{Y-\mu}{1+\lambda(Y-\mu)}\right) = \frac{1}{\lambda} - \frac{1}{\lambda^2}\log\left(\frac{1+\lambda(1-\mu)}{1-\lambda\mu}\right). \tag{5.23}$$

We can now solve

$$0 = \mathrm{E}\left(\frac{Y-\mu}{1+\lambda(Y-\mu)}\right) \tag{5.24}$$

for a fixed $\mu \neq 1/2$ by numerically finding the root of (5.23). In Figure 5.1 we have plotted the solutions for a selection of μ-values.

For each fixed $\mu \in (0,1)$ we need the solution to (5.24) to lie in the set

$$\Lambda_\mu = \{\, \lambda \in \mathbb{R} \mid \Pr[1+\lambda(Y-\mu) \leq 0] = 0 \,\} \tag{5.25}$$

for it to solve (5.22). Since Y is uniformly distributed on $[0,\,1]$, $\Pr[1+\lambda(Y-\mu) \leq 0] = 0$ is satisfied if

$$1+\lambda(0-\mu) > 0 \quad \text{and} \quad 1+\lambda(1-\mu) > 0.$$

This is equivalent to requiring

$$\frac{1}{\mu-1} < \lambda < \frac{1}{\mu}. \tag{5.26}$$

In Figure 5.1 the shaded area is the set of λs satisfying this inequality for the corresponding values of μ. From the plot we see that the solution to (5.24)

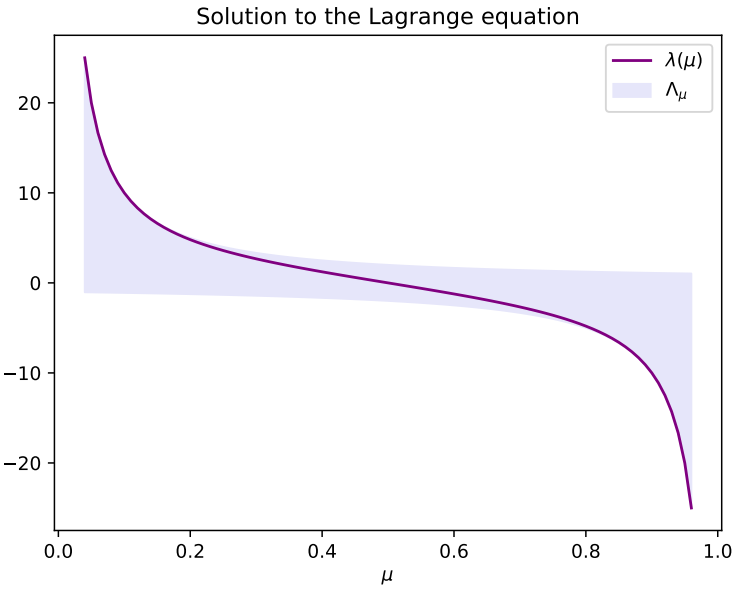


Figure 5.1: A plot of the solution to (5.24) as a function of μ. The shaded area framed indicates where the inequalities in (5.25) hold true.

satisfies (5.26) for each μ. All solutions to (5.24) are therefore solutions to (5.22). Furthermore, as μ moves away from $\mu_0 = 1/2$, $\lambda(\mu)$ approaches the boundary of Λ_μ.

Inspecting the graph, we notice that for $\mu < \mu_0$, the solution to (5.24) is negative, while $\mu > 1/2$ results in a positive root. We also see that the zeros are continuous as a function of μ. These properties actually hold quite generally and can be confirmed theoretically. For each $\mu \in (0, 1)$, we can apply Leibniz integral theorem, see e.g. Lindstrøm 2017, p. 276, to get

$$\frac{\partial}{\partial \mu} \mathrm{E}\left(\frac{Y - \mu}{1 + \lambda(Y - \mu)} \right) = \frac{-1}{[1 + \lambda(Y - \mu)]^2} < 0$$

and

$$\frac{\partial}{\partial \lambda}\Big|_{\lambda(\mu)} \mathrm{E}\left(\frac{Y - \mu}{1 + \lambda(Y - \mu)} \right) = \mathrm{E}\left(-\frac{(Y - \mu)^2}{[1 + \lambda(\mu)(Y - \mu)]^2} \right) < 0.$$

Hence, by the implicit function theorem, see e.g. Lindstrøm 2017, p. 212, the solutions to

$$\mathrm{E}\left(\frac{Y - \mu}{1 + \lambda(Y - \mu)} \right)$$

are differentiable as a function of μ with derivative

$$\lambda'(\mu) = -\frac{\partial}{\partial \mu} \mathrm{E}\left(\frac{Y - \mu}{1 + \lambda(Y - \mu)} \right) \left[\frac{\partial}{\partial \lambda}\Big|_{\lambda(\mu)} \mathrm{E}\left(\frac{Y - \mu}{1 + \lambda(Y - \mu)} \right) \right]^{-1} < 0$$

when they exist. Since $\lambda(\mu_0) = 0$, this implies that $\lambda(\mu)$ is positive for $\mu < \mu_0$ and negative for $\mu > \mu_0$.

The above arguments can easily be modified to similar situations. In many cases the solution to (5.22) will therefore have similar properties as a function of μ to what we have seen here. That being said, analyzing λ is often more complicated than what we have done in this example. The reason for this is that it is often hard, if not impossible, to compute

$$\mathrm{E}\left(\frac{Y-\mu}{1+\lambda(Y-\mu)}\right)$$

analytically. In such cases numerical integration techniques, or approximation of infinite sums by finite ones, can be used to estimate this quantity.

Solving the Lagrange equation

In the previous example we saw that

$$\begin{pmatrix} \mathrm{E}\{M(\mu)/[1+\lambda^T M(\mu)]\} \\ \Pr[1+\lambda^T M(\mu) \leq 0] \end{pmatrix} = 0 \tag{5.27}$$

could be solved for all μ in the support of Y. This property actually holds for many bounded distributions. We will now give a lemma proving, and disproving, the existence of solutions to (5.27) under certain conditions.

Lemma 5.4.1. *Let $m\colon \mathbb{R}^{d+p} \to \mathbb{R}$ be a one dimensional estimating function and fix $\mu \in \mathbb{R}^p$. Let M denote the stochastic variable $m(Y,\mu)$ and $f\colon \mathbb{R} \to \mathbb{R}$ the function*

$$f(\lambda) = \mathrm{E}\left(\frac{M}{1+\lambda M}\right).$$

The following then holds:

(1) *Assume $\mathrm{E}\, M > 0$ and that $M \geq a$ for some $a < 0$ with probability 1. If f is continuous on $[0, -1/a)$, $f(\lambda) = 0$ has a solution in $(-\infty, -1/a)$ if and only if*

$$\lim_{\lambda \to -1/a} \mathrm{E}\left(\frac{1}{1+\lambda M}\right) > 1, \tag{5.28}$$

where the limit is taken from below.

(2) *Assume $\mathrm{E}\, M < 0$ and that $M \leq b$ for some $b > 0$ with probability 1. If f is continuous on $(-1/b, 0]$, $f(\lambda) = 0$ has a solution in $(-1/b, \infty)$ if and only if*

$$\lim_{\lambda \to -1/b} \mathrm{E}\left(\frac{1}{1+\lambda M}\right) > 1,$$

where the limit is taken from above.

Proof. We will only show (1) as (2) follows from this case after replacing M with $-M$, so assume the conditions of (1) hold.

The function

$$\lambda \mapsto \frac{x}{1+\lambda x}$$

is decreasing on $(-\infty, -1/x)$ for all fixed x. Hence

$$\lambda \mapsto \frac{M}{1+\lambda M}$$

is decreasing on the set $(-\infty, -1/a)$ as $M \geq a$ implies that $-1/M \geq -1/a$. Because of this, and monotonicity of expected values, f is a decreasing function on $(-\infty, -1/a)$. Since $f(0) = \mathrm{E}\, M > 0$ by assumption, this implies that f has a root on this interval if and only if it has a positive root. It therefore suffices to check for such a zero.

We notice that

$$f(\lambda) = \mathrm{E}\left(\frac{M}{1+\lambda M}\right) = \frac{1}{\lambda} - \frac{1}{\lambda}\mathrm{E}\left(\frac{1}{1+\lambda M}\right).$$

Hence,

$$\lim_{\lambda \to -1/a} f(\lambda) = -a\left[1 - \lim_{\lambda \to -1/a} \mathrm{E}\left(\frac{1}{1+\lambda M}\right)\right]$$

Assume first that (5.28) holds. Then

$$\lim_{\lambda \to -1/a} f(\lambda) < -a(1-1) = 0,$$

where the limit is taken from below. As $f(0) = \mathrm{E}\, M > 0$ by assumption and f is continuous on $[0, -1/a)$, the intermediate value theorem ensures that f has a root in $\lambda \in [0, -1/a)$.

If, on the other hand,

$$\lim_{\lambda \to -1/a} \mathrm{E}\left(\frac{1}{1+\lambda M}\right) \leq 1,$$

we have

$$\lim_{\lambda \to -1/a} f(\lambda) \geq -a(1-1) = 0.$$

Then $f(\lambda) > 0$ for all $\lambda \in [0, -1/a)$ as it is a decreasing function. Hence, no solution to $f(\lambda) = 0$ exists in $[0, -1/a)$. This concludes the proof. ■

Assume $m(y, \mu) = I(y \leq \mu) - q$ for some $q \in (0, 1)$. Then $-q \leq m(Y, \mu) < 1 - q$ regardless of the distribution on Y. Furthermore

$$f(\lambda) = -\Pr(Y \leq \mu)\frac{q}{1-\lambda q} + [1 - \Pr(Y \leq \mu)]\frac{1-q}{1+\lambda(1-q)},$$

which is continuous on the interval $(-1/(1-q), 1/q)$. Lastly,

$$\mathrm{E}\left(\frac{1}{1+\lambda[I(y \leq \mu) - q]}\right) = \Pr(Y \leq \mu)\frac{1}{1-\lambda q} + [1 - \Pr(Y \leq \mu)]\frac{1}{1+\lambda(1-q)}.$$

As λ goes to both $1/q$ from below and to $-1/(1-q)$ from above, this expression goes to infinity. Hence, the conditions of (1) and (2) both hold for all μ. Because of this Lemma 5.4.1 guarantees that $f(\lambda) = 0$ can be solved in the set $-1/(1-q) < \lambda < 1/q$ for all μ. Since,

$$\Pr\{1 + \lambda[I(Y \leq \mu) - q] \leq 0\} = 0 \quad \Leftrightarrow \quad -\frac{1}{1-q} < \lambda < \frac{1}{q},$$

this shows that

$$\begin{pmatrix} \mathrm{E}\{[I(Y \leq \mu) - q]/[1 + \lambda(I(Y \leq \mu) - q)]\} \\ \Pr\{1 + \lambda[(Y \leq \mu) - q] \leq 0\} \end{pmatrix} = 0$$

can be solved for all values of μ. We can also see this directly as the solution is given explicitly as

$$\lambda(\mu) = \frac{F(\mu) - q}{q(1-q)}.$$

With the estimating function $m(y, \mu) = h(y) - \mu$, the existence of solutions varies with the distribution of the data. To get a better idea of when solutions can and cannot be found, we will go through some specific examples and situations with this estimating function.

The solution to the Lagrange equation for means

In this section we will discuss the existence of solutions to

$$\begin{pmatrix} \mathrm{E}\{(Y - \mu)/[1 + \lambda(Y - \mu)]\} \\ \Pr[1 + \lambda(Y - \mu) \leq 0] \end{pmatrix} = 0 \tag{5.29}$$

in different situations. We will let f_μ denote the function

$$f_\mu(\lambda) = \mathrm{E}\left(\frac{Y - \mu}{1 + \lambda(Y - \mu)}\right)$$

for each μ. This is a decreasing function for each fixed μ as is shown in the proof of Lemma 5.4.1.

It is worth noting that replacing Y with $h(Y)$ allows the analysis of this section to be applied to all estimating functions on the form $m(y, \mu) = h(y) - \mu$ for functions $h\colon \mathbb{R}^d \to \mathbb{R}$.

Situation 1 - When the support of Y is bounded

We start with the case when the support of Y is bounded. So let $Y \in [L, U]$ with probability 1. For each fixed $\mu \in (L, U)$, we then have

$$L - \mu \leq Y - \mu \leq U - \mu$$

with $L - \mu < 0$ and $U - \mu > 0$. Hence, for each fixed μ, $\Pr[1 + \lambda(Y - \mu) \leq 0] = 0$ is equivalent to

$$-\frac{1}{U - \mu} < \lambda < -\frac{1}{L - \mu}. \tag{5.30}$$

So (5.29) has a solution if and only if $f_\mu(\lambda) = 0$ can be solved under condition (5.30).

Let $\mu < \mathrm{E}\, Y$. Then $f_\mu(0) = \mathrm{E}(Y - \mu) > 0$. As f_μ is decreasing, any zero of f_μ must be positive. Because of this, (5.29) can be solved if and only if f_μ has a root in $[0, -1/(L - \mu))$. By case (1) of Lemma 5.4.1 this happens if and only if

$$\lim_{\lambda \to -1/(U-\mu)} \mathrm{E}\left(\frac{1}{1 + \lambda(Y - \mu)} \right) > 1$$

After some algebra, we see that this is equivalent to

$$\mu > U - \left[\mathrm{E}\left(\frac{1}{U - Y} \right) \right]^{-1}.$$

A similar relation holds true when $\mu > \mathrm{E}\, Y$. Since f_μ is decreasing, all potential roots must be negative. By case (2) and (5.30) this happens if and only if

$$\lim_{\lambda \to -1/(L-\mu)} \mathrm{E}\left(\frac{1}{1 + \lambda(Y - \mu)} \right) > 1.$$

Which can be shown to be equivalent to

$$\mu < L + \left[\mathrm{E}\left(\frac{1}{Y - L} \right) \right]^{-1}.$$

In the uniform distribution on [0, 1], both $\mathrm{E}\, Y^{-1}$ and $\mathrm{E}(1 - Y)^{-1}$ diverges to ∞. Hence Lemma 5.4.1 guarantees that a solution to (5.27) exists for all $\mu \in (0, 1)$. This is also what we found in Section 5.4.

Assume now that Y follows a Beta$(2, 2)$ distribution. Numerical integration shows that in this case

$$\mathrm{E}\left(\frac{1}{Y} \right) = \mathrm{E}\left(\frac{1}{1 - Y} \right) = 3$$

The analysis in this section therefore guarantees that a solution to (5.27) should exist only for $\mu \in (1/3,\ 2/3)$. In Figure 5.2, we have displayed the numerically found roots of

$$\mathrm{E}\left(\frac{Y - \mu}{1 + \lambda(Y - \mu)} \right), \tag{5.31}$$

together with the bounds given in (5.30). From the figure, we see that a solution was found only for the expected range of μ-values.

Assume now that Y follows a Beta$(1, 3)$ distribution. As the density in this distribution is non-zero at 0,

$$\mathrm{E}\left(\frac{1}{Y} \right) = \infty.$$

This will be shown in the next section. Because of this (5.27) should have a solution for all $\mu \in (0, \mu_0]$ where $\mu_0 = \mathrm{E}\, Y = 1/4$. On the other hand

$$\mathrm{E}\left(\frac{1}{1 - Y} \right) = \frac{3}{2}.$$

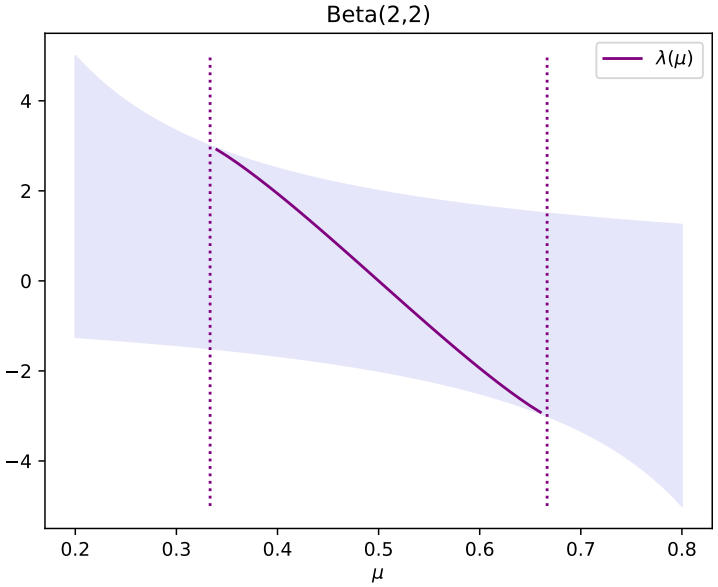


Figure 5.2: A plot of the zeros of (5.31) found numerically as a function of μ. The shaded area indicates the set of λs satisfying (5.30) for different μs. The dotted lines frame the area where $1/3 < \mu < 2/3$, the range for which a solution should theoretically exists.

So, by the above analysis, a solution to (5.27) should only exist when $\mu < 1/3$ for $\mu > \mu_0$. Combining all of this, we get that a solution should exist if and only if $0 < \mu < 1/3$. We have again attempted to find roots of

$$\mathrm{E}\left(\frac{Y-\mu}{1+\lambda(Y-\mu)}\right) \tag{5.32}$$

numerically for all $\mu \in (0,1)$. The result can be fund in Figure 5.3 and agrees with our theoretical analysis.

Situation 2 - When the support of Y is unbounded in one direction

Let Y have support $[L, \infty)$ for some $L \in \mathbb{R}$. As Y is unbounded to the right, $\Pr[1+\lambda(Y-\mu) \leq 0] > 0$ for all $\lambda < 0$. Because of this, we must have $\lambda > 0$ to ensure that the last entry in the Lagrange equation is zero. Since f_μ is a decreasing function, $f_\mu(\lambda) = 0$ does not have a solution satisfying $\Pr[1+\lambda(Y-\mu) \leq 0] > 0$ when $f_\mu(0) > 0$. As $f_\mu(0) = \mathrm{E}\,Y - \mu$, this implies that (5.27) cannot be solved for $\mu > \mu_0$. The opposite is of course true when the support of Y is $(-\infty, U]$ for some $U \in \mathbb{R}$.

In Figure 5.4 we have displayed numerically obtained solutions to $f_\mu(\lambda) = 0$ in two situations. The purple line is the graph of zeros of f_μ when Y follows a Gamma distribution with shape 2 and rate 2. The orange line is a plot of the roots of f_μ when Y instead is exponentially distributed with rate parameter 1. The mean in both of these distributions is 2, but while $\mathrm{E}\,Y^{-1} = 2$ when

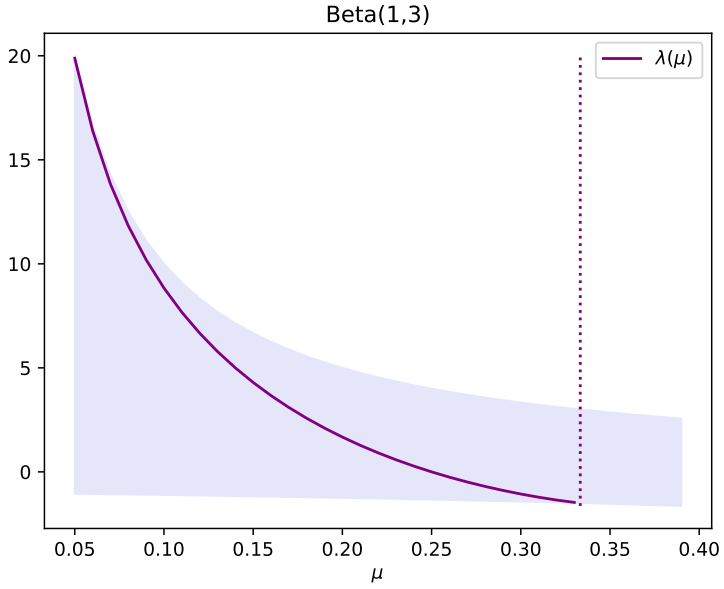


Figure 5.3: A plot of the zeros of (5.32) found numerically as a function of μ. The shaded area indicates the set of λs satisfying (5.30) for different μs and the dotted line where $\mu = 1/3$. This is the theoretical computed upper bound for where a solution should exist.

$Y \sim \text{Gamma}(2, 2)$, the integral diverges if $Y \sim \text{Expo}(1)$. From the figure we see that no solution was found when $\mu > \mathrm{E}\, Y = 2$ for both distributions. In addition, we see that while a solution was found for all $\mu < \mu_0$ in the case of the exponential distribution with rate parameter $1/2$, $f_\mu(\lambda) = 0$ could only be solved when $\mu > 1/\mathrm{E}\, Y^{-1} = 1/2$ for the case of $Y \sim \text{Gamma}(2, 2)$. This is in agreement with our theoretical analysis.

Situation 3 - When the support of Y is unbounded in both directions

When the support of Y is unbounded in both directions, $\Pr[1+\lambda(Y-\mu) \leq 0] > 0$ for all $\lambda \neq 0$. Because of this, no solutions to (5.27) can be found unless

$$\mathrm{E}\left(\frac{Y-\mu}{1+0\cdot(Y-\mu)}\right) = \mathrm{E}\, Y - \mu = 0,$$

and this only holds if μ is the true parameter. Hence (5.27) can only be solved when $\mu = \mathrm{E}\, Y$. This makes it impossible to apply the limit results derived in this chapter. In Section 5.4 we will discuss the use of truncated distributions to deal with this problem.

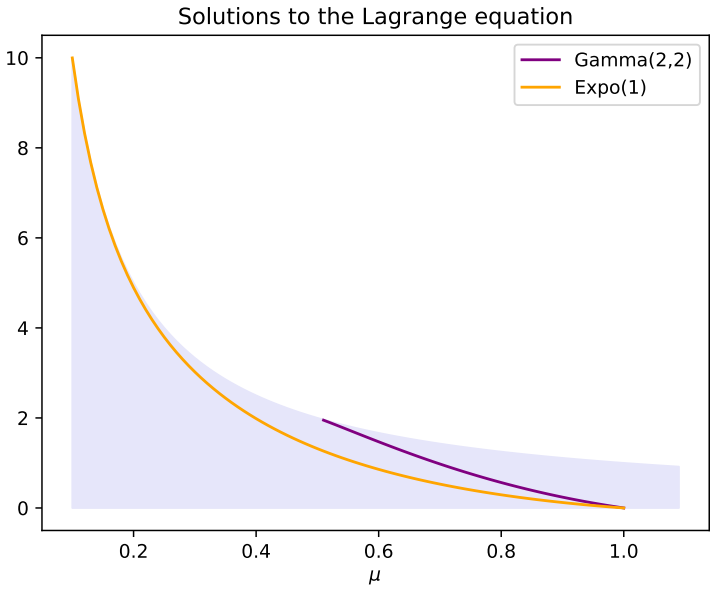


Figure 5.4: A plot of the solutions to $f_\mu(\lambda) = 0$ found numerically as a function of μ in two cases: when $Y \sim \text{Gamma}(2, 2)$ and $Y \sim \text{Expo}(1)$. The shaded area indicates the set of λs satisfying $\Pr[1 + \lambda(Y - \mu) \leq 0] = 0$ for different μs.

Do we need the extra condition?

In the previous examples we found that the roots of

$$0 = \mathrm{E}\left(\frac{M}{1 + \lambda M}\right) \tag{5.33}$$

all satisfied $\Pr(1 + \lambda M \leq 0) = 0$. Because of this, one might ask if we need to impose the extra condition of $\Pr(1 + \lambda(\mu)M \leq 0) = 0$ on the solution, $\lambda(\mu)$. In many cases it is true that this condition is unnecessary. For instance, fix μ and let M follow a continuous distribution with density g. Let $\lambda \neq 0$ and assume $g(-1/\lambda) \neq 0$. Then there exists $\epsilon > 0$ such that the density is bounded from below by $L > 0$ on $[-1/\lambda - \epsilon, -1/\lambda + \epsilon]$. Hence,

$$\mathrm{E}\left|\frac{M}{1 + \lambda M}\right| \geq L \int_{-1/\lambda - \epsilon}^{-1/\lambda + \epsilon} \left|\frac{x}{1 + \lambda x}\right| dx,$$

and this integral diverges. Because of this (5.33) is undefined for many λs that do not satisfy $\Pr(1 + \lambda M \leq 0) = 0$ when the true underlying distribution of the data is continuous. In this case, all roots to (5.33) automatically satisfy the property $\Pr(1 + \lambda M \leq 0) = 0$. That being said, there are cases for which

$$\mathrm{E}\left(\frac{M}{1 + \lambda M}\right) = 0,$$

but $\Pr[1 + \lambda M \leq 0] > 0$. To see this let Y take the values 1 and $1/3$ with probability $1/2$ each. Then

$$\mathrm{E}\left(\frac{Y}{1-2Y}\right) = \frac{1}{2}\cdot\frac{1}{1-2} + \frac{1}{2}\cdot\frac{1/3}{1-2\cdot 1/3} = -\frac{1}{2} + \frac{1}{2} = 0,$$

but

$$\Pr(1 - 2Y \leq 0) = \Pr(Y = 1) = 1/2 > 0.$$

So in general,

$$\mathrm{E}\left(\frac{M}{1+\lambda M}\right)$$

can have roots that do not satisfy $\Pr(1 + \lambda M \leq 0) = 0$.

Distributions with unbounded support

Looking at Lemma 5.4.1, we see that there is one very important situation with a surprising result: when Y is normally distributed and the estimating function is $m(y, \mu) = y - \mu$. This is, perhaps, the simples and most standard situation one can come up with, but since the support of Y is unbounded in both directions,

$$\Pr[1 + \lambda(Y - \mu) \leq 0] > 0,$$

for any $\lambda \neq 0$. Because of this, the set Λ_μ defined in Theorem 5.2.1 consists of a single element, 0, for all μ. Furthermore,

$$\mathrm{E}\left(\frac{Y-\mu}{1+0\cdot(Y-\mu)}\right) = \mu_0 - \mu$$

is equal to 0 if and only if $\mu = \mu_0$. Hence

$$\begin{pmatrix}\mathrm{E}\{(Y-\mu)/[1+\lambda(Y-\mu)]\}\\ \Pr[1+\lambda(Y-\mu)\leq 0]\end{pmatrix} = 0.$$

can only be solved at the true parameter. This is an example of a general trend. If the support of $M(\mu)$ is unbounded, there are many values of μ for which

$$\begin{pmatrix}\mathrm{E}\{M(\mu)/[1+\lambda^T M(\mu)]\}\\ \Pr[1+\lambda^T M(\mu)\leq 0]\end{pmatrix} = 0.$$

has no solution. In such cases the theorems concerning limits of the empirical likelihood function and related quantities derived in this chapter cannot be applied.

One solution to this problem is to work with truncated distributions. In practice, there is very little difference between, for instance, a standard normal distribution and the truncated version with support equal to $[-10, 10]$. Hence, the empirical likelihood function constructed with data following these two distributions will behave very similarly. In Figure 5.5 we have plotted $n^{-1}\log \mathrm{EL}_n(\mu)$ with $m(y, \mu) = y - \mu$ and data following a central normal

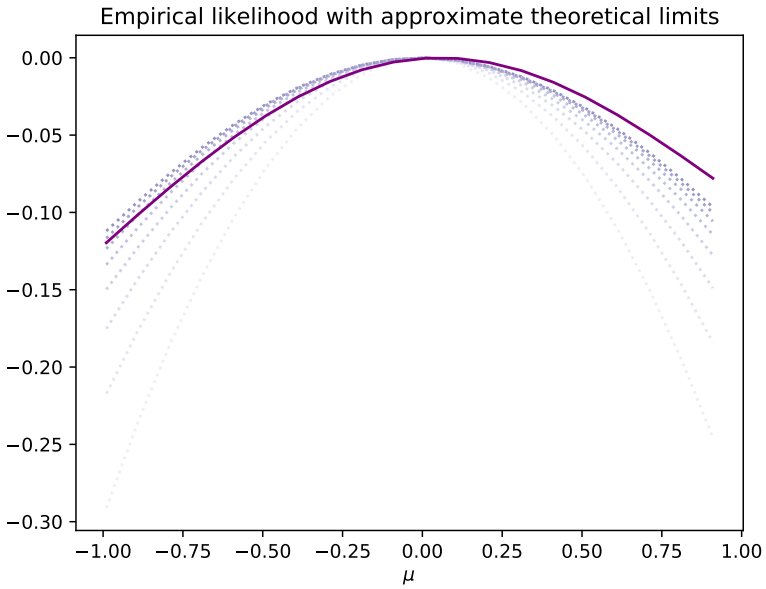


Figure 5.5: The purple line is a plot of the empirical likelihood function, scaled by the sample size, for the mean constructed with 500 simulated data points from the central normal distribution with standard deviation 2. The blue dotted lines are the limits of this function under the assumption that the true underlying distribution is the truncated version of $\mathrm{N}(0, 4)$ on $[-K, K]$ for $K = 2.5, 3, 3.5, \ldots, 6$. The opacity of the lines increases with K.

distribution with standard deviation 2 together with the theoretical limits of this function, for the truncated normal distribution with support on $[-K, K]$ for different values of K. From the plot we see that the theoretical curves draw closer as K increases. Furthermore, the empirical likelihood function is not far off from the theoretical curves.

The trend we see in Figure 5.5 is the typical one. In many cases, approximating unbounded distributions with truncated versions solves the problem with the lack of solution to

$$\begin{pmatrix} \mathrm{E}\{M(\mu)/[1 + \lambda^T M(\mu)]\} \\ \Pr[1 + \lambda^T M(\mu) \leq 0] \end{pmatrix} = 0$$

and gives good approximate results.

CHAPTER 6

The maximum empirical likelihood estimator

The limit results derived in the previous chapter are interesting in their own rights. They provide intuition about what goes on behind the scenes of the empirical likelihood machinery. Furthermore, they guarantee that the theory we know about the empirical likelihood function at the true parameter, generalizes to alternatives. That being said, the results are more than just interesting theory. By Theorem 5.3.1, the empirical likelihood function is very close to an empirical mean. Because of this, its maximizer is almost what we call an M-estimator, the maximizer of functions on the form

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\psi(\theta)$$

where $\psi\colon \mathbb{R}^p \to \mathbb{R}$. A general introduction to this concept can be found in chapter 5 of Vaart 1998.

M-estimators are widely studied, both in the context of maximum likelihood theory and robust statistics. Because of this, there is a loot of theory to build, on and we will now use this theory to show both consistency and asymptotic normality of the maximizer of the empirical likelihood function. In this section we will assume that the estimating equation is one dimensional. This simplifies many of the proofs, but the theorems should be possible to generalize to higher dimensions. In that case, however, many of the arguments require the use of more advanced matrix calculus than what is considered here. This is particularly true for the proofs utilizing Taylor expansions.

Limits of the maximum empirical likelihood estimator, and scaled and centered versions of it, has been derived before, see e.g. Qin and Lawless 1994. Furthermore, the maximizer of the empirical likelihood function will often be the solution to the empirical equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} m(Y_i, \mu),$$

and this is both consistent and asymptotically normal under very general assumptions, se e.g. section 3.2 in Huber 2009. Nevertheless, we will provide our own proofs here based on Theorem 5.2.1 and Theorem 5.3.1. This will serve as an illustration of how the results from the previous chapter can be applied.

Furthermore, the lemmas and theorems proved in this chapter will be useful in Chapter 8, where we will work with the maximizer of a hybrid combination of empirical and parametric likelihoods.

As before, we will let $Y_1, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. variables, following some unknown distribution, F, throughout this chapter. We will also let $m\colon \mathbb{R}^{d+p} \to \mathbb{R}$ be a function and μ_0 the unique p-dimensional vector such that,

$$\mathrm{E}\, m(Y, \mu_0) = 0$$

for $Y \sim F$. Lastly, $M_i(\mu)$ will be short-hand for $m(Y_i, \mu)$ and $M(\mu)$ for $m(Y, \mu)$ where $Y \sim F$.

6.1 The remainder term

Before we can show consistency and asymptotic normality of the maximum empirical likelihood estimator, we need to show that the remainder term in Theorem 5.3.1 goes quickly enough to zero in probability. This will be the topic for this section. Although some additional level of detail will be required here compared to the previous chapter, the ideas behind the proofs are essentially the same now as in Chapter 5. Furthermore, most of the assumptions and results can be seen as stronger versions of those given in the previous chapter.

Our first goal will be to show that $\lambda_n(\mu)$ converges uniformly to $\lambda(\mu)$ in probability over compact sets. In Theorem 5.2.1 we imposed the following condition on $\lambda(\mu)$. We needed there to be a neighborhood of the vector on which

$$\Pr[1 + \lambda m(Y, \mu) > L] = 1$$

for all λ. To show uniform consistency of $\lambda_n(\mu)$ towards $\lambda(\mu)$ in a compact set, $\mathcal{M}$, we will need a similar, but stronger condition. We will assume there exists some $\delta > 0$ such that, for all $\mu \in \mathcal{M}$,

$$\Pr[1 + \lambda m(Y, \mu) > L] = 1$$

for all λs in an open ball centered at $\lambda(\mu)$ with radius δ. This might seem like a strict conditions, but, under sufficient smoothness of the function $(\lambda, \mu) \mapsto B(\lambda, \mu)$ defined in Section 5.2 and λ, this does, indeed, hold.

To illustrate how this condition can be confirmed, we will show that it holds when $m(y, \mu) = h(y) - \mu$ and $h(Y)$ has support in $[a, b]$. We will take $\mathcal{M} = [a, b]$, as showing that the condition holds in this set is sufficient for it to hold in all compact subsets of $[a, b]$. Since $a \leq h(Y) \leq b$ with probability 1,

$$1 + \lambda[h(Y) - \mu] > B(\lambda, \mu) = \begin{cases} 1 + \lambda(b - \mu), & \lambda \leq 0 \\ 1 + \lambda(a - \mu), & \lambda \geq 0 \end{cases}.$$

If λ is a continuous function, $\mu \mapsto B[\lambda(\mu), \mu]$ is continuous as well. Furthermore, $B[\lambda(\mu), \mu] > 0$ for each $\mu \in [a, b]$. Hence, there exists a strictly positive number, L_1, such that $B[\lambda(\mu), \mu] \geq L_1$ for each $\mu \in [a, b]$. Furthermore, Since the support of $h(Y)$ is bounded, $|h(Y) - \mu| \leq K$ for some $K > 0$ with probability 1 and for all $\mu \in [a, b]$. Let $\delta = L_1/2K$. For every $\mu \in [a, b]$ and every λ in the open ball centered at $\lambda(\mu)$ with radius δ, we then have

$$1 + \lambda[h(Y) - \mu] = 1 + \lambda(\mu)[h(Y) - \mu] + [\lambda - \lambda(\mu)][h(Y) - \mu]$$

$$
\begin{aligned}
&> L_1 - \delta \cdot K \\
&= \frac{L_1}{2}
\end{aligned}
$$

with probability 1.

With the above condition, uniform consistency of $\lambda_n(\mu)$ towards $\lambda(\mu)$ is not to hard to prove.

Lemma 6.1.1. *Let $\mathcal{M}$ be a compact set such that the equation*

$$
\begin{pmatrix} \mathrm{E}\{M(\mu)/[1+\lambda M(\mu)]\} \\ \Pr[1+\lambda M(\mu) \leq 0] \end{pmatrix} = 0
$$

has a solution, $\lambda(\mu)$, for each $\mu \in \mathcal{M}$. Assume further that there exists $\delta > 0$ and $L > 0$ such that for all (λ, μ) with $\mu \in \mathcal{M}$ and $|\lambda(\mu) - \lambda| < \delta$,

$$
\Pr[1 + \lambda M(\mu) > L] = 1.
$$

Let $\lambda_n(\mu)$ be a solution to

$$
0 = \Psi_n(\lambda, \mu) = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i(\mu)}{1 + \lambda M_i(\mu)}
$$

and assume that, for almost all y, $\mu \mapsto m(y, \mu)$ is a continuous function with

$$
|m(y, \mu)| \leq p_1(y)
$$

for all $\mu \in \mathcal{M}$ and some p_1 satisfying $\mathrm{E}\, p_1(Y) < \infty$. Then

$$
\sup_{\mu \in \mathcal{M}} |\lambda_n(\mu) - \lambda(\mu)| \xrightarrow{\Pr} 0.
$$

Proof. This proof is slight modification of lemma 5.10 in Vaart 1998, p. 47.

We will start by deriving some properties of the functions λ and

$$
\Phi(\lambda, \mu) = \mathrm{E}\left(\frac{M(\mu)}{1 + \lambda M(\mu)} \right).
$$

Fix $\mu \in \mathcal{M}$ and some $\lambda \in (\lambda(\mu) - \delta, \lambda(\mu) + \delta)$. Let x_n be a sequence converging to this λ. Since $(\lambda(\mu) - \delta, \lambda(\mu) + \delta)$ is a neighborhood of λ, x_n will eventually lie in this set. So, for large enough n and almost all y,

$$
\left| \frac{m(y, \mu)}{1 + x_n m(y, \mu)} \right| \leq \frac{|m(y, \mu)|}{L}.
$$

This is a function with finite expectation, not depending on λ. Because of this, Lebesgue's dominated convergence theorem, see e.g. McDonald and Weiss 2013, p. 169, can be applied to show

$$
\lim_{n \to \infty} \Phi(x_n, \mu) = \Phi\left(\lim_{n \to \infty} x_n, \mu \right) = \Phi(\lambda, \mu).
$$

This shows that $\lambda \mapsto \Phi(\lambda, \mu)$ is continuous in the set $(\lambda(\mu) - \delta, \lambda(\mu) + \delta)$ for each fixed μ. Furthermore, for every y and μ,

$$
\lambda \mapsto \frac{m(y, \mu)}{1 + \lambda m(y, \mu)}
$$

is a strictly decreasing function on

$$\Lambda_\mu = \{ \lambda \in \mathbb{R} \mid \Pr[1 + \lambda M(\mu) \leq 0] = 0 \}.$$

Expectation is monotone, so this implies that $\lambda \mapsto \Phi(\lambda, \mu)$ is a strictly decreasing function. Hence,

$$\lambda \mapsto \Phi(\lambda, \mu)$$

is one-to-one, and $\lambda(\mu)$ is the unique root of the function for each fixed μ. Combining this with continuity of $\lambda \mapsto \Phi(\lambda, \mu)$ on $(\lambda(\mu) - \delta, \lambda(\mu) + \delta)$ and Corollary 1.1 in Kumagai 1980, ensures that λ is continuous as a function of μ.

Define the following set

$$A = \{ (\lambda, \mu) \mid \mu \in \mathcal{M} \text{ and } |\lambda - \lambda(\mu)| \leq \delta/2 \}.$$

This is compact as λ is continuous and compactness is preserved by images of such functions. We will use this to show

$$\sup_{(\lambda,\mu)\in A} |\Psi_n(\lambda, \mu) - \Psi(\lambda, \mu)| \stackrel{a.s.}{\to} 0. \tag{6.1}$$

Since A is compact and

$$(\lambda, \mu) \mapsto \frac{m(y, \mu)}{1 + \lambda m(y, \mu)}$$

is continuous with probability 1, the uniform law of large numbers (Ferguson 1996, p. 108) ensures (6.1), provided

$$\left| \frac{m(y, \mu)}{1 + \lambda m(y, \mu)} \right| \leq p(y)$$

for all $(\lambda, \mu) \in A$ and some p with finite expectation. For every $(\lambda, \mu) \in A$,

$$\left| \frac{m(y, \mu)}{1 + \lambda m(y, \mu)} \right| \leq \frac{m(y, \mu)}{L}$$

with probability 1. By assumption, $|m(y, \mu)|$ is bounded by $p_1(y)$ which is integrable with respect to the probability measure corresponding to F. Hence,

$$\Psi_n(\lambda, \mu) \stackrel{a.s}{\to} \Psi(\lambda, \mu)$$

uniformly in A. Now fix $\epsilon \in (0, \delta/2)$. Then

$$\Pr\left(\sup_{\mu\in\mathcal{M}} |\lambda_n(\mu) - \lambda(\mu)| < \epsilon \right) \geq$$
$$\Pr\left(\sup_{\mu\in\mathcal{M}} \Psi_n[\lambda(\mu) + \epsilon, \mu] < 0 < \inf_{\mu\in\mathcal{M}} \Psi_n[\lambda(\mu) - \epsilon, \mu] \right)$$

as

$$\Psi_n[\lambda(\mu) + \epsilon, \mu] < 0 \quad \text{and} \quad \Psi_n[\lambda(\mu) - \epsilon, \mu] > 0$$

is sufficient for $\lambda \mapsto \Psi_n(\lambda, \mu)$ to have a root between $\lambda(\mu) - \epsilon$ and $\lambda(\mu) + \epsilon$. This follows from continuity of Ψ_n and the intermediate value theorem. Arguing as in the beginning of the proof, we notice that $\lambda \mapsto \Phi_n(\lambda, \mu)$ is a strictly decreasing function for each μ. In particular, it is one-to-one. Because of this, a zero in the interval $[\lambda(\mu) - \epsilon,\, \lambda(\mu) + \epsilon]$ must, indeed, be $\lambda_n(\mu)$.

The uniform convergence proved previously is sufficient for

$$\sup_{\mu \in \mathcal{M}} \Psi_n[\lambda(\mu) + \epsilon, \mu] \xrightarrow{\Pr} \sup_{\mu \in \mathcal{M}} \Psi[\lambda(\mu) + \epsilon, \mu]$$

and

$$\inf_{\mu \in \mathcal{M}} \Psi_n[\lambda(\mu) - \epsilon, \mu] \xrightarrow{\Pr} \inf_{\mu \in \mathcal{M}} \Psi[\lambda(\mu) - \epsilon, \mu].$$

Hence

$$\lim_{n \to \infty} \Pr\left(\sup_{\mu \in \mathcal{M}} |\lambda_n(\mu) - \lambda(\mu)| < \epsilon \right) \geq$$
$$\Pr\left(\sup_{\mu \in \mathcal{M}} \Psi[\lambda(\mu) + \epsilon, \mu] < 0 < \inf_{\mu \in \mathcal{M}} \Psi[\lambda(\mu) - \epsilon, \mu] \right).$$

As was shown before, Ψ is decreasing and continuous in λ for every fixed μ. Furthermore the function is equal to 0 in $\lambda(\mu)$. Since $\mathcal{M}$ is compact, this ensures that the right-hand side of the above inequality is equal to 1. Hence,

$$\sup_{\mu \in \mathcal{M}} |\lambda_n(\mu) - \lambda(\mu)| \xrightarrow{\Pr} 0.$$

■

The above result cannot be applied when $m(y, \mu) = I(y \leq \mu) - q$ as this function is not continuous. Direct computation, however, shows that in this case,

$$\lambda_n(\mu) = \frac{\mathbb{F}_n(\mu) - q}{q(1 - q)}. \tag{6.2}$$

Here $\mathbb{F}_n$ denotes the empirical distribution function of the data. This converges uniformly in probability to

$$\lambda(\mu) = \frac{F(\mu) - q}{q(1 - q)} \tag{6.3}$$

by the Glivenko-Cantelli theorem, see e.g. Vaart 1998, p. 266. Hence,

$$\sup_{\mu \in \mathcal{M}} |\lambda_n(\mu) - \lambda(\mu)| \xrightarrow{\Pr} 0.$$

in this case as well.

Now that we have shown $\lambda_n(\mu) \xrightarrow{\Pr} \lambda(\mu)$ uniformly over compact sets, we can prove that the convergence is of order $O_{\Pr}(1/\sqrt{n})$.

Lemma 6.1.2. *Let $\mathcal{M}$ be compact and assume the conditions of Lemma 6.1.1 hold with this set. Furthermore, let λ be continuously differentiable as a function of μ. If there exists p_2 such that*

$$|m(y, \mu_1) - m(y, \mu_2)| \leq p_2(y)|\mu_1 - \mu_2| \tag{6.4}$$

for all $\mu_1, \mu_2 \in \mathcal{M}$ and

$$\mathrm{E}\, p_1(Y)^4, \mathrm{E}\, p_1(Y)^2 p_2(Y), \mathrm{E}\, p_2(Y)^2 < \infty,$$

we have

$$\sqrt{n}[\lambda_n(\mu) - \lambda(\mu)] \xrightarrow{d} \frac{V(\mu)}{S(\mu)} \tag{6.5}$$

as a process in $\ell^\infty(\mathcal{M})$. Here

$$S(\mu) = \mathrm{E}\left(\frac{M(\mu)^2}{[1 + \lambda(\mu)M(\mu)]^2} \right)$$

is assumed to exists and be finite and non-zero for each $\mu \in \mathcal{M}$, and V is a Gaussian process with mean 0 and variance $S(\mu)$. In particular, (6.5) *implies that*

$$\sup_{\mu \in \mathcal{M}} \left| \sqrt{n}(\lambda_n(\mu) - \lambda(\mu)) \right| = O_{\Pr}(1).$$

Remark 6.1.3. The function λ is continuously differentiable at μ_1 if S is non singular and the map

$$\mu \mapsto \mathrm{E}\left(\frac{M(\mu)}{1 + \lambda(\mu_1)M(\mu)} \right)$$

is continuously differentiable at μ_1. This follows from the implicit function theorem, see e.g. Lindstrøm 2017, p. 212 or Section 5.4 where the behavior of λ as a function of μ was discussed in detail.

Proof. This proof is a slight modification of Theorem 5.41 in Vaart 1998, p. 68.

By assumption there exists δ and L such that for all (λ, μ) with $|\lambda - \lambda(\mu)| < \delta$, $1 + \lambda M(\mu) > L$ with probability 1. Hence,

$$\sup_{\mu \in \mathcal{M}} |\lambda_n(\mu) - \lambda(\mu)| < \delta \implies 1 + \lambda_n(\mu)M(\mu) > L \text{ for all } \mu \in \mathcal{M}.$$

The first event has limiting probability 1 by Lemma 6.1.1. So,

$$1 + \lambda_n(\mu)M(\mu) > L \text{ for all } \mu \in \mathcal{M}$$

happens with probability tending to 1. We can therefore assume $1 + \lambda_n(\mu)M(\mu) > L$ for all $\mu \in \mathcal{M}$ and show that the result holds under this assumption.

Using the first order Taylor expansion of $\lambda \mapsto \Psi_n(\lambda, \mu)$ around $\lambda(\mu)$, we see

$$0 = \Psi_n[\lambda_n(\mu), \mu]$$

$$
\begin{aligned}
=&\Psi_n[\lambda(\mu),\mu] + [\lambda_n(\mu) - \lambda(\mu)]\frac{\partial \Psi_n}{\partial \lambda}[\lambda(\mu),\mu] + \\
&+ \frac{1}{2}[\lambda_n(\mu) - \lambda(\mu)]^2 \frac{\partial^2 \Psi_n}{\partial \lambda^2}[\lambda(\mu)^*,\mu]
\end{aligned}
$$

for some $\lambda(\mu)^*$ on the line segment between $\lambda(\mu)$ and $\lambda_n(\mu)$. By the triangle inequality,

$$
\left|\frac{\partial^2 \Psi_n}{\partial \lambda^2}[\lambda(\mu)^*,\mu]\right| = \left|\frac{1}{n}\sum_{i=1}\frac{2M_i(\mu)^3}{[1+\lambda(\mu)^* M_i(\mu)]^3}\right| \leq \frac{2}{L^3}\cdot\frac{1}{n}\sum_{i=1}^{n}|M_i(\mu)|^3
$$

for all $\mu \in \mathcal{M}$. Here we have used that all λs on the line segment between $\lambda_n(\mu)$ and $\lambda(\mu)$ satisfy the property

$$
|\lambda - \lambda(\mu)| < \delta
$$

and hence also

$$
1 + \lambda M(\mu) > L
$$

with probability 1. As m is continuous, $\mathcal{M}$ compact and $\mathrm{E}p_1(Y)^3 < \infty$, with p_1 as in Lemma 6.1.1, the uniform law of large numbers ensures that

$$
\frac{1}{n}\sum_{i=1}^{n}|M_i(\mu)|^3
$$

converges in probability uniformly in $\mathcal{M}$ to its expected value. In addition,

$$
|m(y,\mu)|^3 \leq p_1(y)^3
$$

for all $\mu \in \mathcal{M}$, so

$$
\mathrm{E}\,|M(\mu)|^3 \leq \mathrm{E}\,p_1(Y)^3,
$$

for all $\mu \in \mathcal{M}$. The right hand side of this inequality is finite by assumption, ensuring

$$
\sup_{\mu\in\mathcal{M}} \mathrm{E}\,|M(\mu)|^3 < \infty.
$$

This proves

$$
\sup_{\mu\in\mathcal{M}}\left|\frac{\partial^2 \Psi_n}{\partial \lambda^2}[\lambda^*(\mu),\mu]\right| = O_{\mathrm{Pr}}(1). \tag{6.6}
$$

Furthermore, $\lambda_n(\mu)$ converges uniformly in probability to $\lambda(\mu)$. Hence, (6.6) implies that the last term in the Taylor expansion is $[\lambda_n(\mu) - \lambda(\mu)]\epsilon_n$ with ϵ_n tending to 0 in probability.

Notice further that

$$
\frac{\partial \Psi_n}{\partial \lambda}[\lambda(\mu),\mu] = -\frac{1}{n}\sum_{i=1}^{n}\frac{M_i(\mu)^2}{[1+\lambda(\mu)M_i(\mu)]^2}.
$$

As $\mu \mapsto m(y,\mu)$ is continuous for almost all y, $\mathcal{M}$ compact and

$$\left| \frac{m(y,\mu)^2}{1+\lambda(\mu)m(y,\mu)} \right| \leq |m(y,\mu)|^2 \leq p_1(y)^2,$$

this converges uniformly in probability to its expected value, $-S(\mu)$, by the uniform law of large numbers.

After manipulating the Taylor expansion as in the proof of 5.41 in Vaart 1998, p. 68 and applying our results, the following holds uniformly in $\mathcal{M}$:

$$\sqrt{n}[\lambda_n(\mu) - \lambda(\mu)] = \frac{\sqrt{n}\Psi_n[\lambda(\mu),\mu]}{S(\mu)} + o_{\Pr}(1). \tag{6.7}$$

Here we have used

$$\sup_{\mu\in\mathcal{M}} |S(\mu)|^{-1} < \infty.$$

This follows from the extreme value theorem and assumption of $S(\mu) \neq 0$ for all $\mu \in \mathcal{M}$, provided S is continuous as a function of μ. It is, and this can be proved similarly to how continuity of $\lambda \mapsto \Phi(\lambda,\mu)$ was shown in the proof of Lemma 6.1.1. We will leave out the details.

Consider

$$\sqrt{n}\Psi_n[\lambda(\mu),\mu] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{M_i(\mu)}{1+\lambda(\mu)M_i(\mu)}.$$

Some algebraic efforts reveal

$$\begin{aligned}
&\left| \frac{m(y,\mu_1)}{1+\lambda(\mu_1)m(y,\mu_1)} - \frac{m(y,\mu_2)}{1+\lambda(\mu_2)^T m(y,\mu_2)} \right| = \\
&\left| \frac{m(y,\mu_1) - m(y,\mu_2) + m(y,\mu_1)m(y,\mu_2)(\lambda(\mu_2)-\lambda(\mu_1))}{[1+\lambda(\mu_1)m(y,\mu_1)][1+\lambda(\mu_2)m(y,\mu_2)]} \right| \leq \\
&\frac{1}{L^2}\{|m(y,\mu_1) - m(y,\mu_2)| + |m(y,\mu_1)m(y,\mu_2)| \cdot |\lambda(\mu_2) - \lambda(\mu_1)|\}.
\end{aligned}$$

By assumption there exists functions of y, p_1 and p_2, such that

$$|m(y,\mu_1) - m(y,\mu_2)| \leq p_2(y)\|\mu_1 - \mu_2\| \quad \text{and} \quad |m(y,\mu_1)m(y,\mu_2)| \leq p_1(y)^2.$$

Furthermore, λ is continuously differentiable and $\mathcal{M}$ compact, so by the mean value theorem there exists $K < \infty$ such that

$$|\lambda(\mu_2) - \lambda(\mu_1)| \leq K\|\mu_1 - \mu_2\|.$$

Hence

$$\left| \frac{m(y,\mu_1)}{1+\lambda(\mu_1)^T m(y,\mu_1)} - \frac{m(y,\mu_2)}{1+\lambda(\mu_2)^T m(y,\mu_2)} \right| \leq \frac{1}{L}\big(p_2(y) + Kp_1(y)^2\big)\|\mu_1 - \mu_2\|.$$

By example 19.7 in Vaart 1998, p. 270, this, in combination with compactness of $\mathcal{M}$, ensures that the class

$$\left\{ \frac{m(y,\mu)}{1+\lambda(\mu)m(y,\mu)} \;\middle|\; \mu \in \mathcal{M} \right\}$$

is P-Donsker, provided.

$$\mathrm{E}\left(\frac{p_2(Y) + Kp_1(Y)^2}{L}\right) = \frac{\mathrm{E}\, p_2(Y)^2 + K\,\mathrm{E}\, p_2(Y)p_1(Y)^2 + K^2\mathrm{E}\, p_1(Y)^4}{L^2} < \infty.$$

By assumption each term in the above expression is finite. Because of this,

$$\sqrt{n}\Psi_n(\mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{M_i(\mu)}{1 + \lambda(\mu)M_i(\mu)}$$

converges to a Gaussian process, $V(\mu)$, in $\ell^\infty(\mathcal{M})$ with mean 0 and variance $S(\mu)$. Combining this with (6.7), proves that

$$\sqrt{n}[\lambda_n(\mu) - \lambda(\mu)] \xrightarrow{d} \frac{V(\mu)}{S(\mu)}$$

as a process in $\ell^\infty(\mathcal{M})$. In particular,

$$\sup_{\mu\in\mathcal{M}} \left|\sqrt{n}[\lambda_n(\mu) - \lambda(\mu)]\right| = \sup_{\mu\in\mathcal{M}} \left|\frac{V(\mu)}{S(\mu)}\right| = O_{\mathrm{Pr}}(1).$$

This follows from the fact that

$$\sup_{\mu\in\mathcal{M}} |S(\mu)|^{-1} < \infty,$$

as shown previously, and

$$\sup_{\mu\in\mathcal{M}} |V_n(\mu)| \xrightarrow{d} \sup_{\mu\in\mathcal{M}} |V(\mu)| = O_{\mathrm{Pr}}(1),$$

by the continuous mapping theorem. ■

As in Lemma 6.1.1, we need m to be continuous to apply Lemma 6.1.2. This excludes the case of $m(Y, \mu) = I(Y, \mu) - q$, but, once again, the conclusion of Lemma 6.1.2 holds for this estimating function as well. This follows from the expressions given in (6.2) and (6.3) derived previously in combination with the Donsker theorem (see Vaart 1998, p. 266).

We have now shown the uniform counterpart to Theorem 5.2.1. What remains to prove is that Theorem 5.3.1 holds uniformly in $\mathcal{M}$ as well. This amounts to showing that the remainder term in (5.20) tends uniformly to 0 in probability.

Lemma 6.1.4. *Fix a compact set $\mathcal{M}$ and assume the conditions of Lemma 6.1.1 and Lemma 6.1.2 hold with this set. Then*

$$\sup_{\mu\in\mathcal{M}} |\delta_n(\mu)| = o_{\mathrm{Pr}}(1),$$

where $\delta_n(\mu)$ is the remainder term in (5.20)*.*

Proof. Inspecting the derivations of Theorem 5.3.1, we see that three thing need to be shown. Firstly, we must prove that $\epsilon_n(\mu) = o_{\mathrm{Pr}}(1)$ uniformly in $\mathcal{M}$,

with $\epsilon_n(\mu)$ being the remainder term in the Taylor expansion given in (5.16). Secondly,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \frac{M_i(\mu)^2}{[1 + \lambda_n(\mu) M_i(\mu)]^2} - \frac{1}{n} \sum_{i=1}^{n} \frac{M_i(\mu)^2}{[1 + \lambda(\mu) M_i(\mu)]^2} \right| = o_{\Pr}(1)$$

uniformly in $\mathcal{M}$ needs to be shown. Lastly, we have to establish that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{M_i(\mu)^2}{[1 + \lambda(\mu) M_i(\mu)]^2} = S(\mu) + o_{\Pr}(1)$$

uniformly in μ. The last assessment was shown in the proof of Lemma 6.1.2 and is therefore omitted.

As in the proof of Lemma 6.1.2, we can assume $\lambda_n(\mu)$ all satisfy

$$\mathbf{1} + \lambda_n(\mu) M(\mu) > L$$

as the probability of this event tends to 1. Hence,

$$|\epsilon_n(\mu)| \leq \frac{1}{L^3} |\lambda_n(\mu) - \lambda(\mu)|^3 \sum_{i=1}^{n} |M_i(\mu)|^3$$

uniformly in μ. This can be shown similarly as in the proof of Theorem 5.2.1. By Lemma 6.1.2

$$\sup_{\mu \in \mathcal{M}} |\lambda_n(\mu) - \lambda| = O_{\Pr}(1/\sqrt{n}).$$

Furthermore,

$$\sup_{\mu \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} |M_i(\mu)|^3 = O_{\Pr}(1) \tag{6.8}$$

was shown in the proof of the same result. Hence,

$$\sup_{\mu \in \mathcal{M}} |\epsilon_n(\mu)| = O_{\Pr}(1/\sqrt{n}).$$

Similarly, we can show

$$\left| \frac{1}{n} \sum_{i=1}^{n} \frac{M_i(\mu)^2}{[1 + \lambda_n(\mu) M_i(\mu)]^2} - \frac{1}{n} \sum_{i=1}^{n} \frac{M_i(\mu)^2}{[1 + \lambda(\mu) M_i(\mu)]^2} \right| = O_{\Pr}(1/\sqrt{n})$$

uniformly in μ by arguing as in the proof of Theorem 5.3.1 and using Lemma 6.1.2 in combinations with (6.8). This is very similar to what was done above, and the details will therefore be left out. ■

Once again the continuity assumption in Lemma 6.1.4 excludes the case of $m(y, \mu) = I(y \leq \mu) - q$. However, the conclusion can be shown to hold in this case as well by using (6.2) and (6.3) in combination with the Glivenko-Cantelli theorem.

6.2 Consistency

We are now ready to show that the maximum empirical likelihood estimator goes in probability to the true parameter as the sample size increases. This estimator will be denoted by $\widehat{\mu}_n$.

Before we show consistency of $\widehat{\mu}_n$ towards μ_0, we will show that the true parameter is, indeed, the maximizer of

$$\Gamma(\mu) = -\mathrm{E}\,\log[1 + \lambda(\mu)M(\mu)].$$

First notice that

$$0 = \mathrm{E}\bigg(\frac{M(\mu)}{1 + \lambda(\mu)M(\mu)}\bigg),$$

implies that

$$0 = \mathrm{E}\bigg(\frac{1 + \lambda(\mu)M(\mu) - 1}{1 + \lambda(\mu)M(\mu)}\bigg),$$

which is equivalent to

$$0 = 1 - \mathrm{E}\bigg(\frac{1}{1 + \lambda(\mu)M(\mu)}\bigg) \quad \text{or} \quad 1 = \mathrm{E}\bigg(\frac{1}{1 + \lambda(\mu)M(\mu)}\bigg).$$

Further manipulation of the expression shows

$$\mathrm{E}\bigg(\frac{1}{1 + \lambda(\mu)M(\mu)}\bigg) = \mathrm{E}\,\exp\{-\log[1 + \lambda(\mu)M(\mu)]\}.$$

Since $x \mapsto \exp(x)$ is a strictly convex function, Jensen's inequality guarantees

$$\begin{aligned} 1 &= \mathrm{E}\,\exp\{-\log[1 + \lambda(\mu)M(\mu)]\} \\ &> \exp(\mathrm{E}\{-\log[1 + \lambda(\mu)M(\mu)]\}) \\ &= \exp[\Gamma(\mu)], \end{aligned}$$

when $\lambda(\mu) \neq 0$. This is equivalent to

$$\Gamma(\mu) < 0,$$

for all μ such that $\lambda(\mu) \neq 0$. Furthermore, μ_0 is the unique solution to

$$0 = \mathrm{E}\,M(\mu),$$

and $\lambda(\mu) = 0$ if and only if

$$0 = \mathrm{E}\bigg(\frac{M(\mu)}{1 + 0 \cdot M(\mu)}\bigg) = \mathrm{E}\,M(\mu).$$

Hence, $\lambda(\mu) = 0$ if and only if $\mu = \mu_0$. Lastly,

$$\Gamma(\mu_0) = \mathrm{E}\{-\log[1 + 0 \cdot M(\mu)]\} = 0,$$

so μ_0 is the unique maximizer of Γ.

Theorem 6.2.1. *For each $n \in \mathbb{N}$ let $\widehat{\mu}_n$ denote a maximizer of EL_n in a compact set $\mathcal{M}$ such that the conditions of Lemma 6.1.1 and Lemma 6.1.2 hold for this set. Then*

$$\widehat{\mu}_n \xrightarrow{\Pr} \mu_0,$$

where μ_0 is the unique solution to

$$\mathrm{E}\, M(\mu) = 0.$$

Proof. We will show that the conditions of Theorem 5.7 in Vaart 1998, p. 45 hold.

First notice that,

$$\frac{1}{n} \log \mathrm{EL}_n(\widehat{\mu}_n) \geq \frac{1}{n} \log \mathrm{EL}_n(\mu_0)$$

by definition of $\widehat{\mu}$. This proves the last assessment in the theorem.

By the proof of Lemma 6.1.2, $V_n(\mu)$ converges to a Gaussian limit process in $\ell^\infty(\mathcal{M})$. In particular, this implies that

$$\frac{1}{\sqrt{n}} V_n(\mu) \xrightarrow{\Pr} 0$$

uniformly in μ. Furthermore,

$$\sup_{\mu \in \mathcal{M}} |S(\mu)|^{-1} < \infty$$

was shown in the proof of Lemma 6.1.2. Hence,

$$\sup_{\mu \in \mathcal{M}} \frac{1}{n} \left| V_n(\mu) S(\mu)^{-1} V_n(\mu) \right| = o_{\Pr}(1).$$

Furthermore, $\delta_n(\mu) = o_{\Pr}(1)$ uniformly in $\mathcal{M}$ by Lemma 6.1.4. So,

$$\sup_{\mu \in \mathcal{M}} \left| \frac{1}{n} \log \mathrm{EL}_n(\mu) - \Gamma(\mu) \right| = \sup_{\mu \in \mathcal{M}} |\Gamma_n(\mu) - \Gamma(\mu)| + o_{\Pr}(1)$$

where

$$\Gamma_n(\mu) = -\frac{1}{n} \sum_{i=1}^{n} \log[1 + \lambda(\mu) M(\mu)].$$

Because of this, it suffices to show that the uniform law of large numbers can be applied to Γ_n, for the first condition in the theorem to be satisfied.

By the mean value theorem

$$\begin{aligned} |\log[1 + \lambda(\mu) m(y, \mu)]| &= |\log[1 + \lambda(\mu) m(y, \mu)] - \log(1)| \\ &= |\lambda(\mu) m(y, \mu)| \cdot \left| \frac{1}{1 + t(y, \mu) \lambda(\mu) m(y, \mu)} \right| \end{aligned}$$

for some $t(y, \mu) \in [0, 1]$. If $\lambda(\mu) m(y, \mu) \leq 0$,

$$1 + t(y, \mu) \lambda(\mu) m(y, \mu) \geq 1 + \lambda(\mu) m(y, \mu) > L,$$

where the existence of L is guaranteed by the assumptions of Lemma 6.1.1. If, on the other hand, $\lambda(\mu)m(y,\mu) \geq 0$,

$$1 + t(y,\mu)\lambda(\mu)m(y,\mu) \geq 1 + 0 \cdot m(y,\mu) = 1.$$

In either case

$$1 + t(y,\mu)\lambda(\mu)m(y,\mu) \geq K = \min(1, L) > 0.$$

Because of this,

$$|\log[1 + \lambda(\mu)m(y,\mu)]| \leq \frac{|\lambda(\mu)m(y,\mu)|}{K} \leq \frac{|\lambda(\mu)| \cdot |m(y,\mu)|}{K}.$$

The function λ is continuous and $\mathcal{M}$ is compact. So λ attains its maximum, C, on $\mathcal{M}$ by the extreme value theorem. Furthermore, $|m(y,\mu)| \leq p_1(y)$ where $\mathrm{E}\, p_1(Y) < \infty$ by the assumptions of Lemma 6.1.1. Combining this with the above shows:

$$|\log[1 + \lambda(\mu)m(y,\mu)]| \leq \frac{Cp_1(y)}{K}.$$

This is an integrable function with respect to the probability measure on Y that do not depend on μ. Since $\mathcal{M}$ is compact and $\mu \mapsto \lambda(\mu)m(y,\mu)$ continuous for almost all y by continuity of both λ and $\mu \mapsto m(y,\mu)$, the uniform law of large numbers can be applied to get

$$\sup_{\mu\in\mathcal{M}} \left| -\frac{1}{n}\sum_{i=1}^{n} \log[1 + \lambda(\mu)M(\mu)] - \Gamma(\mu) \right| \xrightarrow{\Pr} 0.$$

By the arguments above, $\log[1 + \lambda(\mu)m(y,\mu)]$ is bounded by the integrable function $Cp_1(y)/K$. Since λ and $\mu \mapsto m(y,\mu)$, for almost all fixed y, is continuous, Lebesgue's dominated convergence theorem can be applied to show continuity of Γ. This is similar to what was done in the proof of Lemma 6.1.1 and details will be omitted. Furthermore, μ_0 is the unique maximizer of Γ, and $\mathcal{M}$ is compact. Hence,

$$\sup_{\|\mu-\mu_0\|\geq\epsilon} \Gamma(\mu) < \Gamma(\mu_0)$$

for all $\epsilon > 0$ by the extreme value theorem. This proves the remaining condition in theorem 5.7 in Vaart 1998, p. 45 and concludes the proof of Theorem 6.2.1. ■

Theorem 6.2.1 does not cover the case of $m(y,\mu) = I(y \leq \mu) - q$. As we believe this to be an interesting and important estimating function, we will take the time to prove consistency in this particular case.

Theorem 6.2.2. *Let the parameter space be compact and $\widehat{\mu}_n$ a maximizer of $\log \mathrm{EL}_n(\mu)$, constructed with the estimating function*

$$m(y,\mu) = I(y \leq \mu) - q,$$

for some $q \in (0,1)$. Furthermore, let F be continuous. Then $\widehat{\mu}_n$ is a consistent estimator of μ_0.

Proof. As remarked after Lemma 6.1.2 and Lemma 6.1.4, the remainder term, $\delta_n(\mu)$, tends uniformly to 0, and

$$V_n(\mu) \xrightarrow{d} V(\mu)$$

as a Gaussian process in $\mathcal{M}$. Furthermore,

$$S(\mu) = F(\mu)\left(\frac{1-q}{1+\lambda(\mu)(1-q)}\right)^2 + [1-F(\mu)]\left(\frac{-q}{1-\lambda(\mu)q}\right)^2.$$

F is continuous by assumption, ensuring continuity of both $\lambda(\mu)$ given in (6.3) and S. Furthermore $S(\mu) > 0$ for every μ. By compactness of $\mathcal{M}$ and the extreme value theorem, $|S(\mu)|^{-1}$ is uniformly bounded in $\mathcal{M}$, ensuring

$$\sup_{\mu\in\mathcal{M}} \left|V_n(\mu)S(\mu)^{-1}V_n(\mu)\right| = O_{\Pr}(1).$$

Similarly, direct computation shows

$$\Gamma(\mu) = -F(\mu)\log[1+\lambda(\mu)(1-q)] - [1-F(\mu)]\log[1-\lambda(\mu)q]$$

which is continuous, and hence satisfies

$$\sup_{\|\mu-\mu_0\|\geq\epsilon} \Gamma(\mu) < \Gamma(\mu_0)$$

for all $\epsilon > 0$ by arguments similar to those given in the proof of Theorem 6.2.1.

What remains is to show that

$$\sup_{\mu\in\mathcal{M}} |\Gamma_n(\mu) - \Gamma(\mu)| \xrightarrow{\Pr} 0, \tag{6.9}$$

with Γ defined as

$$\Gamma(\mu) = -\mathrm{E}\log\{1+\lambda(\mu)[I(Y\leq\mu)-q]\}$$

and Γ_n as

$$\Gamma_n(\mu) = -\frac{1}{n}\sum_{i=1}^{n}\log\{1+\lambda(\mu)[I(Y_i\leq\mu)-q]\}.$$

First notice that

$$\Gamma(\mu) = -\log[1+\lambda(\mu)(1-q)]\Pr(Y\leq\mu) - \log[1-\lambda(\mu)q]\Pr(Y>\mu)$$

and

$$\Gamma_n(\mu) =$$
$$-\log[1+\lambda(\mu)(1-q)]\frac{1}{n}\sum_{i=1}^{n}I(Y_i\leq\mu) - \log[1-\lambda(\mu)q]\frac{1}{n}\sum_{i=1}^{n}I(Y_i>\mu).$$

The functions

$$\mu \mapsto |-\log[1+\lambda(\mu)(1-q)]| \quad \text{and} \quad \mu \mapsto |-\log[1-\lambda(\mu)q]|$$

are continuous. Because of this, they attain their maximums, K_1 and K_2 respectively, on $\mathcal{M}$. Combining this with the above expressions and applying the triangle inequality shows

$$|\Gamma_n(\mu) - \Gamma(\mu)| \leq$$
$$K_1 \left| \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq \mu) - \Pr(Y \leq \mu) \right| + K_2 \left| \frac{1}{n} \sum_{i=1}^{n} I(Y_i > \mu) - \Pr(Y > \mu) \right|$$

for all $\mu \in \mathcal{M}$. Both of these terms goes in probability to zero uniformly in $\mathcal{M}$ by the Glivenko-Cantelli theorem. This shows (6.9) and concludes the proof. ■

The sample q-quantile always maximizes the empirical likelihood function. Consistency of this estimator is not exactly a revolutionary discovery, but we include the proof for the sake of completion and as an example of a situation where consistency holds even when the conditions of Theorem 6.2.1 do not. Furthermore, we established the following in the proof of Theorem 6.2.2

$$\sup_{\mu \in \mathcal{M}} |\log \mathrm{EL}_n(\mu) - \Gamma(\mu)| \xrightarrow{\Pr} 0.$$

This will be used in Chapter 8 to show consistency of the maximizer of a hybrid combination of parametric and empirical likelihood.

6.3 Asymptotic normality

We are now ready to prove asymptotic normality of the maximum empirical likelihood estimator. To show this, we will first show $\sqrt{n}$-consistency of the estimator towards the true parameter. Afterwards, we will modify the proof of theorem 5.23 Vaart 1998, p. 53 to arrive at a normal limit for $\sqrt{n}(\widehat{\mu} - \mu_0)$. The theorems in this subsection assumes continuity of the estimating function at the true parameter. This excludes the case with $m(y, \mu) = I(y \leq \mu) - q$.

Lemma 6.3.1. *Let $\widehat{\mu}_n$, μ_0 and $\mathcal{M}$ be as in Theorem 6.2.1 and assume the conditions of Lemma 6.1.1 and Lemma 6.1.2 hold true. Then*

$$\sqrt{n}(\widehat{\mu}_n - \mu_0) = O_{\Pr}(1).$$

provided Γ admits a second order Taylor expansion at μ_0 with nonsingular Hessian matrix, $H\Gamma(\mu_0)$, μ_0 is in the interior of $\mathcal{M}$ and

$$\mathrm{E}\, p_1(Y) p_2(Y) < \infty$$

where p_1 and p_2 are as in Lemma 6.1.1 and Lemma 6.1.2 respectively.

Proof. We will use corollary 5.53 in Vaart 1998, p. 77 to show this result.

By the mean value theorem,

$$|\log[1 + \lambda(\mu_1) m(y, \mu_1)] - \log[1 + \lambda(\mu_2) m(y, \mu_2)]| =$$
$$\frac{1}{1 + \xi} \cdot |\lambda(\mu_1) m(y, \mu_1) - \lambda(\mu_2) m(y, \mu_2)|,$$

for some ξ on the line segment between $\lambda(\mu_1)m(y,\mu_1)$ and $\lambda(\mu_2)m(y,\mu_2)$. We can write ξ as

$$\xi = t\lambda(\mu_1)m(y,\mu_1) + (1-t)\lambda(\mu_2)m(y,\mu_2)$$

for some $t \in [0,1]$. Because of this,

$$\begin{aligned} 1+\xi &= t[1+\lambda(\mu_1)m(y,\mu_1)] + (1-t)[1+\lambda(\mu_2)m(y,\mu_2)] \\ &\geq tL + (1-t)L \\ &= L. \end{aligned}$$

Hence,

$$\begin{aligned} &|\log[1+\lambda(\mu_1)m(y,\mu_1)] - \log[1+\lambda(\mu_2)m(y,\mu_2)]| \leq \\ &\frac{1}{L}\cdot|\lambda(\mu_1)m(y,\mu_1) - \lambda(\mu_2)m(y,\mu_2)|. \end{aligned}$$

Addition and subtraction of $\lambda(\mu_1)m(y,\mu_2)$ reveals

$$\begin{aligned} &|\lambda(\mu_1)m(y,\mu_1) - \lambda(\mu_2)m(y,\mu_2)| \leq \\ &|\lambda(\mu_1)|\cdot|m(y,\mu_1) - m(y,\mu_2)| + |\lambda(\mu_1) - \lambda(\mu_2)|\cdot|m(y,\mu_2)|. \end{aligned}$$

By the conditions of Lemma 6.1.1 and Lemma 6.1.2 there exists integrable p_1 and p_2 such that

$$|m(y,\mu_1) - m(y,\mu_2)| \leq p_2(y)\|\mu_1 - \mu_2\| \quad \text{and} \quad |m(y,\mu)| \leq p_1(y).$$

Furthermore, λ is continuous and $\mathcal{M}$ compact. So by the extreme value theorem there exists $K_2 < \infty$ such that $|\lambda(\mu)| \leq K_2$ for all $\mu \in \mathcal{M}$. In addition, the function is continuously differentiable. So by a combination of the mean value theorem for vector valued functions and the extreme value theorem

$$|\lambda(\mu_1) - \lambda(\mu_2)| \leq K_1\|\mu_1 - \mu_2\|$$

for some $K_1 < \infty$. Hence,

$$\begin{aligned} &|\log[1+\lambda(\mu_1)m(y,\mu_1)] - \log[1+\lambda(\mu_2)m(y,\mu_2)]| \leq \\ &\frac{1}{L}[K_2p_2(y)\|\mu_1 - \mu_2\| + K_1\|\mu_1 - \mu_2\|p_1(y)] \leq \\ &\frac{1}{L}[K_2p_2(y) + K_1p_1(y)]\|\mu_1 - \mu_2\|. \end{aligned}$$

Furthermore,

$$\begin{aligned} &\mathrm{E}\left(\frac{[K_1p_2(Y) + K_2p_2(Y)]^2}{L^2}\right) = \\ &\frac{K_1^2}{L^2}\mathrm{E}\,p_1(Y)^2 + \frac{K_2^2}{L^2}\mathrm{E}\,p_2(Y)^2 + \frac{K_1K_2}{L^2}\mathrm{E}p_1(Y)p_2(Y). \end{aligned}$$

By the assumptions of Lemma 6.1.2 and Lemma 6.3.1 each summand in the above expression is finite. This shows that the first assertion of corollary 5.53 in Vaart 1998, p. 77 holds true.

By assumption $\Gamma(\mu)$ admits a second-order Taylor expansion at μ_0, and as

$$n\Gamma_n(\mu) = \log \mathrm{EL}_n(\mu) + r_n(\mu)$$

with

$$r_n(\mu) = V_n(\mu)S(\mu)^{-1}V_n(\mu) + \delta_n(\mu),$$

we have

$$n\Gamma_n(\widehat{\mu}_n) \geq n\Gamma_n(\mu_0) + r_n(\mu) - r_n(\widehat{\mu}_n),$$

by definition of $\widehat{\mu}_n$. Since r_n is uniformly $O_{\mathrm{Pr}}(1)$ by Lemma 6.1.4 and the proof of Lemma 6.1.2, this shows

$$\Gamma_n(\widehat{\mu}_n) \geq \Gamma_n(\mu_0) + O_{\mathrm{Pr}}(n^{-1})$$

and concludes the proof. ■

Now that we know the rate of convergence, we can prove that the maximum empirical likelihood estimator is asymptotically normally distributed after proper scaling and centering.

Theorem 6.3.2. *Let $\widehat{\mu}_n$, μ_0 and $\mathcal{M}$ be as in Theorem 6.2.1 and assume the conditions of Lemma 6.1.1, Lemma 6.1.2 and Lemma 6.3.1 hold true. Let $\mu \mapsto m(y, \mu)$ be differentiable at μ_0 for almost all y with finite expected value. Then,*

$$\sqrt{n}(\widehat{\mu} - \mu_0) = -H\Gamma(\mu_0)^{-1}\lambda'(\mu_0)^T V_n(0) + o_{\mathrm{Pr}}(1).$$

In particular

$$\sqrt{n}(\widehat{\mu} - \mu_0) \xrightarrow{d} N(0, \Sigma)$$

with

$$\Sigma = H\Gamma(\mu_0)^{-1}\lambda'(\mu_0)^T S(0)\lambda'(\mu_0)H\Gamma(\mu_0)^{-1}.$$

Proof. This proof is a slight modification of the proof of theorem 5.23 in Vaart 1998, p. 53.

By the proof of Lemma 6.1.2,

$$\left| \frac{m(y, \mu_1)}{1 + \lambda(\mu_1)m(y, \mu_1)} - \frac{m(y, \mu_2)}{1 + \lambda(\mu_2)m(y, \mu_2)} \right| \leq \frac{1}{L}[p_2(y) + Kp_1(y)^2]\|\mu_1 - \mu_2\|,$$

and by the assumptions of Lemma 6.1.2,

$$\mathrm{E}\left(\frac{p_2(y) + Kp_1(y)^2}{L}\right)^2 = \frac{1}{L^2}\mathrm{E}\, p_2(Y)^2 + \frac{K^2}{L^2}\mathrm{E}\, p_1(Y)^4 + \frac{K}{L^2}p_1(Y)p_2(Y) < \infty.$$

In addition,

$$\mu \mapsto \frac{m(y, \mu)}{1 + \lambda(\mu)m(y, \mu)}$$

is differentiable at μ_0 for almost all y by differentiability of $\mu \mapsto m(y,\mu)$ and λ at this point. Its gradient at μ_0 is

$$\psi(y) = \frac{\partial m}{\partial \mu}(y,\mu_0) + \lambda'(\mu_0)^T m(y,\mu_0)^2.$$

Hence, by theorem 19.31 in Vaart 1998, p. 284,

$$\sum_{i=1}^{n}\left[\frac{M_i(\mu_0+\tilde{s}_n/\sqrt{n})}{1+\lambda(\mu_0+\tilde{s}_n/\sqrt{n})M_i(\mu_0+\tilde{s}_n/\sqrt{n})} - M_i(\mu_0)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\psi(Y_i) - \mathrm{E}\,\psi(Y)] + o_{\mathrm{Pr}}(1),$$

for every sequence $\tilde{s}_n$ bounded in probability. This can be reformulated as

$$\sqrt{n}V_n(\mu_0+\tilde{s}_n/\sqrt{n}) - \sqrt{n}V_n(0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\psi(Y_i) - \mathrm{E}\,\psi(Y)] + o_{\mathrm{Pr}}(1)$$

or

$$V_n(\mu_0+\tilde{s}_n/\sqrt{n}) - V_n(\mu_0) = \frac{1}{n}\sum_{i=1}^{n}[\psi(Y_i) - \mathrm{E}\,\psi(Y)] + o_{\mathrm{Pr}}(1).$$

By the law of large numbers the right hand side converges to 0 in probability, and hence

$$V_n(\mu_0+\tilde{s}_n/\sqrt{n}) = V_n(\mu_0) + o_{\mathrm{Pr}}(1)$$

for all sequences $\tilde{h}_n$ bounded in probability. Furthermore, continuity of S, which was shown in the proof of Lemma 6.3.1, implies that

$$S(\mu_0+\tilde{s}_n/\sqrt{n}) = S(\mu_0) + o_{\mathrm{Pr}}(1)$$

for all sequences $\tilde{s}_n$ bounded in probability, as

$$\mu_0 + \tilde{s}_n/\sqrt{n} \xrightarrow{\mathrm{Pr}} \mu_0.$$

In particular, this ensures that

$$V_n(\mu_0+\tilde{s}_n/\sqrt{n})S(\mu_0+\tilde{s}_n/\sqrt{n})^{-1}V_n(\mu_0+\tilde{s}_n/\sqrt{n}) = \tag{6.10}$$
$$V_n(\mu_0)S(\mu_0)^{-1}V_n(\mu_0) + o_{\mathrm{Pr}}(1). \tag{6.11}$$

The result now follows more or less directly from the proof of theorem 5.23 in Vaart 1998, p. 53. Arguing as is done there we arrive at the following expression:

$$n\Gamma_n(\mu_0+\tilde{s}_n/\sqrt{n}) = \frac{1}{2}\tilde{s}_n^T H\Gamma(\mu_0)\tilde{s}_n + \tilde{s}_n^T\lambda'(\mu_0)^T V_n(\mu_0) \tag{6.12}$$

for all sequences, $\tilde{s}_n$, bounded in probability. Here we have used that

$$\frac{\partial}{\partial\mu}\Big|_{\mu_0}\log[1+\lambda(\mu)m(y,\mu)] = \frac{\lambda'(\mu_0)m(y,\mu_0) + \partial m/\partial\mu(y,\mu_0)\lambda(\mu_0)}{1+\lambda(\mu_0)^T m(y,\mu_0)}$$

$$= \frac{\lambda'(\mu_0)m(y,\mu_0) + 0}{1+0}.$$

for almost all y.

The remainder term, $\delta_n(\mu)$, from (5.20) converges in probability to 0 uniformly in $\mathcal{M}$. Combining this with (6.11), shows

$$\log \mathrm{EL}_n(\mu_0 + \tilde{s}_n/\sqrt{n}) =$$
$$n\Gamma_n(\mu_0 + \tilde{s}_n/\sqrt{n}) + V_n(\mu_0)S(\mu_0)^{-1}V_n(\mu_0) + o_{\mathrm{Pr}}(1).$$

(6.12) then ensures

$$\log \mathrm{EL}_n(\mu_0 + \tilde{s}_n/\sqrt{n}) = \tag{6.13}$$
$$\frac{1}{2}\tilde{s}_n^T H\Gamma(\mu_0)\tilde{s}_n + \tilde{s}_n^T\lambda'(\mu_0)^T V_n(\mu_0) + V_n(\mu_0)S(\mu_0)^{-1}V_n(\mu_0) + o_{\mathrm{Pr}}(1) \tag{6.14}$$

for all sequences $\tilde{h}_n$ bounded in probability. In particular, this holds for the sequences

$$\widehat{s}_n = \sqrt{n}(\widehat{\mu} - \mu_0) \quad \text{and} \quad s_n^* = -H\Gamma(\mu_0)^{-1}\lambda'(\mu_0)^T V_n(\mu_0).$$

The vector $\widehat{\mu}_n$ maximizes $\log \mathrm{EL}_n(\mu)$, and hence

$$\frac{1}{2}\widehat{s}_n^T H\Gamma(\mu_0)\widehat{s}_n + \widehat{s}_n^T\lambda'(\mu_0)^T V_n(\mu_0) + V_n(\mu_0)S(\mu_0)^{-1}V_n(\mu_0) + o_{\mathrm{Pr}}(1) \geq$$
$$\frac{1}{2}(s_n^*)^T H\Gamma(\mu_0)s_n^* + (s_n^*)^T\lambda'(\mu_0)^T V_n(\mu_0) + V_n(\mu_0)S(\mu_0)^{-1}V_n(\mu_0) + o_{\mathrm{Pr}}(1).$$

by (6.14). Manipulating the above expression, shows

$$\frac{1}{2}(\widehat{s}_n - s_n^*)^T\Gamma(\mu_0)(\widehat{s}_n - s_n^*) + o_{\mathrm{Pr}}(1) \geq 0. \tag{6.15}$$

By the arguments preceding Theorem 6.2.1, μ_0 is the maximizer of Γ. As μ_0 is in the interior of $\mathcal{M}$, this maximum is a local maximum. This implies that the eigenvalues of $H\Gamma(\mu_0)$ are all strictly negative. The only way for (6.15) to hold asymptotically is therefore if

$$\widehat{s}_n = s_n^* + o_{\mathrm{Pr}}(1).$$

This concludes the proof. ■

The limit distribution proved in the previous theorem is not particularly informative as it involves the derivative of the, in general, unknown function λ. We will therefore present a corollary where we provide alternative characterizations of certain quantities.

Corollary 6.3.3. *Assume the conditions of Theorem 6.3.2 hold true. Then*

$$\sqrt{n}(\widehat{\mu} - \mu_0) = -H\Gamma(\mu_0)^{-1}\xi_0^T S(0)V_n(0) + o_{\mathrm{Pr}}(1). \tag{6.16}$$

where

$$\xi_0 = \frac{\partial}{\partial\mu}\Big|_{\mu_0} \mathrm{E}\, M(\mu).$$

If Leibniz integral theorem can be applied twice to Γ*, this simplifies to*

$$\sqrt{n}(\widehat{\mu} - \mu_0) = -\left(\xi_0^T S(\mu_0)\xi_0\right)^{-1}\xi_0 S(\mu_0)V_n(0) + o_{\Pr}(1), \tag{6.17}$$

ensuring the following limit distribution:

$$\sqrt{n}(\widehat{\mu} - \mu_0) \xrightarrow{d} N(0, \Sigma)$$

with

$$\Sigma = \left(\xi_0 S(\mu_0)^{-1}\xi_0\right)^{-1}. \tag{6.18}$$

Remark 6.3.4. Suppose λ is twice continuously differentiable and there is an open subset U of $\mathcal{M}$ on which $\mu \mapsto m(y, \mu)$ is twice differentiable. Assume further that every first and second order partial derivative of $\mu \mapsto m(y, \mu)$ is bounded by some integrable function on U, not depending on μ. It can then be shown that the conditions of both applications of Leibniz integral theorem are satisfied. This can be achieved by computing derivatives and arguing as we have done multiple times in this section. Details are omitted.

Proof. Application of the implicit function theorem, see e.g. Lindstrøm 2017, p. 212, shows

$$\lambda'(\mu_0) = S(\mu_0)^{-1}\xi_0.$$

This is enough for (6.16).

For (6.17), notice that

$$\frac{\partial}{\partial \mu}\log[1 + \lambda(\mu)M(\mu)] = \frac{M'(\mu)\lambda(\mu) + \lambda'(\mu)M(\mu)}{1 + \lambda(\mu)M(\mu)}.$$

So the hessian matrix of $\log(1 + \lambda(\mu)M(\lambda(\mu)))$ is given by

$$H(\mu) = \frac{\partial}{\partial \mu}\frac{M'(\mu)^T\lambda(\mu) + \lambda'(\mu)^T M(\mu)}{1 + \lambda(\mu)M(\mu)}.$$

Using the standard rules of matrix calculus and the fact that $\lambda(\mu_0) = 0$, we arrive at the following expression

$$\begin{aligned} H(\mu_0) =& H\lambda(\mu_0)M(\mu_0) + M'(\mu_0)^T\lambda'(\mu_0) + \lambda'(\mu_0)^T M'(\mu_0) \\ & - \lambda'(\mu_0)^T M(\mu_0)^2\lambda'(\mu_0). \end{aligned}$$

Taking the expected value of the above equation and entering in $\lambda'(\mu_0) = S(\mu_0)^{-1}\xi_0$, $\mathrm{E}\, M(\mu_0) = 0$, $\mathrm{E}\, M(\mu_0)^2 = S(\mu_0)$ and $\mathrm{E}\, M'(\mu_0) = \xi_0$, results in

$$\mathrm{E}\, H(\mu_0) = 0 + \xi_0^T S^{-1}\xi_0 + \xi_0^T S^{-1}\xi_0 - \xi_0 S^{-1} S S^{-1}\xi_0 = \xi_0^T S^{-1}\xi_0.$$

By assumption

$$H\Gamma(\mu_0) = \mathrm{E}\, H(\mu_0).$$

So,

$$H\Gamma(\mu_0) = \xi_0^T S^{-1}\xi_0,$$

showing (6.17) and (6.18).

In the above we have used

$$\xi_0 = \frac{\partial}{\partial \mu}\Big|_{\mu_0} \mathrm{E}\, M(\mu) = \mathrm{E}\, M'(\mu_0). \tag{6.19}$$

The condition (6.4) ensures that the derivative of $\mu \mapsto m(y, \mu)$ is bounded by p_2. This has finite expectation, and hence (6.19) follows from Leibniz integral theorem. ■

As mentioned before, asymptotic normality of the maximum empirical likelihood estimator is not a new discovery. The first, and most famous, article establishing a normal limit of

$$\sqrt{n}(\widehat{\mu} - \mu_0)$$

is Qin and Lawless 1994. In this article it is shown that

$$\sqrt{n}(\widehat{\mu} - \mu_0) \xrightarrow{d} N(0, \Sigma_0)$$

where

$$\Sigma_0 = \mathrm{E}\, M'(\mu_0)^T \mathrm{E}\, M(\mu_0)^2 \mathrm{E}\, M'(\mu_0). \tag{6.20}$$

In the proof of Theorem 6.3.2 we showed

$$\xi_0 = \frac{\partial}{\partial \mu}\Big|_{\mu_0} \mathrm{E}\, M(\mu) = \mathrm{E}\, M'(\mu_0).$$

Hence, Σ given in (6.18) is equal to Σ_0 from (6.20). Because of this, the conclusion of Qin and Lawless 1994s result agrees with ours.

As noted at the beginning of this chapter, a solution to the equation

$$0 = \frac{1}{n} \sum_{i=1} M_i(\mu) \tag{6.21}$$

will always maximize the empirical likelihood function. The maximum empirical likelihood estimator is therefore also a Z-estimator in many situations. Such estimators are both consistent for the solution to

$$\mathrm{E}\, M(\mu) = 0$$

and asymptotically normal, after proper scaling and centering. Again we refer to section 3.2 in Huber 2009 for the relevant definitions and theorems. If we use this approach to find the limit distribution of the maximum empirical likelihood estimator, we get

$$\sqrt{n}(\widehat{\mu}_n - \mu_0) \xrightarrow{d} \mathrm{N}(0, \Sigma_0)$$

with Σ_0 as in (6.20). Hence, Theorem 6.3.2 is in agreement with the results we get using asymptotic properties of Z-estimators.

That being said, one thing distinguishes the results of this chapter from both general theory concerning solutions to (6.21) and theorems derived by Qin and Lawless 1994. We have utilized an alternative characterization of the

empirical likelihood function and showed that maximization of it is very similar to an M-estimation problem. This alternative way of looking at the empirical likelihood function allows us to work with hybrid combinations of EL_n with other quantities. In the upcoming part of the thesis we will use the empirical likelihood function to robustify the, perhaps, most famous class of M-estimators: maximum likelihood estimators. The theorems from this chapter will be crucial in Chapter 8 when neither the result of Qin and Lawless 1994 nor general theory concerning solutions to (6.21) suffice.

PART II

Hybrid Likelihood

CHAPTER 7

Under model conditions

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables following some unknown distribution. Assume we want to fit a family, indexed by a parameter $\theta \in \mathbb{R}^p$, to these variables. A standard way of doing this, is to estimate θ by the maximizer of the likelihood function. When the parametric model is specified correctly, maximum likelihood estimates are both consistent and has a limiting distribution with variance attaining the Cramér-Rao lower bound. Hence, this technique is a good choice when the true density is a member of the parametric family. There is, however, no guarantee that the method works well when this assumption does not hold true. Furthermore, there might be situations where robust estimation of certain parameters are more important than the overall model fit. In such cases, we might want to pay special attention to these variables and ensure that their estimates are extra robust. In this part of the thesis we will discuss one particular way of achieving this. The method uses the hybrid likelihood function, combining parametric and empirical likelihoods.

The hybrid likelihood function was introduced in Hjort, I. McKeague, and Van Keilegom 2018, and in this chapter will summarize and discuss the results and definitions given there. At the end of the chapter, we will formulate and prove a profiling result for the hybrid likelihood function. We will also provide some examples, illustrating how the theory can be applied. Among other things, we will take a second look at the Correlation of War data set from Section 4.3 and see if the added strength from a parametric model can strengthen our results.

7.1 The definition

Before we can properly define the hybrid likelihood function, we need to introduce some concepts and notation. Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables following some unknown distribution with density function, f. Assume we wish to fit some parametric family,

$$\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \},$$

to this data, but that robust estimation of certain parameters, $\mu(f) \in \mathbb{R}^q$, are of extra importance. We will call these control parameters, and pay special attention to them when fitting the parametric model. We are not necessarily interested in estimating these control parameters themselves, but we want to estimate θ in a way ensuring that the parametric estimator of μ is a robust

one. In mathematical notation, this means seeking $\widehat{\theta}$ such that $\mu(f_{\hat{\theta}})$ is a robust estimate of $\mu(f)$.

We will assume the control parameters can be characterized with the estimating equation

$$\mathrm{E}\, m(Y, \mu) = 0,$$

for some estimating function, $m\colon \mathbb{R}^{d+q} \to \mathbb{R}^q$, and $Y \sim f$. Suppose further that for each fixed $\theta \in \Theta$, there is a corresponding $\mu(\theta) \in \mathbb{R}^q$ such that

$$\mathrm{E}\, m[Y, \mu(\theta)] = 0, \tag{7.1}$$

when $Y \sim f_\theta$ and m is as before.

To ensure robust estimates of the control parameters, we will use the empirical likelihood function,

$$\mathrm{EL}_n(\mu) = \max\left\{ \prod_{i=1}^n n w_i \;\middle|\; \sum_{i=1}^n w_i m(Y_i, \mu) = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}. \tag{7.2}$$

Estimating θ with $\widehat{\theta}$ such that the empirical likelihood function, evaluated at $\mu(\widehat{\theta})$ is large, will ensure that $\mu(\widehat{\theta})$ is a robust estimate of the control parameter. This follows from the arguments and theory given in the previous part of the thesis. In every other aspect, however, there is no guarantee that $f_{\hat{\theta}}$ will be a good estimate of the true density. Another, and popular, way of estimating θ is to maximize the parametric likelihood function,

$$L_n(\theta) = \prod_{i=1}^n f_\theta(Y_i),$$

and in this part of the thesis, we will use the parametric likelihood function to ensure a good overall model fit.

We will combine the ideas from the previous paragraph to fit parametric families, while taking both overall model fit and robust estimation of the control parameters into account. This will be done by estimating θ in a way ensuring that both the parametric and empirical likelihood function, evaluated at the corresponding values, are large. We will combine the maps to form what we call the hybrid likelihood function to achieve this. The ensuing definition formalizes the idea.

Definition 7.1.1 (Hjort, I. McKeague, and Van Keilegom 2018)**.** Let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables, $\theta \in \mathbb{R}^p$ and f_θ, for $\theta \in \Theta$, a parametric family we wish to fit to the data. With control parameters $\mu(\theta) \in \mathbb{R}^q$ characterized by (7.1), the hybrid likelihood function is defined as

$$H_n(\theta) = L_n(\theta)^{1-a} \mathrm{EL}_n[\mu(\theta)]^a,$$

where $a \in [0, 1)$ is a balance parameter, $L_n(\theta) = \prod_{i=1}^n f_\theta(Y_i)$ is the parametric likelihood of the data and $\mathrm{EL}_n[\mu(\theta)]$ is as in (7.2).

The maximizer of this function is called the maximum hybrid likelihood estimator and will be denoted by $\widehat{\theta}_{hl}$ in this thesis.

The logarithm is an increasing function. Because of this, maximizing $H_n(\theta)$ is equivalent to maximizing $h_n(\theta) = \log H_n(\theta)$. So,

$$\widehat{\theta}_{hl} = \operatorname*{argmax}_{\theta\in\Theta} h_n(\theta) = \operatorname*{argmax}_{\theta\in\Theta} \{ a\,\ell_n(\theta) + (1-a)\log \mathrm{EL}_n[\mu(\theta)] \}. \qquad (7.3)$$

Here ℓ_n is the log-likelihood. From (7.3) it is clear that $\hat{\theta}_{hl}$ is the maximizer of a convex combination of the log-likelihood function and logarithm of the empirical likelihood function, with the balance parameter deciding how much weight should be put on each term. For small values of a, $\ell_n(\theta)$ will be the dominant term, resulting in $\widehat{\theta}_{hl}$ being close to the maximum likelihood estimator of θ. With larger values of the balance parameter, $\log \mathrm{EL}_n(\mu(\theta))$ will dominate, and $\hat{\theta}_{hl}$ will be close to a value of θ such that $\mu(\theta)$ equals the maximum empirical likelihood estimator. For most values of a, however, $\hat{\theta}_{hl}$ will be a trade-off between the two, and $f_{\hat{\theta}_{hl}}$ will be a good overall parametric fit and $\mu(\widehat{\theta}_{hl})$ a robust estimator of the control parameters.

Before we continue, we want to comment quickly on how to compute $\mathrm{EL}_n[\mu(\theta)]$. Although the empirical likelihood function is a function of the control parameters only, $\mathrm{EL}_n[\mu(\theta)]$ is a function of θ. The value is computed by first finding $\mu(\theta)$ such that (7.1) holds. $\mathrm{EL}_n[\mu(\theta)]$ can then be calculated as explained in Section 2.2 using (7.2) with $\mu = \mu(\theta)$. To illustrate, let $\mathcal{F}$ be the family of Weibull(λ, k) densities,

$$f_{\lambda,k}(y) = k\lambda^{-k} y^{k-1} \exp\bigl(-\lambda^{-k} y^k\bigr) \quad \text{for } y, \lambda, k > 0,$$

and the control parameter the third quartile. In this example, $\theta = (\lambda, k)^T$. Furthermore, $\mu(\theta) = \lambda(\log 4)^{\frac{1}{k}}$, as this is the third quartile in a Weibull(λ, k) distribution. Lastly, we would use $m(y, \mu) = I(y \le \mu) - 0.75$ as this results in an estimating equation characterizing the third quartile.

7.2 The main results

The definition and concepts from the previous section are less useful if we know nothing about the distribution of $h_n(\theta)$ or $\hat{\theta}_{hl}$. Luckily, both these quantities have large sample distributions after proper scaling and centering. This is stated and proved in Hjort, I. McKeague, and Van Keilegom 2018, and in this section we will present the main results from the paper.

We start with some notation and minor observations. As before, we will let $Y_1, \ldots, Y_n$ be i.i.d. random variables, following some distribution with density f, and

$$\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^p \}$$

the parametric family we wish to fit to these variables. In addition, we will assume the control parameters can be characterized with the estimating equation

$$\mathrm{E}\, m(Y, \mu) = 0$$

for some $m\colon \mathbb{R}^{d+q} \to \mathbb{R}^q$.

In this section, we will assume $f = f_{\theta_0}$ for some $\theta_0 \in \Theta$. In this case, the solution to (7.1) is $\mu(\theta_0)$. Furthermore, for $Y \sim f_{\theta_0}$, $\mathrm{E}\, u(Y, \theta_0) = 0$, where

$$u(y, \theta) = \frac{\partial}{\partial\theta} \log f_\theta(y)^T$$

is the score function. By the central limit theorem,

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u(\,Y_i, \theta_0) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m[\,Y_i, \mu(\theta_0)] \end{pmatrix} = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} u(\,Y_i, \theta_0) \\ m[\,Y_i, \mu(\theta_0)] \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) \xrightarrow{d} \begin{pmatrix} U \\ V \end{pmatrix}$$

where

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathrm{N}_{p+q}\,(0, \Sigma) \quad \text{with} \quad \Sigma = \mathrm{Var} \begin{pmatrix} u(\,Y, \theta_0) \\ m[\,Y, \mu(\theta_0)] \end{pmatrix} = \begin{pmatrix} J & C \\ C^T & W \end{pmatrix}. \tag{7.4}$$

Since we have assumed that f is, indeed, equal to f_{θ_0},

$$J = \mathrm{Var}\, u(Y, \theta_0) = -\mathrm{E}\Big(\frac{\partial^2}{\partial\theta \partial\theta^T}\Big|_{\theta_0} \log f_\theta(\,Y) \Big),$$

the Fisher matrix. Furthermore $C = \mathrm{Cov}[u(\,Y, \theta_0), m(\,Y, \theta_0)]$ and $W = \mathrm{Var}\, m(\,Y, \theta_0)$.

Assume Σ is positive definite and define the following process on $\mathbb{R}^p$

$$A_n(s) = h_n\Big(\theta_0 + \frac{s}{\sqrt{n}}\Big) - h_n(\theta_0).$$

The main result of section 2 in Hjort, I. McKeague, and Van Keilegom 2018 involves convergence of A_n to a process, A, in the normed space $\ell^\infty(K)$ for arbitrary compact sets, K. To present the theory from Hjort, I. McKeague, and Van Keilegom 2018, we will therefore start by going through the most important conditions needed for $A_n \xrightarrow{d} A$ in $\ell^\infty(K)$.

Firstly, we need to make some assumptions about the behavior of m. Let V_n be defined as

$$V_n(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m\Big[\,Y_i, \mu\Big(\theta_0 + \frac{s}{\sqrt{n}}\Big)\Big].$$

We will need this quantity to be more or less linear for large values of n. In Hjort, I. McKeague, and Van Keilegom 2018 this is achieved by assuming that over compact sets, K,

$$\sup_{s \in K} ||V_n\,(s) - V_n\,(0) - \xi_n s|| \xrightarrow{\Pr} 0, \tag{7.5}$$

for some $q \times p$-matrix, ξ_n, going in probability to some $q \times p$-matrix, ξ_0. As explained in Hjort, I. McKeague, and Van Keilegom 2018, p. 2396, this holds with

$$\xi_n = n^{-1} \sum_{i=1}^{n} \frac{\partial m}{\partial \theta}[\,Y_i, \mu(\theta_0)] \quad \text{and} \quad \xi_0 = E\Big(\frac{\partial m}{\partial \theta}[\,Y, \mu\,(\theta_0)] \Big)$$

if m is sufficiently smooth in θ. For quantiles, (7.5) holds with

$$\xi_n = \xi_0 = f_{\theta_0}[\mu(\theta_0)] \nabla \mu(\theta_0).$$

Before stating the main result of section 2 in Hjort, I. McKeague, and Van Keilegom 2018, we need to define three quantities:

$$U^* = (1-a)U - a\xi_0^T W^{-1} V, \tag{7.6}$$

$$J^* = (1-a)J + a\xi_0^T W^{-1}\xi_0^T \text{ and} \tag{7.7}$$

$$K^* = (1-a)^2 J + a^2 \xi_0^T W^{-1}\xi_0 - a(1-a)\left(CW^{-1}\xi_0 + \xi_0^T W^{-1} C^T\right). \tag{7.8}$$

U^*, J^* and K^* play similar roles in hybrid likelihood theory as U, J and

$$K = \operatorname{Var} u(Y, \theta_0)$$

do in maximum likelihood theory. This should be evident from the following theorem.

Theorem 7.2.1 (Hjort, I. McKeague, and Van Keilegom 2018)**.** *Under the conditions stated above, as well as additional smoothness conditions on* $\log f_\theta$ *and* $EL_n(\theta)$ *given in Hjort, I. McKeague, and Van Keilegom 2018,* A_n *converges in distribution to*

$$A(s) = s^T U^* - \frac{1}{2} s^T J^* s$$

in the space $\ell^\infty(K)$*, equipped with the uniform topology, for each compact subset* K *of* $\mathbb{R}^p$*.*

Theorem 7.2.1 ensures, not only pointwise convergence, but convergence of A_n to A as stochastic processes in $\ell^\infty(K)$. We will use Theorem 7.2.1 when working with the profile hybrid likelihood function in the next section, but the result also has the following corollary.

Corollary 7.2.2 (Hjort, I. McKeague, and Van Keilegom 2018)**.** *Under the conditions stated in Hjort, I. McKeague, and Van Keilegom 2018, p. 11 the follwing three properties hold*

(1) $\widehat{\theta}_{hl}$ *is consistent for* θ_0*.*

(2) $\sqrt{n}\left(\widehat{\theta}_{hl} - \theta_0\right) \xrightarrow{d} (J^*)^{-1}U^* \sim N_p\big(0, (J^*)^{-1}K^*(J^*)^{-1}\big)$

(3) $2\left(h_n(\widehat{\theta}_{hl}) - h_n(\theta_0)\right) \xrightarrow{d} (U^*)^T (J^*)^{-1} U^*$.

This result is very similar to corresponding versions in traditional maximum likelihood theory, see e.g. Appendix A.5 of Schweder and Hjort 2016, p. 27.

7.3 Choosing the balance parameter

When fitting a model to data using maximum hybrid likelihood, we need to decide what control and balance parameter to use. The hybrid likelihood machinery is applied only when we want model robust estimates of the control parameter. What value to choose will depend on what we want to infer from our data and is highly problem specific. Constructing a general procedure for choosing the control parameter, is therefore not something that neither can, nor should, be done. The balance parameter, however, is a tuning parameter

that ought be set to give the best trade-off between robustness and efficiency. The optimal value of a will depend on how well the parametric model fits to the data and how well relevant quantities are estimated. In this section we will present a general way of choosing a when the model is specified correctly. As in the previous section the ideas presented here are credited to Hjort, I. McKeague, and Van Keilegom 2018. We therefore refer to this paper for further reading and discussion.

Maximum hybrid likelihood is used in cases where estimation of the control parameter is of extra importance. Because of this, we should select an a resulting in a good estimate of μ. A natural way of doing this is to choose balance parameter such that the mean squared error of the resulting maximum hybrid likelihood estimate of the control parameter is small. Provided the necessary regularity conditions, the maximum likelihood estimate of the control parameter is unbiased and its variance attains the Cramer-Rao lower bound as the sample size goes to infinity. Because of this, minimizing the asymptotic mean squared error would always result in a choice of $a = 0$, when the model is specified correctly. That being said, we use the hybrid likelihood machinery because we are willing to give up on a certain amount of efficiency, in favor of extra robust estimates of μ. To choose the balance parameter, we should therefore start by deciding just how much we are willing to give up in terms of efficiency. Afterwards, the largest value of a resulting in an estimator of μ with variance less than this threshold can be chosen. The following scheme describes this procedure in more detail:

(1) Decide on a focus parameter, ψ, such that this has the value $g(\theta)$ in the parametric model and $g\colon \mathbb{R}^p \to \mathbb{R}$ is a differentiable function.

(2) Decide on an acceptable increase in variance.

(3) For each value of a in a grid of points between 0 and 1, estimate the asymptotic variance of $g(\widehat{\theta}_{hl})$ with the formula

$$\kappa_a^2 = \frac{\nabla g(\widehat{\theta}_{hl})^T (J^*)^{-1} K^* (J^*)^{-1} \nabla g(\widehat{\theta}_{hl})}{n}.$$

(4) Choose the largest value of a for which $\widehat{\kappa}_a$ is less than the threshold set in (2).

To generalize the procedure as much as possible, we have allowed for all focus parameters in (1). In practice, however, the most natural choice will often be the control parameter. In step (2) a general way of setting a treshold is to use relative increase in variance. Accepting an increase of 5% in standard deviation from the maximum likelihood estimator of ψ, would for example result in confidence intervals that are 10% stretched. Step (4) would then involve computing κ_a/κ_0 for each a and choosing a balance parameter for which this quantity is no larger than 1.05.

The procedure presented in this section attempts to choose a such that the limiting mean squared error has certain desirable properties. In some situations, however, other loss functions might be of interest. Prediction error or absolute loss are two examples of functions that are important in certain applications. The procedure above can in many cases be modified to such situations.

Methods for risk estimation, like cross validation or bootstrap are, of course, also options when choosing what balance parameter to use. Such methods are, however, computer-intensive as they require multiple models to be fit. Using the procedure presented here, we need only fit as many models as the number of balance parameters we wish to consider. This significantly reduces computation time, and can therefore be desirable in many situations.

We will use this procedure to choose the balance parameter in Section 7.5. In Chapter 8 we will consider the situation where the true underlying distribution is not necessarily a member of the parametric family fit to the data. In such cases, the arguments concerning limiting efficiency and unbiasedness from this section does not hold true. Because of this, more complicated procedures than the ones presented in this section are needed. This will be the topic for Chapter 9.

7.4 Profile hybrid likelihood

As explained in Chapter 3, we are in often interested in a focus parameter, $\psi(\theta)$, rather than the full parameter vector, θ. In standard maximum likelihood theory we have the concept of profile likelihoods and Wilks theorem to help in such situations. In this section, we will present analogous definitions and results for the hybrid likelihood function.

We start by defining the profile hybrid likelihood function. This is done similarly as in standard maximum likelihood theory.

Definition 7.4.1 (Profile hybrid likelihood)**.** Let $h_n(\theta)$ be the logarithm of the hybrid likelihood function. For a focus parameter $\psi = g(\theta)$, we define the profile hybrid likelihood function, $h_{n,prof}(\psi)$, in the following way

$$h_{n,prof}(\psi) = \max_{g(\theta)=\psi} h_n(\theta). \tag{7.9}$$

The profile likelihood function is used frequently in maximum likelihood theory. The reason for this is that the deviance can be constructed based on this quantity. Under certain conditions, this has a known limit distribution. We can therefore use the profile likelihood function to make inference about ψ. The following theorem allows us to use Definition 7.4.1 in the same way.

Theorem 7.4.2. *[Hjort, I. McKeague, and Van Keilegom 2018, Appendix S.6] As before let $h_n(\theta)$ be the logarithm of the hybrid likelihood function. Furthermore assume $g\colon \mathbb{R}^p \to \mathbb{R}$ is a map for which the second order partial derivatives are all continuous and the focus parameter is given by $\psi = g(\theta)$.*

Let D_n denote the following quantity

$$D_n(\psi) = 2\left(h_n(\widehat{\theta}_{hl}) - h_{n,prof}(\psi)\right), \tag{7.10}$$

$\psi_0 = g(\theta_0)$, θ_0 be the true value of θ and U^*, J^* and K^* be as defined in Section 7.2. Assume the conditions of Theorem 7.2.1 and Corollary 7.2.2 hold true. Then

$$\kappa \cdot D_n(\psi_0) \xrightarrow{d} \chi^2_1$$

where

$$\kappa = \frac{b^T (J^*)^{-1} b}{b^T (J^*)^{-1} K^* (J^*)^{-1} b}. \tag{7.11}$$

and b denotes the gradient of g at θ_0.

The name D_n is chosen to emphasize that this quantity is the hybrid counterpart to the deviance function used in maximum likelihood theory. With this formulation, the similarity between Theorem 7.4.2 and Wilks theorem is clear. See Theorem 2.4 in Schweder and Hjort 2016, p. 35 for a statement of this result within and appendix A.5 in the same book for a statement of the result outside of model conditions.

Theorem 7.4.2 is stated without proof in Hjort, I. McKeague, and Van Keilegom 2018. We will now prove the result using a slight modification of the arguments given in Remark 2.5 in Schweder and Hjort 2016, pp. 36–37.

Proof. We will assume $g(\theta) = b^T \theta$. The result can be extended to hold for all g with continuous second order partial derivatives by arguing as in the proof of Theorem 3.0.5

Let K be the compact set

$$K = \{\, s \in \mathbb{R}^p \mid \|s\| \leq M \,\},$$

for some $M > 0$ and

$$T = \left\{\, b^T s \;\middle|\; s \in K \,\right\}.$$

Notice that T is the image of K under $s \mapsto b^T s$. This is a continuous map, and compactness is preserved by images of such functions. Hence, T is a compact set.

Now let $f \colon \ell^\infty(K) \to \ell^\infty(T)$ be the map

$$f(h)(t) = \max_{b^T s = t} h(s).$$

This is a continuous function. Since $A_n \xrightarrow{d} A$ in $\ell^\infty(K)$ by Theorem 7.2.1, the continuous mapping theorem ensures $E_n \xrightarrow{d} E$ in $\ell^\infty(T)$, where

$$E_n(t) = f(A_n)(t) = \max_{b^T s = t} A_n(s) \quad \text{and} \quad E(t) = f(A)(t) = \max_{b^T s = t} A(s).$$

Furthermore, $h \mapsto h - h(0)$ is a continuous operator on $\ell^\infty(T)$. Hence,

$$B_n = E_n - E_n(0) \xrightarrow{d} E - E(0) = B$$

in $\ell^\infty(T)$ by the continuous mapping theorem. Combining this with the fact that $h \mapsto \max_{t \in T} h$ is continuous as a function from $\ell^\infty(T)$ into $\mathbb{R}$, shows

$$\max_{t \in T} B_n(t) \xrightarrow{d} \max_{t \in T} B(t),$$

as random variables in $\mathbb{R}$.

In the above, the compact set K was arbitrary. So, using similar arguments as in the proof of Theorem 3.0.5, we can show that

$$\max_{t\in\mathbb{R}} B_n(t) \xrightarrow{d} \max_{t\in\mathbb{R}} B(t)$$

when the maximizers of B_n are stochastically bounded. Notice that,

$$B_n(t) = \max_{b^T s=t} A_n(s) - \max_{b^T s=0} A_n(s) = h_{n,prof}\left(\psi_0 + \frac{t}{\sqrt{n}}\right) - h_{n,prof}(\psi_0).$$

Since $h_{n,prof}(\psi)$ is maximized for $\psi = b^T\widehat{\theta}_{hl}$, the maximizer of B_n in $\mathbb{R}$ is

$$b^T\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big).$$

Furthermore, $\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big)$ is stochastically bounded by Corollary 7.2.2, so

$$b^T\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) = O_{\Pr}(1).$$

Hence,

$$\max_{t\in\mathbb{R}} B_n(t) \xrightarrow{d} \max_{t\in\mathbb{R}} B(t) \tag{7.12}$$

by the previous argument.

Theorem 7.4.2 follows more or less directly from this. Since

$$B_n(t) = h_{n,prof}\left(\psi_0 + \frac{t}{\sqrt{n}}\right) - h_{n,prof}(\psi_0),$$

and $h_{n,prof}(\psi)$ is maximized for $\psi = b^T\widehat{\theta}_{hl}$,

$$\begin{aligned}\max_{t\in\mathbb{R}} B_n(t) &= h_{n,prof}\left(b^T\widehat{\theta}_{hl}\right) - h_{n,prof}(\psi_0)\\ &= h_n\left(\widehat{\theta}_{hl}\right) - h_{n,prof}(\psi_0)\\ &= \frac{1}{2}D_n(\psi_0).\end{aligned}$$

Combining all of this, shows

$$D_n(\psi_0) = 2\cdot\max_{t\in\mathbb{R}} B_n(t) \xrightarrow{d} 2\cdot\max_{t\in\mathbb{R}} B(t).$$

Computing the limit is all that remains to complete the argument. Notice first that

$$2\cdot\max_{t\in\mathbb{R}} B(t) = 2\cdot\max_{t\in\mathbb{R}}\max_{b^T s=t} A(s) - 2\cdot\max_{b^T s=0} A(s).$$

A is a quadratic function, so its maximum over the set

$$\left\{\, s\in\mathbb{R}^p \;\middle|\; b^T s = t \,\right\}$$

can be found using the method of Lagrange multipliers. The solution is given by

$$\max_{b^T s=t} A(s) = \frac{1}{2}U^*(J^*)^{-1}U^* - \frac{1}{2}\cdot\frac{\left(t - b^T(J^*)^{-1}U^*\right)^2}{b^T(J^*)^{-1}b}.$$

The second term is always non-positive, so $\max_{b^T s=t} A(s)$ is at its largest when this is equal to zero. Hence

$$\max_t \max_{b^T s=t} A(s) = \max_{b^T s=b^T(J^*)^{-1}U^*)} A(s) = \frac{1}{2}U^*(J^*)^{-1}U^*.$$

Furthermore,

$$\max_{b^T s=0} A_n(s) = \frac{1}{2}U^*(J^*)^{-1}U^* - \frac{1}{2}\cdot\frac{\left(b^T(J^*)^{-1}U^*\right)^2}{b^T(J^*)^{-1}b}.$$

Hence,

$$\begin{aligned} 2\cdot\max_{t\in\mathbb{R}} B(t) &= U^*(J^*)^{-1}U^* - U^*(J^*)^{-1}U^* + \frac{\left(b^T(J^*)^{-1}U^*\right)^2}{b^T(J^*)^{-1}b} \\ &= \frac{\left(b^T(J^*)^{-1}U^*\right)^2}{b^T(J^*)^{-1}b}. \end{aligned}$$

We have now shown

$$D_n(\psi) \xrightarrow{d} \frac{\left(b^T(J^*)^{-1}U^*\right)^2}{b^T(J^*)^{-1}b}.$$

Since

$$\mathrm{Var}\left(b^T(J^*)^{-1}U^*\right) = b^T(J^*)^{-1}K^*(J^*)^{-1}b$$

and $U^* \sim \mathrm{N}(0, K^*)$,

$$\frac{\left(b^T(J^*)^{-1}U^*\right)^2}{b^T(J^*)^{-1}b} \sim \frac{b^T(J^*)^{-1}K^*(J^*)^{-1}b}{b^T(J^*)^{-1}b}\cdot\chi_1^2.$$

Therefore,

$$D_n(\psi_0) \xrightarrow{d} \frac{b^T(J^*)^{-1}K^*(J^*)^{-1}b}{b^T(J^*)^{-1}b}\cdot\chi_1^2,$$

or equivalently

$$\kappa\cdot D_n(\psi_0) \xrightarrow{d} \chi_1^2,$$

with κ defined as in (7.11). ■

7.5 Examples

We will now illustrate how the theorems and definitions developed in this chapter can be used to make inference. First, we will consider two examples with simulated data. Afterwards we will go back to Section 4.3 and investigate whether a hybrid approach can strengthen our results regarding the median number of battle deaths.

The third quartile in a Weibull distribution

We will start by working with a simulated data set of 100 i.i.d. variables following a Weibull distribution with shape parameter 2 and scale parameter 15. The results from the previous sections will be used to make inference about the third quartile in the distribution of the data.

The density in a Weibull(λ, k) distribution is

$$f_{\lambda,k}(y) = k\lambda^{-k} y^{k-1} \exp\bigl(-\lambda^{-k} y^k\bigr).$$

So the log-density is given by

$$f_{\lambda,k}(y) = \log k - k \log \lambda + (k-1) \log y - \lambda^{-k} y^k,$$

which results in a log-likelihood on the form

$$\ell_n(\theta, k) = n \log k - nk \log \lambda + (k-1) \sum_{i=1}^{n} \log y_i - \lambda^{-k} \sum_{i=1}^{n} y_i^k.$$

The maximizer of this function has no closed form expression, but can be found numerically. For our simulated data set, the result was $(\hat{\lambda}_{ml}, \hat{k}_{ml}) \approx (2.117, 15.557)$.

The maximum likelihood estimate for the third quartile is

$$\hat{\lambda}_{ml} (\log 4)^{1/\hat{k}_{ml}} \approx 18.153$$

as

$$g(\lambda, k) = \lambda [-\log(1-q)]^{1/k}$$

is the q-quantile in a Weibull(λ, k)-distribution. The empirical 0.75-quantile of the data is 17.830 which is closer to the true value: 17.661. In this example, we will use the hybrid likelihood function to control for this quantity in an attempt to decrease the bias of the estimate of the third quartile. To achieve this, we will use the control parameter $\mu(\lambda, k) = \lambda (\log 4)^{1/k}$ and estimating function $m(y, \mu) = I(Y \leq \mu) - 0.75$. Since the data really is Weibull(λ, k)-distributed,

$$\mathrm{E}\, m[Y, \mu(\lambda_0, k_0)] = \Pr\Bigl(Y \leq \lambda_0 (\log 4)^{1/k_0}\Bigr) - q = 0,$$

for the true values, λ_0 and k_0, of λ and k and $Y \sim$ Weibull(λ_0, k_0). Furthermore, (7.5) is satisfied with

$$\xi_n = \xi_0 = f_{\lambda_0, k_0}[\mu(\lambda_0, k_0)] \cdot \nabla \mu(\lambda_0, k_0).$$

This was explained in the previous chapter, right before the statement of Theorem 7.2.1. For additional arguments, see Stute 1982 or argue similarly as in example 19.29 in Vaart 1998, p. 283.

Before we can construct the hybrid likelihood function, we need to choose what balance parameter to use. We will do this as explained in Section 7.3, choosing a in such a way that the variance of the maximum hybrid likelihood estimate of the third quartile is no more than 10% greater than that of the maximum likelihood estimate. Confidence intervals for μ, obtained with

Corollary 7.2.2 and the delta method, will then be 20% wider than the corresponding ones obtained with maximum likelihood. After implementing the procedure of Section 7.3 for our data set, we end up with the value $a = 0.68$.

We can now construct confidence curves and intervals for focus parameters $\psi = g(\lambda, k)$ in two ways. The first approach is to use the normal approximation in Corollary 7.2.2 and the delta method. This results in symmetric confidence intervals centered around the maximum hybrid likelihood estimate. Another approach is to use Theorem 7.4.2. In the following we will utilize both methods to make inference about the third quartile.

We start with the normal approximation. By Corollary 7.2.2,

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) \xrightarrow{d} N\big[0, (J^*)^{-1}K^*(J^*)^{-1}\big], \tag{7.13}$$

where $\widehat{\theta}_{hl}$ is the maximum hybrid likelihood estimator and θ_0 the true value of (λ, k). The matrices J^* and K^* are as defined in Section 7.2. The true value of the third quartile is given by $g(\theta_0)$ as the data really follows a Weibull distribution.

Applying the delta method to (7.13) shows

$$\sqrt{n}\Big(g(\widehat{\theta}_{hl}) - g(\theta_0)\Big) \xrightarrow{d} \nabla g(\theta_0) N\big(0, (J^*) {-} 1 K^* (J^*)^{-1}\big).$$

Because of this,

$$g(\widehat{\theta}_{hl}) \overset{d}{\approx} \mathrm{N}\left(g(\theta_0), \frac{1}{n}\nabla g(\theta_0)^T (J^*)^{-1} K^* (J^*)^{-1} \nabla g(\theta_0)\right).$$

We can use this approximate distribution to construct confidence intervals and curves for the third quartile. In this example, we will only present the confidence curve as it gives a visual summary of our results and confidence intervals of all levels can be read off it. With

$$\tau^2 = \frac{1}{n}\nabla g(\theta_0)^T (J^*)^{-1} K^* (J^*)^{-1} \nabla g(\theta_0)$$

the confidence curve is given by

$$cc_1(\mu) = \left|1 - 2\Phi\left(\frac{g(\widehat{\theta}_{hl}) - g(\theta_0)}{\tau}\right)\right|,$$

where Ψ is the cumulative distribution function in the standard normal distribution. See section 3.4 in Schweder and Hjort 2016 for more details. The matrices J^* and K^*, as well as the gradient of g at θ_0, can be estimated by their canonical estimators as explained in Hjort, I. McKeague, and Van Keilegom 2018. All of this was done for our simulated data set, and the resulting approximate confidence curve can be found in Figure 7.1 together with a curve obtained similarly using corresponding results in maximum likelihood theory.

From Figure 7.1, we notice that the confidence curve constructed using hybrid likelihood is slightly wider than the one obtained using maximum likelihood theory. This is not surprising as maximum likelihood estimators are asymptotically the most efficient in terms of mean squared error. An increase in variance when using maximum hybrid likelihood rather than maximum

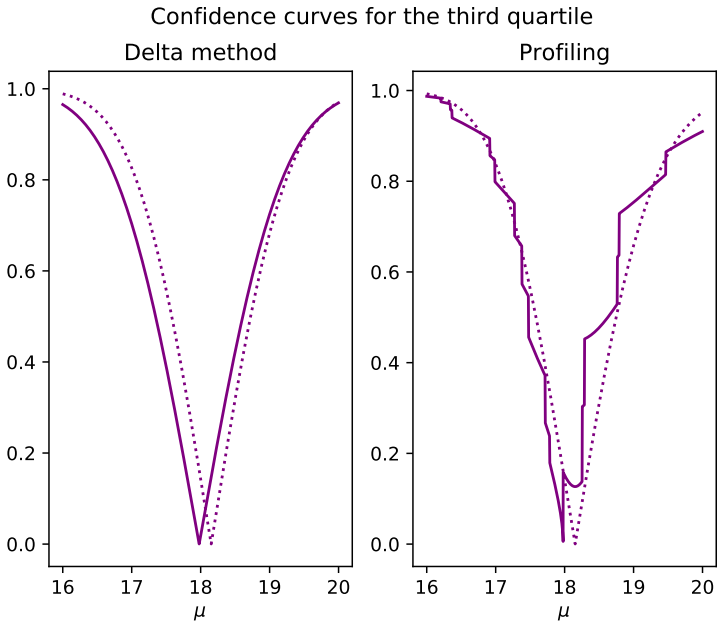


Figure 7.1: Confidence curves for the third quartile. The full drawn lines are based on hybrid likelihood, while the dotted graphs are confidence curves obtained with traditional maximum likelihood results.

likelihood estimators is therefore to be expected. Looking at Figure 7.1, we also notice that the curve constructed using maximum hybrid likelihood is shifted slightly to the left compared to the one obtained using maximum likelihood theory. As we saw previously in this example, the maximum likelihood estimate of the control parameter overshoots the true value. The confidence curve obtained using maximum hybrid likelihood is shifted to the left as a result of the increased robustness.

Theorem 7.4.2 can also be used to construct confidence intervals and curves for the control parameter. By this result,

$$\kappa D_n(\mu_0) \xrightarrow{d} \chi^2_1 \tag{7.14}$$

at the true value, μ_0, of μ. Here, as in the theorem,

$$\kappa = \frac{\nabla g(\theta_0)^T (J^*)^{-1} \nabla g(\theta_0)}{\nabla g(\theta_0)^T (J^*)^{-1} K^* (J^*)^{-1} \nabla g(\theta_0)}$$

and

$$D_n(\mu) = 2\Big(h_n(\widehat{\theta}_{hl}) - h_{n,prof}(\mu)\Big).$$

As before, κ can be estimated by plugging in the the canonical estimates for J^*, K^* and $\nabla g(\theta_0)$. Furthermore, closer inspection of D_n shows

$$D_n(\mu) = 2\left(h_n(\widehat{\theta}_{hl}) - \sup_{\lambda(\log 4)^{1/k}=\mu} h_n(\lambda, k)\right)$$

$$= 2\Big[h_n(\widehat{\theta}_{hl}) - \sup_k h_n\Big(\mu(\log 4)^{-1/k}, k\Big)\Big].$$

This quantity can be computed by minimizing

$$x \mapsto 2\Big[h_n(\widehat{\theta}_{hl}) - h_n\Big(\mu(\log 4)^{-1/x}, x\Big)\Big]$$

for each fixed μ in a grid of relevant values.

To compute a confidence curve, we utilize (7.14). Since

$$\kappa D_n(\mu_0) \overset{d}{\approx} \chi^2_1,$$

Slutsky's theorem ensures

$$\widehat{\kappa} D_n(\mu_0) \overset{d}{\approx} \chi^2_1, \tag{7.15}$$

where $\widehat{\kappa}$ is the consistent estimate of κ described in the previous paragraph. Using the approximation in (7.15), one can show that

$$cc_2(\mu) = \Gamma_1[\hat{\kappa} D_n(\mu)]$$

is an approximate confidence curve for the third quartile. This can be seen by arguing as in Section 4.1 or section 3.4 of Schweder and Hjort 2016. This confidence curve is displayed in Figure 7.1 together with the one obtained using maximum likelihood and Wilks theorem.

Inspecting Figure 7.1, we see the same pattern in the right as in the left figure. The confidence curve constructed using hybrid likelihood theory is wider and shifted to the right compared to the curve based on maximum likelihood theory. As before, this is a result of the added robustness and decreased efficiency when using maximum hybrid likelihood rather than maximum likelihood. In addition, we notice that the confidence curve constructed using Theorem 7.4.2 is quite jagged. This is a consequence of the discontinuity in

$$m[y, \mu(\lambda, k)] = I[y \leq \mu(\lambda, k)] - 0.75,$$

making h_n, and therefore also $h_{n,prof}$, a discontinuous function.

In this example, the focus and control parameter were the same. This will often be the case, as both control and focus parameters are quantities whose estimates are of extra importance. That being said, choosing the same quantity as both control and focus parameter is by no means a requirement. Confidence intervals and curves can be constructed for any $\psi = g(\lambda, k)$ as long as g is sufficiently smooth. In Section 7.5, we will revisit the example from Section 4.3 concerning battle deaths. Then the control and focus parameters will, indeed, be different.

Correlation in a bivariate normal distribution

In the previous example, we worked with random variables in $\mathbb{R}$. The theorems and result from this section do, however, also apply for data points in $\mathbb{R}^d$. To illustrate this, we will use the hybrid likelihood machinery to make inference about the correlation in a bivariate normal distribution.

We will work with n i.i.d. random variables, $Y_1, \ldots, Y_n \in \mathbb{R}^2$, following a bivariate normal distribution with mean $\mu_0 = (0, 0)$. We will in this example treat μ_0 as known and work only with the parameters $\theta = (\sigma_1, \sigma_2, \rho)$. Since $Y_1, \ldots, Y_n$ are independent and follow a bivariate normal distribution, the log-likelihood function takes the form

$$\ell_n(\sigma_1, \sigma_2, \rho) = \\ -n \log 2\pi - \frac{n}{2} \log|\Sigma(\sigma_1, \sigma_2, \rho)| - \frac{1}{2} \sum_{i=1}^{n} Y_i^T \Sigma(\sigma_1, \sigma_2, \rho)^{-1} Y_i,$$

where

$$\Sigma(\sigma_1, \sigma_2, \rho) = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \tag{7.16}$$

and $|\Sigma|$ denotes the determinant of the variance matrix.

To construct the hybrid likelihood function, we need to decide on a control and balance parameter. In this example, we will use

$$\mu = \begin{pmatrix} \Pr(Y_{i,1} > 1) \\ \Pr(Y_{i,2} \leq -1) \end{pmatrix}$$

for the first and $a = 0.5$ for the latter. These parameters are set somewhat arbitrary as the goal of this example is to illustrate how the theory can be used for multidimensional data rather than conducting any advanced analysis.

The following estimating equation identifies $\mu = (\mu_1, \mu_2)$:

$$m(y_1, y_2, \theta) = (I(y_1 > 1) - \mu_1, I(y_2 \leq -1) - \mu_2)^T.$$

Furthermore, the control parameter is given by

$$\mu(\sigma_1, \sigma_2, \rho) = (p_1(\sigma_1), p_2(\sigma_2))^T$$

in a central bivariate normal distribution with variance matrix $\Sigma(\sigma_1, \sigma_2, \rho)$ as given in (7.16). Here $p_j(\sigma_j)$ for $j = 1, 2$ is the probability that a $N(0, \sigma_j^2)$ distributed variable lands in the interval $(1, \infty)$ or $(-\infty, -1]$ respectively. With these functions, we can construct the empirical likelihood function, used in definition of the hybrid likelihood function, in the following way:

$$\mathrm{EL}_n[\mu(\theta)] = \max \left\{ \prod_{i=1}^{n} n w_i \;\middle|\; \sum_{i=1}^{n} w_i = 0, \sum_{i=1}^{n} w_i m[Y_i, \mu(\theta)] = 0, \, w_i \geq 0 \right\}.$$

Combining this with the log-likelihood function, allows us to construct the hybrid log-likelihood function,

$$h_n(\theta) = (1 - a)\ell_n(\theta) + a \log \mathrm{EL}_n[\mu(\theta)]. \tag{7.17}$$

Inference about θ, and functions of this parameter, can now be made using the results from this chapter. To illustrate how this can be done, we will use Theorem 7.4.2 to make inference about ρ. The profile hybrid log-likelihood function, $h_{n,prof}(\rho)$, can be computed by maximizing the function $x \mapsto h_n(x, \rho)$

for each fixed ρ. With $\widehat{\theta}_{hl}$ denoting the maximizer of h_n, Theorem 7.4.2 guarantees that

$$2\kappa\Big(h_n(\widehat{\theta}_{hl}) - h_{n,prof}(\rho)\Big) \overset{d}{\approx} \chi^2_1,$$

where κ is as defined in the theorem. The value of κ is unknown, but can be estimated consistently by

$$\widehat{\kappa} = \frac{(0,0,1)(\widehat{J}^*)^{-1}(0,0,1)^T}{(0,0,1)(\widehat{J}^*)^{-1}\widehat{K}^*(\widehat{J}^*)^{-1}(0,0,1)^T},$$

where $\widehat{J}^*$ and $\widehat{K}^*$ are the canonical estimators of J^* and K^*. Hence, by Slutsky's theorem,

$$2\widehat{\kappa}\Big(h_n(\widehat{\theta}_{hl}) - h_{n,prof}(\rho)\Big) \overset{d}{\approx} \chi^2_1.$$

This approximation can be used to construct approximate confidence intervals and curves for ρ. The latter is given by

$$cc(\rho) = \Gamma_1\Big[2\widehat{\kappa}\Big(h_n(\widehat{\theta}_{hl}) - h_{n,prof}(\rho)\Big)\Big].$$

We computed the quantities as described above for a simulated data set of $n = 100$ i.i.d. random variables,following a central bivariate normal distribution with variance matrix,

$$\Sigma_0 = \begin{pmatrix} 0.5^2 & 0.5 \cdot 0.5 \cdot 0.5 \\ 0.5 \cdot 0.5 \cdot 0.5 & 0.5^2 \end{pmatrix} = \begin{pmatrix} 0.25 & 0.125 \\ 0.125 & 0.25 \end{pmatrix}.$$

In other words, both components have mean 0 and standard deviation 0.5, and the correlation between them is also 0.5. A plot of $cc(\rho)$ for this data set can be found in Figure 7.2. In addition, we computed the profile log-likelihood function and added the confidence curve based on this and Wilks theorem to the plot.

From Figure 7.2, we see that the confidence curve based on hybrid likelihood is slightly wider than the one based on maximum likelihood. This is a consequence of the decreased efficiency of maximum hybrid likelihood estimators compared to those maximum likelihood estimators.

Revisiting the deadly example

We will now try out the hybrid likelihood machinery on the Correlations of War data set presented in Section 4.3. Our goal will be to compare the median number of deaths in conflicts before and after the Korean war, and compare our results to the ones obtained using empirical and parametric likelihood. As before we will let $X_1, \ldots, X_{n_1}$ denote the number of deaths in each conflict before, and including, the Korean war, and $Y_1, \ldots, Y_{n_2}$ be the same for after this conflict. Furthermore, we will refer to the true medians in the distributions as ν_1 and ν_2 respectively.

Before we start with the mathematics, we need to decide what family to model the distribution of the data as. In Cunen, Hjort, and Nygård 2020, it

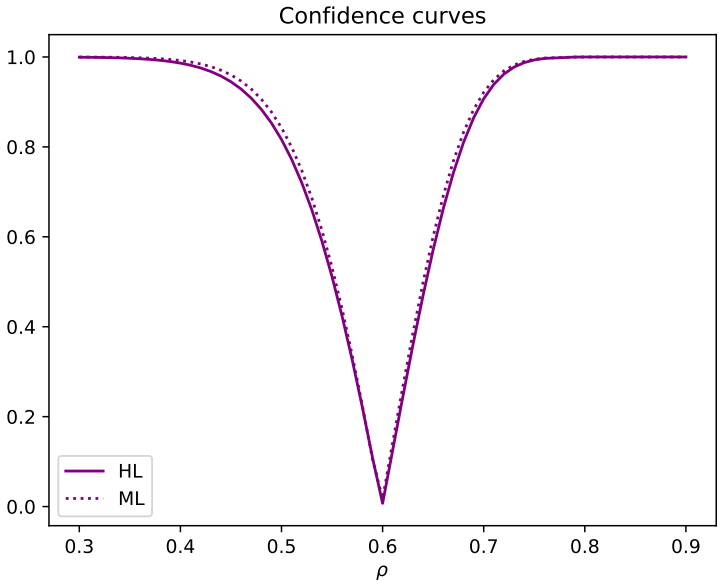


Figure 7.2: Confidence curves for ρ. The full drawn line is obtained with hybrid likelihood and Theorem 7.4.2, while the dotted curve is constructed with maximum likelihood and Wilks theorem.

is argued that, after subtraction of the smallest data point (1001), the inverse Burr distribution is a good fit for these data sets. This model will therefore be used. The inverse Burr distribution has three parameters, θ, α, $b > 0$, and cumulative distribution function

$$F_{\theta,\alpha,b}(y) = \left(1 + e^{b\alpha} y^{-\alpha}\right)^{-\theta}. \tag{7.18}$$

The density function in this distribution is

$$f_{\theta,\alpha,b}(y) = \alpha\theta e^{b\alpha} y^{-(\alpha+1)} \left(1 + e^{b\alpha} y^{-\alpha}\right)^{-(\theta+1)},$$

for y, θ and α positive and zero otherwise. Zero is not in the support of this density, so, just as in Cunen, Hjort, and Nygård 2020, we will replace the smallest data point with 1001.01 to avoid problems with the smallest observation. Notice also that we work with the logarithm of the scale parameter rather than this quantity itself. This is done to ensure numerical stability of the optimization algorithms.

The maximum likelihood estimates for $(\theta, \alpha, b)^T$ in the two distributions can be found by numerically maximizing the log likelihood functions, $\ell_{n_1,1}$ and $\ell_{n_2,2}$. Here

$$\ell_{n_1,1}(\theta, \alpha, b) = \sum_{i=1}^{n_1} \log f_{\theta,\alpha,b}(X_i)$$

and

$$\ell_{n_2,2}(\theta,\alpha,b) = \sum_{i=1}^{n_2} \log f_{\theta,\alpha,b}(Y_i),$$

for $f_{\theta,\alpha,b}$ defined as above. The estimates we obtain are $(\hat{\theta}_1, \hat{\alpha}_1, \hat{b}_1) = (0.416, 0.773, 11.030)$ for the data set corresponding to conflicts before the Korean War and $(\hat{\theta}_2, \hat{\alpha}_2, \hat{b}_2) = (0.619, 0.931, 8\,174.403)(0.619, 0.931, 9.009)$ for the data set with more recent conflicts. From maximum likelihood theory we know

$$\sqrt{n}\left[\begin{pmatrix}\hat{\theta}_j\\ \hat{\alpha}_j\\ \hat{b}_j\end{pmatrix} - \begin{pmatrix}\theta_{0,j}\\ \alpha_{0,j}\\ b_{0,j}\end{pmatrix}\right] \xrightarrow{d} \mathrm{N}\left(0, J_j^{-1}\right) \tag{7.19}$$

for $j = 1, 2$. Here $(\theta_{0j}, \alpha_{0j}, b_{0j})^T$ and J_j for $j = 1, 2$ are the true parameter vectors and Fisher matrices in the respective distributions.

We are interested in the median number of casualties. In an inverse Burr distribution with parameters (θ, α, b), the median is given by

$$e^b \left(2^{1/\theta} - 1\right)^{-1/\alpha}$$

So with

$$g(\theta,\alpha,b) = e^b \left(2^{1/\theta} - 1\right)^{-1/\alpha} + 1001, \tag{7.20}$$

we are interested in making inference about $\nu_j = g(\theta_{0,j}, \alpha_{0,j}, b_{0,j})$. 1001 is added to correct for the subtraction of the smallest data point done in the beginning of this example.

Using the delta method and (7.19), we get the following limit result

$$\begin{aligned}\sqrt{n}\left(g(\hat{\theta}_j,\hat{\alpha}_j,\hat{b}_j) - \nu_j\right) &= \sqrt{n}\left(g(\hat{\theta}_j,\hat{\alpha}_j,\hat{b}_j) - g\left(\theta_{0,j},\alpha_{0,j},b_{0,j}\right)\right)\\ &\overset{d}{\approx} \nabla g\left(\theta_{0,j},\alpha_{0,j},b_{0,j}\right)^T \mathrm{N}\left(0, J^{-1}\right)\\ &\sim \mathrm{N}\left(0, \nabla g\left(\theta_{0,j},\alpha_{0,j},b_{0,j}\right)^T J^{-1} \nabla g\left(\theta_{0,j},\alpha_{0,j},b_{0,j}\right)\right),\end{aligned}$$

for $j = 1, 2$. Therefore

$$g(\hat{\theta}_j,\hat{\alpha}_j,\hat{b}_j) \overset{d}{\approx} \mathrm{N}\left(\nu_j, \frac{\nabla g\left(\theta_{0,j},\alpha_{0,j},b_{0,j}\right)^T J^{-1} \nabla g\left(\theta_{0,j},\alpha_{0,j},b_{0,j}\right)}{n}\right)$$

for $j = 1, 2$. As the maximum likelihood estimator is consistent for the true value and ∇g is continuous, $\nabla g(\theta_{0j}, \alpha_{0j}, b_{0j})$ can be approximated by $\nabla g(\hat{\theta}_j, \hat{\alpha}_j, \hat{b}_j)$. The matrices J_j^{-1}/n, for $j = 1, 2$, can be estimated consistently by the the negative inverse of $H\ell_{n,j}$ evaluated at the point $(\hat{\theta}_j, \hat{\alpha}_j, \hat{b}_j)$. Hence, for $j = 1, 2$,

$$g(\hat{\theta}_j,\hat{\alpha}_j,\hat{b}_j) \overset{d}{\approx} \mathrm{N}\left(\nu_j, \nabla g(\hat{\theta}_j,\hat{\alpha}_j,\hat{b}_j)^T \left(-H\ell_{n,j}(\hat{\theta}_j,\hat{\alpha}_j,\hat{b}_j)\right)^{-1} g(\hat{\theta}_j,\hat{\alpha}_j,\hat{b}_j)\right).$$

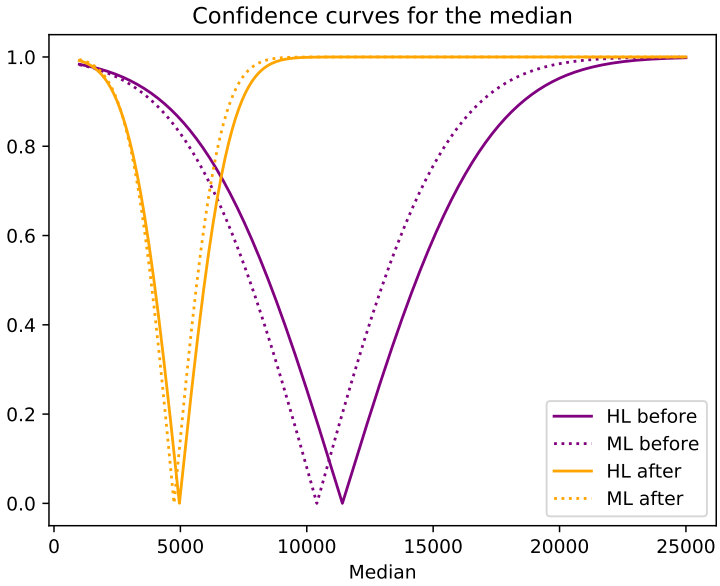


Figure 7.3: Confidence curves for the medians number of battle deaths before and after the Korean war. We have used purple for the older conflicts and orange for the more recent ones. The dotted graphs are confidence curves based on maximum likelihood, while the full drawn lines are obtained using hybrid likelihood.

Using this approximation we can construct confidence intervals and curves for the median in the two populations. A plot of the curves together with the ones based on hybrid likelihood can be found in Figure 7.3.

Before fitting the parametric models using hybrid likelihood, we will examine how well the inverse Burr distribution fit the data. In Figure 7.4 we have displayed a QQ-plot. From the figure, we notice that despite working well for the smaller observations, the fit is worse for the larger data points. The reason for this is that both data sets have many data points that are quite small (between 1001 and 2000). In the maximum likelihood fit, these will dominate, ensuring a good fit in this region at the cost of a more lacking fit in the upper part. This "pulls" the weight of the distribution towards the smaller values and results in underestimation of the median. To get more robust estimates of the parameter of interest, we will use using maximum hybrid likelihood with control parameter $\mu_j = \Pr(Y \geq q_j)$ for $j = 1, 2$ to fit inverse-Burr models. μ_j is the probability of observing a value greater than q_j so this will ensure that the expected number of data points greater than this q_j is not underestimated, hence "pulling" the weight of the distribution towards the larger values. For each j we will use q_j equal to the empirical 0.25 quantile. By (7.18),

$$\mu(\theta, \alpha, b) = 1 - \left(1 + e^{b\alpha} q_j^{-\alpha}\right)^{-\theta}.$$

Furthermore $\mathrm{E}[I(Y > q_j) - \mu_j] = 0$ for the true value of μ_j, so $m(y, \mu) = I(y > q_j) - \mu$ for $j = 1, 2$ will be our estimating functions.

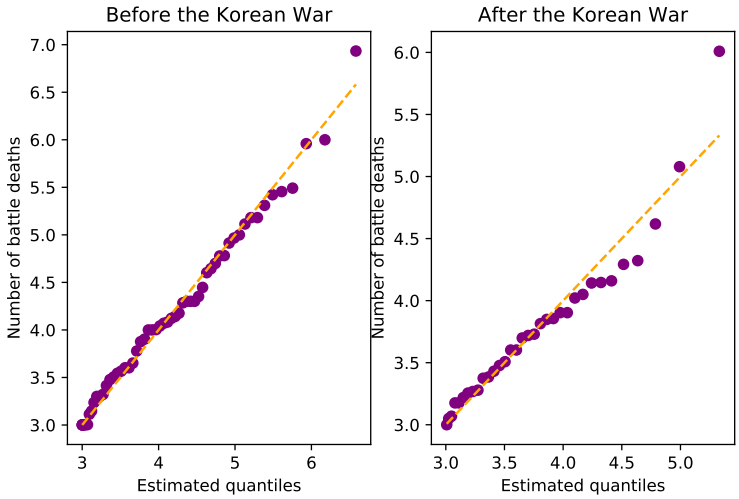


Figure 7.4: QQ-plots of the ordered number of battle deaths plotted against the maximum likelihood estimate of the theoretical quantiles in the distribution. We have plotted the values on $\log_{10}$ scale to make the figures more readable.

We have now decided on the control parameters, μ_j, but before we can construct the hybrid likelihood functions, we need to choose the balance parameters, a_j. We believe that the model is correct and that the data does follow an inverse Burr distribution. If the model is specified correctly, maximum likelihood estimates are asymptotically optimal in terms of mean squared error, under sufficient regularity. So to choose the tuning parameters a_j, we will find values such that the loss of efficiency is not too great when compared to the maximum likelihood estimates. In this example we will use the median as the focus parameter and choose a such that the standard deviation of the maximum hybrid likelihood estimate of the median is no more than 10% larger than that of the maximum likelihood estimate. This corresponds to requiring confidence intervals of the median, based on Corollary 7.2.2 and the delta method to be stretched no more than 10% in each direction. This is the procedure described in Section 7.3, and after implementation, we found that $a = 0.39$ gave the desired trade-off for the first and $a = 0.38$ for the second data set.

We are now ready to fit the parametric family to the data sets using maximum hybrid likelihood. With the control parameters and values of a found above, the resulting hybrid likelihood functions can be constructed and numerically maximized. For our data sets, the maximum hybrid likelihood estimates we computed to be $(\hat{\theta}_{hl,1}, \hat{\alpha}_{hl,1}, \hat{b}_{hl,1}) = (0.422, 0.783\,11.072)$ and $(\hat{\theta}_{hl,2}, \hat{\alpha}_{hl,2}, \hat{b}_{hl,2}) = (0.626, 0.939, 9.036)$, for the first and second data set respectively. With these parameters, the median number of battle deaths is estimated to be 11409 before the Korean war and 4961 afterwards. These values are slightly closer to the empirical estimates (11375 and 5240 respectively) than the corresponding numbers obtained with maximum likelihood, 10399 and 4749 respectively.

Using Corollary 7.2.2 and the delta method we can construct confidence

intervals and curves for ν_1 and ν_2. The arguments are similar to the ones presented previously in this example in the context of maximum likelihood theory. Details are therefore omitted, but a plot of the resulting confidence curves can be found in Figure 7.3.

We now turn our attention to the focus parameter

$$\psi_0 = \frac{\nu_1}{\nu_2} = f\left(\theta_{0,1}, \alpha_{0,1}, b_{0,1}, \theta_{0,2}, \alpha_{0,2}, b_{0,2}\right).$$

Where f is the function

$$f\left(\theta_1, \alpha_1, b_1, \theta_2, \alpha_2, b_2\right) = \frac{g\left(\theta_1, \alpha_1, b_1\right)}{g\left(\theta_2, \alpha_2, b_2\right)},$$

with g as in (7.20).

Since the X_is and Y_js are assumed to be independent, we construct the hybrid log-likelihood function for the joint sample in the following way:

$$h_{n_1,n_2}\left(\theta_1, \alpha_1, b_1, \theta_2, \alpha_2, b_2\right) = h_{n_1}\left(\theta_{0,1}, \alpha_{0,1}, b_{0,1}\right) + h_{n_2}\left(\theta_{0,2}, \alpha_{0,2}, b_{0,2}\right).$$

Here h_{n_j} is the hybrid log-likelihood function based on the j^{th} data set, constructed with the balance and control parameters described previously in this section. Similarly, as in Section 4.3, a profiling result holds for the joint hybrid likelihood function as well. We will give a sketch of a proof here, but a full argument will not be provided.

Let η_j denote the vector $(\theta_{0,j}, \alpha_{0,j}, b_{0,j})$ for $j = 1, 2$ and $\eta = (\eta_1, \eta_2)$. Since the X_is and Y_js are independent, Theorem 7.2.1 ensures the process

$$\begin{aligned} A_n(s) &= h_{n_1,n_2}\left(\eta + \frac{s}{\sqrt{n}}\right) - h_{n_1,n_2}(\eta) \\ &= h_{n_1}\left(\eta_1 + \frac{s_1}{\sqrt{n}}\right) - h_{n_1}(\eta_1) + h_{n_2}\left(\eta_2 + \frac{s_2}{\sqrt{n}}\right) - h_{n_2}(\eta_2), \end{aligned}$$

where $s = (s_1, s_2)$, converges as a process to

$$A(s) = A_1(s_1) + A_2(s_2)$$

with

$$A_j(s_j) = s_j^T U_j^* - \frac{1}{2} s_j^T J_j^* s_j \quad \text{for } j = 1, 2,$$

and U_1^* and U_2^* independent. Some algebra shows

$$\begin{aligned} A(s) &= s_1^T U_1^* - \frac{1}{2} s_1 J_1^* s_1 + s_2^T U_2^* - \frac{1}{2} s_2 J_2^* s_2 \\ &= s^T U^* - \frac{1}{2} s^T J^* s \end{aligned}$$

with

$$U^* = \begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix} \quad \text{and} \quad J^* = \begin{pmatrix} J_1^* & 0 \\ 0 & J_2^* \end{pmatrix}.$$

With this, and arguments similar to those given in the proof of Theorem 7.4.2, one can show that a profiling result holds also for h_{n_1,n_2} with U^* and J^* defined as above and

$$K^* = \operatorname{Var} U^* = \begin{pmatrix} K_1^* & 0 \\ 0 & K_2^* \end{pmatrix}.$$

Here J_j^*, K_j^* and U_j^* are of course the quantities defined in (7.6) corresponding to data set j.

Both J^* and K^* are unknown, but replacing them with their canonical estimates, gives us a consistent estimator, $\hat{\kappa}$, of κ. Hence, $\hat{\kappa} D_n(\psi_0) \xrightarrow{d} \chi_1^2$ by Slutsky's theorem and Theorem 7.4.2. We can now construct the corresponding approximate confidence curve,

$$cc(\psi) = \Gamma_1[\hat{\kappa} D_n(\psi)],$$

where Γ_1 is the cumulative distribution function in a χ_1^2-distribution. A plot of this curve can be found in Figure 7.5 together with the one obtained with traditional maximum likelihood theory. The latter was constructed in a similar fashion using the deviance function and Wilks theorem.

From Figure 7.5, we see that the confidence curve based on profile hybrid likelihood is slightly wider than the one based on profile likelihood. As hybrid likelihood methods are asymptotically less efficient than the corresponding maximum likelihood versions, this is not surprising. Furthermore, the curve corresponding to hybrid likelihood is shifted to the right compared to its maximum likelihood counterpart. This is a consequence of the added robustness of the estimate of the control parameter.

We can use the confidence curves to find confidence intervals for ψ and p-values for testing hypothesis on the form $\psi_0 = \psi$. For the profile maximum likelihood function we find that [0.99, 4.79] and [0.85, 5.57] are 90%- and 95%-confidence intervals respectively. We are therefore able to conclude that $\psi_0 > 1$ on neither a 5% nor 10% level with using this method.

The results are a little more positive for the hybrid likelihood function. Using the more robust method, we find that [1.06, 5.01] is a 90% confidence interval. Increasing the level to 95% results in the interval [0.91, 5.83]. So, we are able to reject the hypothesis $\psi_0 = 1$ on a 10% but not on a 5% level. This is a stronger result than we got using the profile empirical likelihood function in Section 4.3. Since knowing the distribution of the data adds strength to the estimation procedure, this does not come as a surprise.

Lastly we want to comment on how the result compares to the findings of Cunen, Hjort, and Nygård 2020. The authors of this article also compute the p-value for testing $H_0 \colon \psi_0 = 1$ against the alternative $\psi > 1$, albeit with another method than ours. They find a p-value slightly bigger than 5. This is similar to what we get, a p-value of 7.9%.

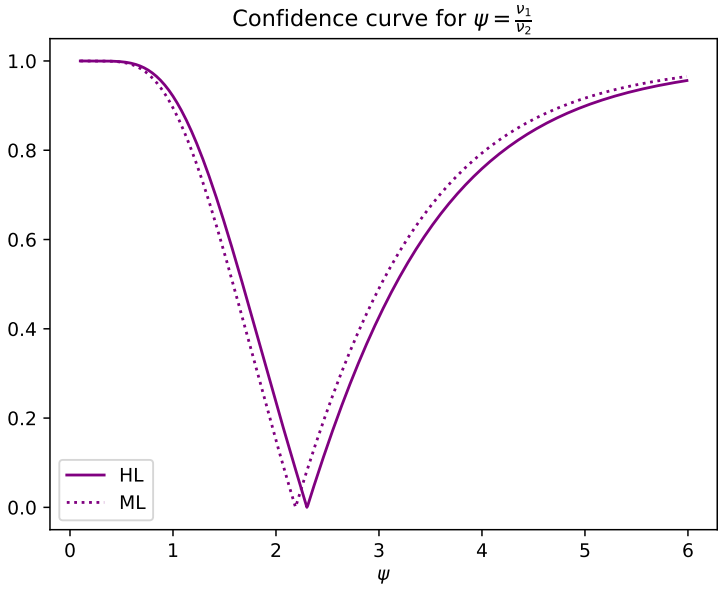


Figure 7.5: Confidence curves for $\psi = \frac{\nu_1}{\nu_2}$. The full drawn line is based on the approximate χ^2_1-distribution of the scaled profile hybrid likelihood function. The dotted line is constructed using Wilks theorem for the profile likelihood function.

CHAPTER 8

Outside model conditions

The results from the previous chapters are elegant and easy to use. They are, however, based on one the assumption that the model is specified correctly. Hybrid likelihood was developed as a way to make the estimator of the control parameter robust against model misspecification. Assuming that the model is correctly specified is therefore a little counter-intuitive, as it undermines the need for model robust estimates of the control parameter. In this chapter, we will discard the problematic assumption and take a closer look at what happens outside of model conditions. For this, the results derived in Chapter 5 and Chapter 6 will be essential.

We will start by finding what value the maximizer of the hybrid likelihood function is aiming for and prove consistency of $\widehat{\theta}_{hl}$ towards it. This will be the topic of Section 8.1. Afterwards, we will turn our attention to the variable

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big). \tag{8.1}$$

In Section 8.2, we will show that (8.1) has a normal limit, also when the model is specified incorrectly. We will also propose some consistent estimators of the matrices involved in the limit distribution of (8.1). This will be done in Section 8.3. In Section 8.4, we will see that the results of this chapter reduces to those of Chapter 7 when the model is, indeed, specified correctly.

As before, we will let $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables. In this chapter, however, we will assume their true distribution, with cumulative distribution function F, is not necessarily a member of the parametric family, f_θ for $\theta \in \Theta$, fit to the data. Y will always be a general random variable following the true distribution, a the balance parameter and $\mu \in \mathbb{R}^r$ the control parameter. We will use $\mu(\theta)$ to refer to the corresponding value of μ in the distribution f_θ. $m\colon \mathbb{R}^{d+r} \to \mathbb{R}^q$ will denote the estimating function used in construction of the empirical likelihood function. As in Chapter 6 we will assume $q = 1$, but generalizations to higher dimensions should be possible with sufficient mathematical efforts. Furthermore, h_n and $\widehat{\theta}_{hl}$ will always refer to the log-hybrid likelihood function and its maximizer respectively, and ℓ_n and EL_n to the log-likelihood and empirical likelihood functions. We will use $M_i(\theta)$ as short-hand for $m(Y_i, \theta)$, and $M(\theta)$ for $m(Y, \theta)$. Lastly, $\lambda_n(\theta)$ will denote the unique solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{M_i(\theta)}{1 + \lambda M_i(\theta)} = 0,$$

with $1 + \lambda M_i(\theta) > 0$ for $i = 1, \ldots, n$ and $\lambda(\theta)$ equal to the unique root of

$$\begin{pmatrix} \mathrm{E}\{M(\theta)/[1 + \lambda M(\theta)]\} \\ \Pr[1 + \lambda M(\theta) \leq 0] \end{pmatrix}.$$

8.1 Consistency

In the previous chapter, we assumed the true underlying distribution was a member of the parametric family. In this case, there exists a "true" parameter, θ_0, such that the underlying density is equal to f_{θ_0}. Furthermore, Corollary 7.2.2 guarantees the maximizer of the hybrid likelihood function is consistent for this value. In this chapter, however, we do not assume the true underlying distribution is a member of the parametric family. Because of this, a "true" parameter does not exist. This leaves us with some very important questions: What does the maximizer of the hybrid likelihood function aim for when the model is wrongly specified? Is $\widehat{\theta}_{hl}$ consistent for any value, and in the case that it is, what interpretation does such a value have? In this section, we will use the results of Section 6.2 to address these question

Fix some $\theta \in \Theta$ such that $\mu(\theta)$ satisfies the conditions of Theorem 5.3.1. Then

$$\frac{1}{n} \log \mathrm{EL}_n[\mu(\theta)] \xrightarrow{\Pr} -\mathrm{E} \log[1 + \lambda(\theta) M(\theta)]$$

by the same result. Furthermore,

$$\frac{1}{n} \ell_n(\theta) \xrightarrow{\Pr} \mathrm{E} \log\, f_\theta(Y)$$

provided the right hand side exists. Hence, for all $\theta \in \Theta$ satisfying the conditions above,

$$\frac{1}{n} h_n(\theta) \xrightarrow{\Pr} (1 - a)\mathrm{E} \log\, f_\theta(Y) - a\mathrm{E} \log[1 + \lambda(\theta) M(\theta)]. \tag{8.2}$$

Because of the convergence in (8.2), we would expect the maximizer of h_n to be consistent for the maximizer of the limit. Under the sufficient conditions, this does indeed hold true. This will be shown in Theorem 8.1.1. To simplify notation, we will introduce the random function $\Gamma_n \colon \Theta \to \mathbb{R}$ defined as

$$\Gamma_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{(1 - a) \log f_\theta(Y_i) - a \log[1 + \lambda(\theta) M_i(\theta)]\},$$

as well as the non-random population version

$$\Gamma(\theta) = (1 - a)\mathrm{E} \log f_\theta(Y) - a\mathrm{E} \log[1 + \lambda(\theta) M(\theta)]. \tag{8.3}$$

Theorem 8.1.1. *Let Θ_0 be a compact subset of Θ such that the conditions of Lemma 6.1.1 and Lemma 6.1.2 hold for the image of Θ_0 under μ. Let $\widehat{\theta}_{hl}$ and θ_0 denote the maximizer of h_n and Γ respectively in this set. Then*

$$\widehat{\theta}_{hl} \xrightarrow{\Pr} \theta_0,$$

provided μ *and* $\log\, f_\theta(y)$*, for almost all* y*, are continuous as a functions of* θ*,* θ_0 *is the unique maximizer of* Γ *and there exists a function* p_3 *such that*

$$|\log\, f_\theta(y)| \leq p_3(y) \tag{8.4}$$

for all y *in the support of* Y *with* $\mathrm{E}\, p_3(Y) < \infty$*.*

Proof. We will use theorem 5.7 in Vaart 1998, p. 45.

By the proof of Theorem 6.2.1,

$$\mu \mapsto \mathrm{E}\, \log[1 + \lambda(\mu)M(\mu)]$$

is continuous. Furthermore, μ is continuous by assumption, and condition (8.4) can be used to show continuity of

$$\theta \mapsto \mathrm{E} \log\, f_\theta(Y),$$

in a similar way as we did for Φ in the proof of Lemma 6.1.1. Because of this, Γ is a continuous function. Furthermore, Θ_0 is compact and θ_0 the unique maximizer of Γ, so

$$\sup_{\theta \in \Theta_0} \{\, \Gamma(\theta) \mid \|\theta_0 - \theta\| \geq \epsilon \,\} < \Gamma(\theta_0)$$

for all $\epsilon > 0$. In addition,

$$\frac{1}{n} h_n(\widehat{\theta}_{hl}) \geq \frac{1}{n} h_n(\theta_0)$$

by definition of $\widehat{\theta}_{hl}$. The only thing remaining to check is

$$\sup_{\theta \in \Theta_0} \left| \frac{1}{n} h_n(\theta) - \Gamma(\theta) \right| \xrightarrow{\mathrm{Pr}} 0.$$

By the triangle inequality

$$\begin{aligned} \sup_{\theta \in \Theta_0} \left| \frac{1}{n} h_n(\theta) - \Gamma(\theta) \right| \leq & (1-a) \sup_{\theta \in \Theta_0} \left| \frac{1}{n} \sum_{i=1}^{n} \log\, f_\theta(Y_i) - \mathrm{E} \log\, f_\theta(Y) \right| + \\ & + a \sup_{\theta \in \Theta_0} |\log \mathrm{EL}_n[\mu(\theta)] - \mathrm{E}\, \log[1 + \lambda(\theta)M(\theta)]| \end{aligned}$$

The second term goes in probability to 0 by the proof of Theorem 6.2.1, so if we can show

$$\sup_{\theta \in \Theta_0} \left| \frac{1}{n} \sum_{i=1}^{n} \log\, f_\theta(Y_i) - \mathrm{E} \log\, f_\theta(Y) \right| \xrightarrow{\mathrm{Pr}} 0 \tag{8.5}$$

the argument is complete. By the condition (8.4), compactness of Θ_0 and continuity of $\theta \mapsto \log\, f_\theta(y)$, the uniform law of large numbers can be applied to ensure (8.5). This concludes the proof. ■

The conditions stated in the previous theorem are sufficient, but not necessary for consistency of $\widehat{\theta}_{hl}$ towards θ_0. Assumptions like continuity of

$\theta \mapsto m(y, \mu(\theta))$ and compactness of the parameter space are convenient to ensure

$$\sup_{\theta \in \Theta_0} \left| \frac{1}{n} h_n(\theta) - \Gamma(\theta) \right| \xrightarrow{\Pr} 0,$$

but this might hold true even when the conditions of Theorem 8.1.1 do not. One example of this is the case of $m[y, \mu(\theta)] = I[y \leq \mu(\theta)] - q$ for $q \in (0, 1)$. This function is not continuous in θ. Because of this, the previous result cannot be applied. Nevertheless, the maximum hybrid likelihood estimator is still consistent for the minimizer of Γ. This is proved in the ensuing theorem.

Theorem 8.1.2. *Let Θ_0 be a compact subset of Θ and $\widehat{\theta}_{hl}$ a maximizer of $h_n(\theta)$ in this set, where the hybrid likelihood function is constructed with the estimating function*

$$m(y, \mu) = I(y \leq \mu) - q,$$

for some $q \in (0, 1)$. Assume F, μ and $\theta \mapsto \log f_\theta(y)$ for almost all y are continuous functions, θ_0 is the unique maximizer of Γ in Θ_0 and that there exists a function p_3 such that

$$|\log f_\theta(y)| \leq p_3(y),$$

for all y in the support of Y, with $\mathrm{E}\, p_3(Y) < \infty$. Then

$$\widehat{\theta}_{hl} \xrightarrow{\Pr} \theta_0.$$

Proof. We have

$$\Gamma(\theta) = (1 - a)\mathrm{E} \log f_\theta(Y) - a\mathrm{E} \log(1 + \lambda(\theta)\{I[Y \leq \mu(\theta)] - q\}).$$

The first summand is continuous as a function of θ by the arguments in the proof of Theorem 8.1.1. Continuity of the second part was shown in Theorem 6.2.2. Hence, Γ is a continuous function. This, in combination with compactness of Θ_0 and θ_0 being the unique maximizer of Γ in Θ_0, ensures that

$$\sup_{\theta \in \Theta_0} \{ \Gamma(\theta) \mid \|\theta_0 - \theta\| \geq \epsilon \} < \Gamma(\theta_0)$$

holds for all $\epsilon > 0$. Furthermore,

$$\frac{1}{n} h_n(\widehat{\theta}_{hl}) \geq \frac{1}{n} h_n(\theta_0)$$

by definition of $\widehat{\theta}_{hl}$. In addition,

$$\begin{aligned}
&\sup_{\theta \in \Theta_0} \left| \frac{1}{n} h_n(\theta) - \Gamma(\theta) \right| \leq \\
&(1 - a) \sup_{\theta \in \Theta_0} \left| \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(Y_i) - \mathrm{E} \log f_\theta(Y) \right| + \\
&+ a \sup_{\theta \in \Theta_0} |\log \mathrm{EL}_n[\mu(\theta)] - \mathrm{E} \log(1 + \lambda(\theta)\{I[Y \leq \mu(\theta)] - q\})|.
\end{aligned}$$

The first term goes to 0 in probability by the proof of Theorem 8.1.1, and the second term by the arguments in Theorem 6.2.2. This concludes the proof. ■

In this section, we have used Theorem 5.7 in Vaart 1998 to prove consistency, but there are other ways to proceed. If $\theta \in \mathbb{R}$, Lemma 5.10 in the same source can be applied in more general settings than the ones listed here. Furthermore, if the hybrid likelihood function is concave, consistency can be shown using e.g. the results in Hjort and Pollard 1993. The general idea, however, is to use that the hybrid likelihood function converges pointwise in probability to Γ. Conditions for consistency of $\widehat{\theta}_{hl}$ are nothing but requirements for this convergence to be "quick enough" for the maximizers to converge as well. This might happen even when the conditions of Theorem 8.1.1 or Theorem 8.1.2 fail to hold.

The limit of the maximum hybrid likelihood estimate

By Theorem 8.1.1 and Theorem 8.1.2, the maximum hybrid likelihood estimator is consistent for the maximizer of

$$\Gamma(\theta) = (1-a)\mathrm{E}\,\log f_\theta(\,Y) - a\mathrm{E}\,\log[1+\lambda(\theta)M(\theta)].$$

To better understand what happens with the hybrid likelihood theory outside of model conditions, we need to investigate both Γ and its maximizer θ_0. In this subsection, we will assume the true underlying distribution is continuous with density f, but the analysis generalizes to discrete distributions in the natural way.

The maximizer of Γ is also the minimizer of $-\Gamma$. Furthermore, addition of a constant to a function does not change its argmin. Hence

$$\theta_0 = \underset{\theta\in\Theta}{\operatorname{argmin}}\ \{\,-\Gamma(\theta) + (1-a)\mathrm{E}\,\log f(\,Y)\,\} \tag{8.6}$$

$$= \underset{\theta\in\Theta}{\operatorname{argmin}}\ \{\,(1-a)\mathrm{KL}(f,f_\theta) + a\mathrm{E}\,\log[1+\lambda(\theta)M(\theta)]\,\}\,, \tag{8.7}$$

where

$$\mathrm{KL}(f,f_\theta) = \int_{\mathbb{R}^d} f(y)\log\frac{f(y)}{f_\theta(y)}\,dy$$

is the Kullback-Leibler divergence from f to f_θ. From (8.7), we notice that θ_0 is the minimizer of a convex combination of two quantities: the Kullback-Leibler divergence and

$$d_\mu(f,f_\theta) = \mathrm{E}\,\log[1+\lambda(\theta)M(\theta)].$$

The Kullback-Leibler divergence is a measure of how much two distributions differ, and, under mild conditions, the maximum likelihood estimator aims for the minimizer of this divergence. See e.g. appendix A.5 in Schweder and Hjort 2016 for details about the Kullback-Leibler divergence. The function $\theta \mapsto d_\mu(f,f_\theta)$ is the pointwise limit of $-n^{-1}\log\mathrm{EL}_n[\mu(\theta)]$ by Theorem 5.3.1. In the arguments preceding Theorem 6.2.1, we showed that this map is minimized when $\mu(\theta) = \mu_0$ and μ_0 is the true value of the control parameter. Hence, when minimizing the divergence

$$d_{a,\mu}(f,f_\theta) = (1-a)\mathrm{KL}(f,f_\theta) + a\,d_\mu(f,f_\theta),$$

we find a value of θ such that both the overall model fit, measured by the Kullback-Leibler divergence, and the quality of the estimate of the control parameter, measured by $d_\mu(f, f_\theta)$, are taken into account. The balance parameter decides how much weight is put on overall model fit and how much is put on estimation of the control parameter. This fits well with the original idea behind hybrid likelihood: We are willing to give up some amount of overall model fit in favor of robust estimation of the control parameter. How much we are willing to pass up on is expressed in terms of the balance parameter.

For small values of a, the Kullback-Leibler divergence will dominate $d_{a,\mu}$. Hence, the minimizer, θ_0, of $d_{a,\mu}$ will be close to what the maximum likelihood estimator is aiming for. Since the true distribution is not assumed to be a member of the parametric family, however, $\mu(\theta_0)$ may be far away from the true value of the control parameter. On the other hand, $d_{a,\mu}$ and d_μ will be very similar when the balance parameter is close to 1. Hence, $\mu(\theta_0)$ and μ_0 will be similar, but the parametric fit may be lacking in other aspects. For most values of a, however, minimization of $d_{a,\mu}$ will result in a density close to f, in terms of Kullback-Leibler divergence, whose value of the control parameter is not too far away from μ_0.

To get some intuition about how the choice of balance parameter affects what the hybrid likelihood estimator is aiming for, we have plotted the minimizer of $d_{a,\mu}$ for different values of a in different situations. Consider first the case where the true underlying distribution, f, is a Weibull distribution with shape parameter 2 and scale parameter 7 and the parametric family is the collection of all gamma distributions with shape-rate parametrization. In Figure 8.1 we have plotted the minimizer of $d_{a,\mu}$ with two different choices of control parameter. In the plot to the left, μ is the median in the distribution. For $a = 0$ the minimizer is $(3.136, 0.506)$, which is the minimizer of the Kullback-Leibler divergence from the parametric family to the true distribution. As $a \to 1$, the minimizer moves towards $(3.206, 0.494)$. The median in a gamma distribution with shape parameter 3.206 and rate parameter 0.494 is approximately 5.827, which is, indeed, the median in a Weibull(2,7) distribution. So, as a increases from 0 to 1, the minimizer $d_{a,\mu}$ moves from the minimizer of the Kullback-Leibler divergence towards the value of θ resulting in $\mu(\theta) = \mu_0$. The plot on the right-hand side in Figure 8.1 is a similar parametrized curve, but with control parameter equal to $\Pr(Y \leq 3)$ instead of the median. This curve also starts in the point $(3.136, 0.506)$, the minimizer of the Kullback-Leibler divergence from the parametric family to the true distribution, but as a increases, the minimizer of $d_{a,\mu}$ moves towards $(3.169, 0.509)$. The probability of a Gamma$(3.169, 0.509)$-distributed variable being greater than 3 is 0.168, which is, indeed, the true value of $\Pr(Y \leq 3)$ when $Y \sim \text{Weibull}(2, 7)$.

In Figure 8.2, we present another situation. We have computed minimizers of $d_{a,\mu}$ for different choices of control and balance parameters and plotted them all in the same figure. The true underlying density is a gamma distribution with shape parameter 2 and rate parameter 1/2. The normal distribution is used as the parametric family. As in Figure 8.1, we notice that with $a = 0$, the minimizer of $d_{a,\mu}$ is the same regardless of control parameter. This is not surprising as a balance parameter of 0 corresponds to no weight being put on the empirical likelihood part of $d_{a,\mu}$. Hence, the minimizer of $d_{a,\mu}$ is the minimizer of the Kullback-Leibler distance regardless of control parameter.

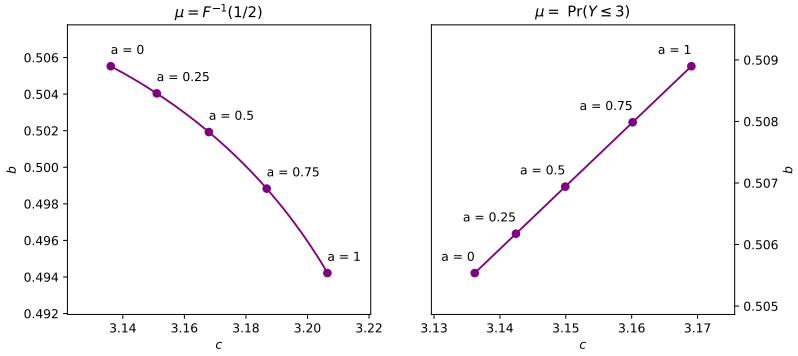


Figure 8.1: The plots display the minimizer of $d_{a,\mu}$ with different values of balance parameter, a, and control parameter, μ. The true distribution is a Weibull distributed with shape parameter 2 and scale parameter 7. The parametric family used is the Gamma distribution with shape parameter c and rate parameter b.

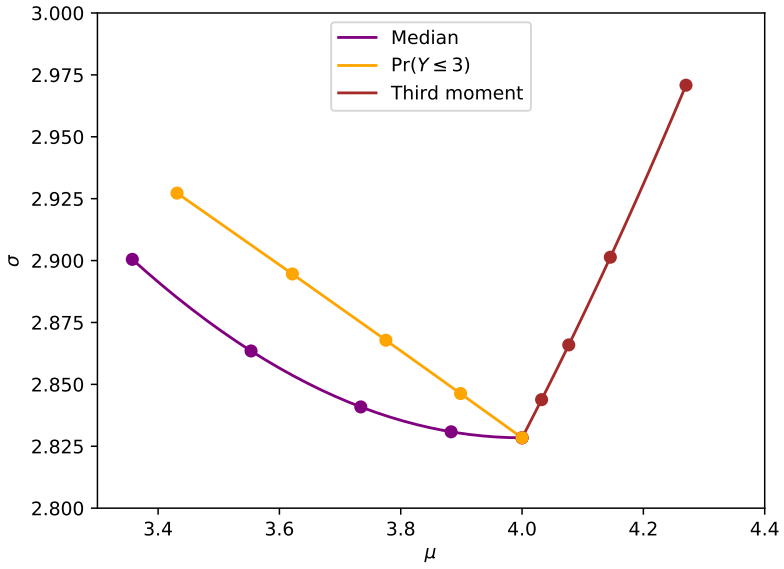


Figure 8.2: The plot displays minimizers of $d_{a,\mu}$ with different values of balance parameter, a, and control parameter shown in the legend. The dots on the curves indicate where $a = 0, 0.25, 0.5, 0.75, 1$ respectively, with the curves meeting at $a = 0$. The data is assumed to be Gamma distributed with shape parameter 2 and rate parameter $1/2$. The parametric family used is the normal distribution.

8.2 Limit distributions

In the previous section, we proved consistency of the maximum hybrid likelihood estimator towards the minimizer, θ_0, of a certain distance function. This ensures

$$\widehat{\theta}_{hl} - \theta_0 \xrightarrow{\Pr} 0,$$

but we know nothing about the speed of this convergence. The following theorem ensures that indeed

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) = O_{\Pr}(1).$$

This is needed to prove asymptotic normality of the maximum hybrid likelihood estimator, which is the goal for this section.

Lemma 8.2.1. *Let $\widehat{\theta}_{hl}$, θ_0 and Θ_0 be as in Theorem 8.1.1 and let the conditions of this result hold true. Assume furthermore that θ_0 lies in the interior of Θ_0 and that Γ admits a second-order Taylor expansion at this point with non-singular Hessian matrix, $H\Gamma(\theta_0)$. If there exists a function p_4, such that*

$$|\log\, f_{\theta_1}(y) - \log\, f_{\theta_2}(y)| \leq p_4(y)\|\theta_1 - \theta_2\| \tag{8.8}$$

for almost all y and all θ_1, θ_2 in a neighborhood of θ_0 on which μ is continuously differentiable,

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) = O_{\Pr}(1),$$

provided

$$\mathrm{E}\, p_4(Y)^2, \mathrm{E}\, p_4(Y)p_2(Y), \mathrm{E}\, p_4(Y)p_1(Y) < \infty,$$

where p_1 and p_2 as defined in Lemma 6.1.2 and Lemma 6.1.1 respectively.

Proof. We will use corollary 5.53 in Vaart 1998, p. 77 and argue very similarly as in the proof of Lemma 6.3.1. Here we derived the following:

$$\begin{aligned}&|\log\{1 + \lambda(\theta_1)m[y, \mu(\theta_1)]\} - \log\{1 + \lambda(\theta_2)m[y, \mu(\theta_2)]\}| \leq \\ &\frac{1}{L}(K_2 p_2(y) + K_1 p_1(y))\|\mu(\theta_1) - \mu(\theta_2)\|,\end{aligned}$$

for constants $L, K_1, K_2 < \infty$ and θ_1, θ_2 in the neighborhood of θ_0 described in Lemma 8.2.1. Since μ is continuously differentiable on this set, a combination of the mean value theorem for functions of several variables and the extreme value theorem ensures the existence of $K_3 < \infty$ such that

$$\|\mu(\theta_1) - \mu(\theta_2)\| \leq K_3\|\theta_1 - \theta_2\|.$$

Let

$$\psi(y, \theta) = (1 - a)\log\, f_\theta(y) - a\log\{1 + \lambda(\theta)m[y, \mu(\theta)]\}.$$

Combining the above with condition (8.8), shows

$$|\psi(y, \theta_1) - \psi(y, \theta_2)| \leq \left((1 - a)p_4(y) + a\frac{K_3}{L}[K_1 p_2(y) + K_2 p_1(y)]\right)\|\theta_1 - \theta_2\|$$

for all θ_1, θ_2 in a neighborhood of θ_0. Since

$$\mathrm{E}\bigg((1-a)p_4(Y) + a\frac{K_3}{L}[K_2p_2(Y) + K_1p_1(Y)]\bigg)^2$$

is finite by assumption, this proves the first condition of corollary 5.53 in Vaart 1998.

The vector $\widehat{\theta}_{hl}$ maximizes the hybrid likelihood function and

$$\Gamma_n(\theta) = \frac{1}{n}h_n(\theta) + O_{\mathrm{Pr}}(1/n)$$

uniformly in θ by Lemma 6.1.2 and Lemma 6.1.4. Because of this,

$$\Gamma_n(\widehat{\theta}_{hl}) \geq \Gamma_n(\theta_0) + O_{\mathrm{Pr}}(1/n).$$

The remaining conditions of corollary 5.53 hold by assumption, so this concludes the proof. ■

Lemma 8.2.1 ensures that the sequence

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big)$$

is bounded in probability. What remains to show is that this quantity goes to a normal limit. This is guaranteed by the following theorem.

Theorem 8.2.2. *Let $\widehat{\theta}_{hl}$, θ_0 and Θ_0 be as in Theorem 8.1.1 and assume the conditions of this result and Lemma 8.2.1 hold true. Assume*

$$\theta \mapsto m[y, \mu(\theta)] \quad \textit{and} \quad \theta \mapsto \log f_\theta(y)$$

are differentiable at θ_0 for almost every y, and let

$$J^* = -H\Gamma(\theta_0) \quad \textit{and} \quad U_n^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\Big|_{\theta_0}\psi(Y_i, \theta) \tag{8.9}$$

with

$$\psi(y, \theta) = (1-a)\log\, f_\theta(y) - a\log\{1 + \lambda(\theta)m[y, \mu(\theta)]\}.$$

Then

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) = (J^*)^{-1}U_n^* + o_{\mathrm{Pr}}(1). \tag{8.10}$$

In particular,

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) \xrightarrow{d} N\big(0, (J^*)^{-1}K^*(J^*)^{-1}\big), \tag{8.11}$$

where

$$K^* = \mathrm{Var}\bigg(\frac{\partial}{\partial\theta}\Big|_{\theta_0}\psi(Y, \theta)\bigg).$$

Proof. The result follows more or less directly from the proof of Theorem 6.3.2. Using arguments from this results and the proof of theorem 5.23 in Vaart 1998, p. 53, we arrive at the following identity

$$h_n\big(\theta_0 + \tilde{s}_n/\sqrt{n}\big) - s_n(\theta_0) = -\frac{1}{2}\tilde{s}_n J^* \tilde{s}_n + \tilde{s}_n^T U_n^* + o_{\Pr}(1),$$

for all sequences $\tilde{s}_n$ bounded in probability. In particular, this holds for the sequences

$$\widehat{s}_n = \sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) \quad \text{and} \quad s_n^* = (J^*)^{-1} U_n^*.$$

Since $\widehat{\theta}_{hl}$ maximizes h_n,

$$-\frac{1}{2}\widehat{s}_n^T J^* \widehat{s}_n + \widehat{s}_n^T U_n^* + o_{\Pr}(1) \geq -\frac{1}{2}(s_n^*)^T J^* s_n^* + (s_n^*)^T U_n^* + o_{\Pr}(1).$$

Manipulating this expression, leaves us with

$$-\frac{1}{2}(\widehat{s}_n - s_n^*)^T J^* (\widehat{s}_n - s_n^*) + o_{\Pr}(1) \geq 0.$$

By arguments similar to those given in the proof of Theorem 6.3.2, this implies

$$\widehat{s}_n = s_n^* + o_{\Pr}(1) = (J^*)^{-1} U_n^* + o_{\Pr}(1),$$

showing (8.10).

Since θ_0 is a maximizer of Γ in the interior of Θ,

$$0 = \Gamma'(\theta_0) = \frac{\partial}{\partial\theta}\Big|_{\theta_0} \mathrm{E}\,\psi(Y,\theta) = \mathrm{E}\,\frac{\partial}{\partial\theta}\Big|_{\theta_0} \psi(Y,\theta). \tag{8.12}$$

Hence,

$$U_n^* \xrightarrow{d} \mathrm{N}(0, K^*)$$

with

$$K^* = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_0} \psi(Y,\theta_0)\right). \tag{8.13}$$

This shows (8.11).

In (8.12) we interchanged differentiation and expectation. By Leibniz integral theorem, this is unproblematic if there exists a function g such that $\mathrm{E}\,g(Y) > \infty$ and

$$\left|\frac{\partial}{\partial\theta}\Big|_{\theta_0} \psi(y,\theta)\right| \leq g(y)$$

for all θ in a neighborhood of θ_0 and almost all y. This condition holds true by assumption as there is a neighborhood of θ_0 on which

$$|\psi(y,\theta) - \psi(y,\theta_0)| \leq g(y)\|\theta - \theta_0\|$$

for a function g with finite expectation. This was shown in the proof of Lemma 8.2.1. Hence,

$$\left|\frac{\partial}{\partial\theta}\Big|_{\theta_0} \psi(y,\theta)\right| \leq g(y)$$

as desired. ■

In the proof of Theorem 8.2.2, we showed

$$h_n(\theta_0 + \tilde{s}_n/\sqrt{n}) - \frac{1}{n}s_n(\theta_0) = -\frac{1}{2}\tilde{s}_n^T J^* \tilde{s}_n + \tilde{s}_n^T U_n^* + o_{\Pr}(1),$$

for all sequences $\tilde{s}_n$ bounded in probability. Inspecting the arguments, in particular those given in the proof of lemma 19.31 in Vaart 1998, p. 284, we notice something slightly stronger has been shown, namely,

$$\begin{aligned} A_n(s) &= h_n(\theta_0 + s/\sqrt{n}) - h_n(\theta_0) \\ &= -\frac{1}{2}s^T J^* s + s^T U_n^* + o_{\Pr}(1) \end{aligned}$$

uniformly over compacts sets K. Furthermore,

$$Z_n(s) = -\frac{1}{2}s^T J^* s + s^T U_n^*$$

is a convex process converging pointwise to

$$A(s) = -\frac{1}{2}s^T J^* s + s^T U^*.$$

Hence, by Arcones 1998,

$$A_n \xrightarrow{d} A$$

as a process in $\ell^\infty(K)$ for each compact set K. See the proof of Theorem 3.0.5 for more details about arguments like these. Using the process convergence and Lemma 8.2.1, we can prove a profiling result for the hybrid likelihood function outside model conditions. The argument is similar to the one given in the proof of Theorem 7.4.2 and will be omitted, but we state the result in the ensuing theorem.

Theorem 8.2.3. *Assume the conditions of Theorem 8.2.2 hold true, and let $g\colon \mathbb{R}^p \to \mathbb{R}$ be a map for which the second order partial derivatives are all continuous. With $h_{n,prof}$ defined as in Definition 7.4.1 and D_n as in Theorem 7.4.2,*

$$\kappa \cdot D_n(\psi_0) \xrightarrow{d} \chi_1^2$$

where

$$\kappa = \frac{b^T (J^*)^{-1} b}{b^T (J^*)^{-1} K^* (J^*)^{-1} b},$$

b denotes the gradient of g at θ_0 and $\psi_0 = g(\theta_0)$. Here θ_0 is the maximizer of Γ as defined in (8.3) *and J^* and K^* are as in Theorem 8.2.2.*

Theorem 8.2.3 can be used to construct approximate confidence intervals and curves for focus parameters on the form $g(\theta)$ when g is sufficiently smooth. In Section 10.2 we will use the result to, yet again, make inference about the ratio of median battle deaths in wars before and after the Korean War.

8.3 Consistent estimators

The limit distributions derived in the previous section involve the matrices J^* and K^*. In practice, these quantities are unknown. Because of this, we need to estimate them before Theorem 8.2.2 or Theorem 8.2.3 can be applied. In this section, we will define some estimators of the matrices and give conditions for consistency. All notation will be the same as in the previous sections.

We start with the matrix

$$J^* = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta\partial\theta^T}\Big|_{\theta_0} \{(1-a)\log f_\theta(Y) - a\log[1+\lambda(\theta)M(\theta)]\} \right),$$

or equivalently

$$J^* = -H\Gamma(\theta_0),$$

where Γ is defined as in Equation (8.3) on page 108 and $H\Gamma(\theta_0)$ denotes its Hessian matrix at θ_0, the maximizer of Γ. From Section 8.1, we know that for each $\theta \in \Theta$

$$\frac{1}{n} h_n(\theta) = \Gamma(\theta) + \epsilon_n(\theta) \tag{8.14}$$

with $\epsilon_n(\theta)$ tending uniformly to 0 in probability. Because of this, we would expect the Hessian matrix of h_n/n and Γ to be close. This is true provided

$$\|H\epsilon_n(\theta)\| = o_{\mathrm{Pr}}(1) \tag{8.15}$$

for $\theta \in \Theta$. Here $H\epsilon_n(\theta)$ denotes the Hessian matrix of ϵ_n at θ. In the following we will assume (8.15).

Assuming (8.15), we have

$$\frac{1}{n} Hh_n(\theta) = H\Gamma(\theta) + o_{\mathrm{Pr}}(1), \tag{8.16}$$

for every fixed $\theta \in \Theta$. Hence, $-Hh_n(\theta_0)/n$ estimates J^* consistently. In practice, however, θ_0 is not known. We would therefore like to replace it with its canonical estimator, $\widehat{\theta}_{hl}$. Since the maximum hybrid likelihood estimator converges in probability to θ_0, we would expect $-Hh_n(\widehat{\theta}_{hl})/n$ to converge to $-H\Gamma(\theta_0) = J^*$. Sadly, this does not hold true in general. We need additional regularity to have $-Hh_n(\widehat{\theta}_{hl}) \xrightarrow{\mathrm{Pr}} J^*$. One sufficient condition involves the notion of stochastic equicontinuity.

Definition 8.3.1 (Stochastically equicontinuity Pollard 1984, p. 139)**.** Let T be a metric space and $Z_n \colon T \to \mathbb{R}^k$ for $n = 1, 2, \ldots$ a sequence of stochastic processes. Fix $t_0 \in T$. $\{ Z_n \}_{n=1}^{\infty}$ is stochastically equicontinuous at t_0 if, for every $\eta, \epsilon > 0$, there is a neighborhood U of t_0 such that

$$\limsup_{n\in\mathbb{N}} \Pr\left(\sup_{t\in U} \|Z_n(t) - Z_n(t_0)\| > \eta \right) < \epsilon.$$

Stochastic equicontinuity guarantees that, for large values of n, $Z_n(t)$ and $Z_n(t_0)$ are close when t and t_0 are not far apart. Hence, if t_n is a sequence converging in probability to t_0, $Z_n(t_n)$ and $Z_n(t_0)$ should eventually take quite similar values. We state this formally in the ensuing lemma.

Lemma 8.3.2 (Pollard 1984, p. 140)**.** *Let* $\{ Z_n \}_{n=1}^{\infty}$ *be a sequence of stochastic processes from a metric space,* T*, to* $\mathbb{R}^k$*. Assume further that* $t_n \in T$ *is a sequence of estimators converging in probability to* t_0 *and that* $\{ Z_n \}_{n=1}^{\infty}$ *is stochastically equicontinuous at this limit. Then*

$$Z_n(t_n) - Z_n(t_0) = o_{\Pr}(1).$$

The above result follows more or less directly from Definition 8.3.1. An argument is given in Pollard 1984, p. 139–140.

Lemma 8.3.2 grantees that stochastic equicontinuity of Hh_n/n at θ_0 is enough to have $-Hh_n(\widehat{\theta}_{hl})/n$ and $-Hh_n(\theta_0)/n$ asymptotically equivalent. As this last quantity tends to J^* in probability by (8.16), the following is a consistent estimator of J^*:

$$\widehat{J^*} = -\frac{1}{n} Hh_n(\widehat{\theta}_{hl}), \tag{8.17}$$

provided $\{ Hh_n/n \}_{n=1}^{\infty}$ is an equicontinuous sequence of processes and the Hessian matrix of the remainder term in (8.14) tends to 0 in probability at θ_0.

(8.17) is a practical estimator. The quantity can easily be calculated and is, in fact, returned by most numerical optimization methods. In addition, its form is similar to that of the observed information matrix,

$$\widehat{J} = -\frac{1}{n} H\ell_n(\widehat{\theta}_{ml}),$$

used in maximum likelihood theory. In the above, ℓ_n is the likelihood and $\widehat{\theta}_{ml}$ its maximizer.

Now that estimation of J^* is taken care of, we will address $\lambda(\theta_0)$, the solution to

$$0 = \mathrm{E}\left(\frac{M(\theta_0)}{1 + \lambda M(\theta_0)} \right).$$

By Theorem 5.2.1, $\lambda_n(\theta)$ goes pointwise in probability to $\lambda(\theta)$ for $\theta \in \Theta$. Hence, $\lambda_n(\theta_0) \xrightarrow{\Pr} \lambda(\theta_0)$. Similarly as before, this quantity cannot be used in practice, as θ_0 is usually unknown. We would therefore like to use the estimator $\lambda_n(\widehat{\theta}_{hl})$ instead. This vector is found as part of the constrained optimization problem that arises in computation of $\mathrm{EL}_n[\mu(\widehat{\theta}_{hl})]$ (see Section 2.2). It is therefore available if the hybrid likelihood function can be computed. Under the conditions of Lemma 6.1.1,

$$\sup_{\theta \in \Theta} |\lambda_n(\theta) - \lambda(\theta)| = o_{\Pr}(1). \tag{8.18}$$

By the triangle inequality,

$$\left| \lambda_n(\widehat{\theta}_{hl}) - \lambda(\theta_0) \right| \leq \left| \lambda_n(\widehat{\theta}_{hl}) - \lambda(\widehat{\theta}_{hl}) \right| + \left| \lambda(\widehat{\theta}_{hl}) - \lambda(\theta_0) \right|$$

The first term goes to 0 in probability by (8.18) and the second term by consistency of $\widehat{\theta}_{hl}$ towards θ_0 and continuity of λ and μ. This ensures that $\lambda_n(\widehat{\theta}_{hl})$ is a consistent estimator of $\lambda(\theta_0)$.

We will also need a consistent estimator of $\lambda'(\theta_0)$. By the implicit function theorem,

$$\lambda'(\theta_0) = -\left[\frac{\partial}{\partial\lambda}\Big|_{\lambda(\theta_0)} \mathrm{E}\left(\frac{M(\theta_0)}{1+\lambda M(\theta_0)}\right)\right]^{-1} \frac{\partial}{\partial\theta}\Big|_{\theta_0} \mathrm{E}\left(\frac{M(\theta)}{1+\lambda(\theta_0)M(\theta)}\right).$$

Furthermore, the assumptions of Lemma 6.1.1 and Lemma 6.1.2 guarantees

$$\left|\frac{m[y,\mu(\theta)]^2}{\{1+\lambda(\theta)m[y,\mu(\theta)]\}^2}\right| \leq \frac{p_1(y)^2}{L^2}$$

for some $L < \infty$ and p_1 with $\mathrm{E}\, p_1(Y) < \infty$. This is enough for Leibniz integral theorem to be applicable, ensuring

$$\frac{\partial}{\partial\lambda}\Big|_{\lambda(\theta_0)} \mathrm{E}\left(\frac{M(\theta_0)}{1+\lambda M(\theta_0)}\right) = -\mathrm{E}\left(\frac{M(\theta_0)}{1+\lambda(\theta_0)M(\theta_0)}\right)^2.$$

Similarly, we can show

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta_0} \mathrm{E}\left(\frac{M(\theta)}{1+\lambda(\theta_0)M(\theta)}\right) = \\
&\mathrm{E}\left(\frac{M'(\theta_0)+M'(\theta_0)\lambda(\theta_0)M(\theta_0)-M(\theta_0)\lambda(\theta_0)M'(\theta_0)}{[1+\lambda(\theta_0)M(\theta_0)]^2}\right) = \\
&\mathrm{E}\left(\frac{M'(\theta_0)}{[1+\lambda(\theta_0)M(\theta_0)]^2}\right),
\end{aligned}$$

using arguments like those above and as given in the proof of Corollary 6.3.3. Because of this,

$$\lambda'(\theta_0) = \left[\mathrm{E}\left(\frac{M(\theta_0)}{1+\lambda(\theta_0)M(\theta_0)}\right)^2\right]^{-1} \mathrm{E}\left(\frac{M'(\theta)}{[1+\lambda(\theta_0)M(\theta)]^2}\right). \tag{8.19}$$

Let $Z_n\colon \mathbb{R}^{p+1} \to \mathbb{R}$ be defined as

$$Z_n(\theta,\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i(\theta)}{1+\lambda M_i(\theta)}\right)^2.$$

Assuming equicontinuity of $\{\, Z_n \,\}_{n=1}^{\infty}$ at $(\theta_0, \lambda(\theta_0))$, is enough to have

$$Z_n\Big(\widehat{\theta}_{hl}, \lambda_n(\widehat{\theta}_{hl})\Big) - Z_n[\theta_0, \lambda(\theta_0)] = o_{\mathrm{Pr}}(1).$$

As before, this follows from Lemma 8.3.2 and consistency of $\widehat{\theta}_{hl}$ and $\lambda_n(\widehat{\theta}_{hl})$ towards θ_0 and $\lambda(\theta_0)$ respectively. Furthermore, $Z_n[\theta_0, \lambda(\theta_0)]$ converges in probability to

$$\mathrm{E}\left(\frac{M(\theta_0)}{1+\lambda(\theta_0)M(\theta_0)}\right)^2,$$

ensuring that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i(\widehat{\theta}_{hl})}{1+\lambda_n(\widehat{\theta}_{hl})M_i(\widehat{\theta}_{hl})}\right)^2 = \mathrm{E}\left(\frac{M(\theta_0)}{1+\lambda(\theta_0)M(\theta_0)}\right)^2 + o_{\mathrm{Pr}}(1),$$

under the assumption of equicontinuity of $\{ Z_n \}_{n=1}^{\infty}$ at $(\theta_0, \lambda(\theta_0))$. Similarly, one can show

$$\frac{1}{n}\sum_{i=1}^{n}\frac{M_i'(\widehat{\theta}_{hl})}{[1+\lambda_n(\widehat{\theta}_{hl})M_i(\widehat{\theta}_{hl})]^2} = \mathrm{E}\left[\frac{M'(\theta_0)}{[1+\lambda(\theta_0)M(\theta_0)]^2}\right] + o_{\mathrm{Pr}}(1),$$

provided equicontinuity at $(\theta_0, \lambda(\theta_0))$ of the corresponding sequence of processes. The continuous mapping theorem now ensures that

$$\widehat{\lambda'(\theta_0)} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i(\widehat{\theta}_{hl})}{1+\lambda_n(\widehat{\theta}_{hl})M_i(\widehat{\theta}_{hl})}\right)^2\right]^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{M_i'(\widehat{\theta}_{hl})}{[1+\lambda_n(\widehat{\theta}_{hl})M_i(\widehat{\theta}_{hl})]^2} \tag{8.20}$$

is a consistent estimator of $\lambda'(\theta_0)$, provided the conditions outlined above.

We are now ready to find a consistent estimator of K^*. By definition,

$$K^* = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_0}\psi(Y,\theta)\right)$$

with

$$\psi(y,\theta) = (1-a)\log f_\theta(y) - a\log\{1+\lambda(\theta)m[y,\mu(\theta)]\}.$$

Define the following function

$$\eta(y,\theta,\lambda,\lambda') = (1-a)u(y,\theta) - a\frac{(\lambda')^T m[y,\mu(\theta)] + \frac{\partial m}{\partial\theta}[y,\mu(\theta)]^T\lambda}{1+\lambda m[y,\mu(\theta)]}$$

Direct computation shows

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0}\psi(y,\theta)^T = \eta[y,\theta_0,\lambda(\theta_0),\lambda'(\theta_0)],$$

and so

$$K^* = \mathrm{E}\,\eta[Y,\theta_0,\lambda(\theta_0),\lambda'(\theta_0)]\eta[Y,\theta_0,\lambda(\theta_0),\lambda'(\theta_0)]^T$$

By arguing as previous in this chapter, this ensures that

$$\widehat{K}^* = \frac{1}{n}\sum_{i=1}^{n}\eta\Big[Y_i,\widehat{\theta}_{hl},\lambda_n(\widehat{\theta}_{hl}),\widehat{\lambda'(\theta_0)}\Big]\eta\Big[Y_i,\widehat{\theta}_{hl},\lambda_n(\widehat{\theta}_{hl}),\widehat{\lambda'(\theta_0)}\Big]^T \tag{8.21}$$

is a consistent estimator of K^* provided equicontinuity of the process

$$(\theta,\lambda,\lambda') \mapsto \frac{1}{n}\sum_{i=1}^{n}\eta(Y_i,\theta,\lambda,\lambda')^T\eta(Y_i,\theta,\lambda,\lambda')$$

at $(\theta_0, \lambda(\theta_0), \lambda'(\theta_0))$.

8.4 What if the model is correct?

In the previous sections, we derived several limit results. At no point was the true underlying distribution assumed to be a member of the parametric family, f_θ for $\theta \in \Theta$. The theorems can therefore be applied in situations where no member of the parametric family is the true density of the data. That being said, the results hold true when this is the case as well. In this section, we will assume the true underlying distribution, indeed, has a density f_{θ_0} for some $\theta_0 \in \Theta$ and show that the limit distributions from the current chapter agree with those of Chapter 7. For this, it suffices to prove that the limit of U_n^* given in (8.9) is U^* from (7.6), and that J^* from (8.9) and (7.7) are the same matrices.

We start by assuming the true underlying density, f, is a member of the parametric family fit to the data. Then there exists $\theta_0 \in \mathbb{R}^p$, such that $f = f_{\theta_0}$. Since θ_0 minimizes both the Kullback-Leibler divergence and $\mu(\theta_0)$ is the true value of the control parameter, θ_0 maximizes both the parametric and the empirical likelihood part of Γ given in (8.3). See Section 8.1 for further explanation. Because of this, application of both Corollary 7.2.2 and Theorem 8.1.1 ensures consistency of the hybrid likelihood estimate towards the θ_0.

Let u be the score function of f_θ and μ_0 the true value of the control parameter, $\mu(\theta_0)$. Then

$$\begin{aligned}\frac{\partial}{\partial\theta}\Big|_{\theta_0}\psi(y,\theta) &= (1-a)u(y,\theta_0)^T - a\frac{\lambda'(\theta_0)^T m(y,\mu_0)+0}{1+0}\\ &= (1-a)u(y,\theta_0) - a\lambda'(\theta_0)^T m(y,\mu_0),\end{aligned}$$

where we have used $\lambda(\theta_0) = 0$, shown in the arguments preceding Theorem 6.2.1. By (8.19) and $\lambda(\theta_0) = 0$,

$$\lambda'(\mu_0) = \frac{\mathrm{E}\, M'(\theta_0)}{S(\theta_0)},$$

where $S(\theta_0) = \mathrm{E}\, M(\theta_0)^2$. By the remarks the preceding Theorem 7.2.1 on page 88, the nominator in this expression is equal to the matrix ξ_0 used in Chapter 7, under weak conditions. Hence,

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0}\psi(y,\theta)^T = (1-a)u(y,\theta_0) - a\xi_0^T S(\mu_0)^{-1}m(y,\mu_0),$$

resulting in

$$U_n^* = (1-a)U_n(\theta_0) - a\xi_0^T S(\theta_0)^{-1}V_n(\theta_0)$$

with

$$U_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n u(Y_i,\theta_0) \quad \text{and} \quad V_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n m(Y_i,\mu_0).$$

Since,

$$\begin{pmatrix}U_n(\theta_0)\\ V_n(\theta_0)\end{pmatrix} \xrightarrow{d} \begin{pmatrix}U\\ V\end{pmatrix},$$

with U and V as in (7.4), the limit of U_n^* defined in (8.9) is U^* given in (7.6).

Consider now the matrix J^* given in (8.9). By definition,

$$J^* = (1-a)\frac{\partial^2}{\partial\theta\partial\theta^T}\Big|_{\theta_0} \mathrm{E}\,\log\, f_\theta(Y) - a\frac{\partial^2}{\partial\theta\partial\theta^T}\Big|_{\theta_0} \mathrm{E}\,\log[1+\lambda(\theta)M(\theta)].$$

The first term in this expression is the Fisher matrix, J, as $f = f_{\theta_0}$. For the second term, notice

$$\begin{aligned}&\frac{\partial^2}{\partial\theta\partial\theta^T}\Big|_{\theta_0} \mathrm{E}\,\log\{1+\lambda(\theta)M(\theta)\} = \\ &\mu'(\theta_0)^T\left(\frac{\partial^2}{\partial\mu\partial\mu^T}\Big|_{\mu_0} \mathrm{E}\,\log[1+\lambda(\mu)M(\mu)]\right)\mu'(\theta_0)\end{aligned}$$

as

$$\frac{\partial}{\partial\mu}\Big|_{\mu_0} \mathrm{E}\,\log[1+\lambda(\mu)M(\mu)] = 0$$

since μ_0 is the maximizer of the expression by the arguments preceding Theorem 6.2.1. By the derivations from Corollary 6.3.3,

$$\frac{\partial^2}{\partial\mu\partial\mu^T}\Big|_{\mu_0} \mathrm{E}\,\log[1+\lambda(\mu)M(\mu)] = \mathrm{E}\,M'(\mu_0)^T S(\theta_0)^{-1}\mathrm{E}\,M'(\mu_0),$$

and hence,

$$\begin{aligned}\frac{\partial^2}{\partial\theta\partial\theta^T}\Big|_{\theta_0} \mathrm{E}\,\log\{1+\lambda(\theta)M(\theta)\} &= \mu'(\theta_0)^T\mathrm{E}\,M'(\mu)^T S(\theta_0)^{-1}\mathrm{E}\,M'(\mu)\mu(\theta_0)\\ &= \mathrm{E}\,M'(\theta_0)^T S(\theta_0)^{-1}\mathrm{E}\,M'(\theta_0)\\ &= \xi_0^T S(\theta_0)^{-1}\xi_0\end{aligned}$$

by the chain rule. This shows

$$J^* = (1-a)J - a\xi_0 S(\theta_0)^{-1}\xi_0,$$

which is the same expression as the one given in (7.7).

In the previous paragraph, we abused notation somewhat to make calculations more readable. We let $M(\mu)$ refer to $m(Y,\mu)$ and $M(\theta)$ to $m[Y,\mu(\theta)]$. We hope the intentional use of the symbols μ and θ removes confusion about this, otherwise ambiguous, notation.

The above ensures that using both Corollary 7.2.2 and Theorem 8.2.2 leads to the conclusion

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) \xrightarrow{d} (J^*)^{-1}U^*$$

with θ_0 equal to the true parameter and U^* and J^* defined in (7.6) and (7.7) respectively. Similarly, the calculations in this section ensures the limit distributions in Theorem 7.4.2 and Theorem 8.2.3 agree when the model is specified correctly.

CHAPTER 9

Focused information criterion for hybrid likelihood

When fitting parametric families to data, there are many decisions to make. What family should we choose? Are there conditions on the parameters? How many of the covariates in a regression setting are needed? When faced with decisions like these, it is popular to make use of information criteria. In this chapter, we will take a closer look at one particular such criterion and use it to address the elephant in the room: How do we choose the balance parameter used in construction of the hybrid likelihood function?

We will use the same short-hand notation here as in the previous chapter.

9.1 The focused information criterion

Information criteria are numbers that can be assigned to most models. For a fixed information criterion, the number represent the quality of the model fit in some way. We can evaluate different models by comparing their corresponding values of the information criterion. Ranking model fits with a single quantity is not only intuitively simple, but also quite convenient as comparing numbers is easier than full models. There are many information criteria available. Among the most popular are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), but there exists numerous others, like the deviance information criterion (DIC) and Mallow's C_p. See Claeskens and Hjort 2008b for definitions and discussions of these and other criteria. One thing all of the quantities mentioned here have in common, is that they evaluate the overall fit of the model. When using maximum hybrid likelihood, we do, however, give up some of the overall model fit in favor of robust estimation of the control parameter. Evaluation of different models fit using maximum hybrid likelihood, should therefore use a criterion giving credit to the robust estimation of the control parameter. Luckily such a criterion already exists. This is called the Focused Information criterion (FIC) and will be the topic of this chapter.

The focused information criterion, or FIC for short, was first introduced in Claeskens and Hjort 2003. Results concerning the criterion for sub-models fit with maximum likelihood are derived in this paper. Since publication of the original article, FIC has been further developed and extended. There are too many contributers for a complete list to be given her, but some important articles include Claeskens, Croux, and Van Kerckhoven 2006, who derived an

expression for FIC using more general loss functions than the mean squared error, Claeskens and Hjort 2008a, who worked with expressions for a weighted version of the criterion, and Jullum and Hjort 2017, who derived a version of FIC allowing for parametric and non-parametric models to be compared. In addition, numerous authors have worked with the criterion for specific classes of models. Examples include Zhang and Liang 2011 for generalized partial linear models and Claeskens, Croux, and Van Kerckhoven 2007 for autoregressive models. We will briefly explain the focused information criterion in this section, but a more comprehensive introduction can be found in Claeskens and Hjort 2008b along with an overview of literature.

FIC differs from most other information criteria in a monumental way. Rather than assessing the overall fit of the model, FIC ranks models by the quality of their estimator of a certain quantity. This quantity is called "the focus parameter" and in this section we will denote it by ψ. The choice of focus parameter depends on the setting and what we are interested in inferring from our data. Taking the data set from Section 4.1 as an example, one statistician might be interested in in the mean salary in Oslo, while another might want to estimate the median well. There is no general rule for what to use as a focus parameter, as what quantities are of interest is entirely dependent on the specific application. Hence, each statistician is free to choose a focus parameter tailored to their uses.

FIC ranks models by comparing the quality of estimators of the focus parameter. To use the information criterion, we therefore need to properly define what we mean by the "quality" of an estimator. For a general loss function, $L(\psi, \hat{\psi})$, describing how much is lost by estimating ψ by $\widehat{\psi}$, we define the risk of the estimator $\widehat{\psi}$ to be

$$R(\psi, \widehat{\psi}) = \mathrm{E}\, L(\psi, \widehat{\psi}).$$

In principle, FIC can measure quality of an estimator using any risk function, but in this thesis, we will only consider squared error loss. This is by far the most popular loss function and has the following formulation:

$$L(\psi, \widehat{\psi}) = \Big(\psi - \widehat{\psi}\Big)^2.$$

Its risk function is called the mean squared error, or MSE for short, and is defined as:

$$\mathrm{MSE}(\psi, \widehat{\psi}) = \mathrm{E}\Big(\psi - \widehat{\psi}\Big)^2.$$

In this chapter, we will mostly be interested in MSE evaluated at the true parameter value, ψ_0. We will therefore use $\mathrm{MSE}(\widehat{\psi})$ as short-hand for $\mathrm{MSE}(\psi_0, \widehat{\psi})$.

For an estimator, $\widehat{\psi}$, of ψ mean squared error decomposes as

$$\mathrm{MSE}(\widehat{\psi}) = \Big(\mathrm{E}\,\widehat{\psi} - \psi\Big)^2 + \mathrm{Var}\,\widehat{\psi}. \tag{9.1}$$

The first term is the bias of the estimator, and the second is the variance. To estimate the mean squared error of an estimator $\widehat{\psi}$, it therefore suffices to estimate each of these terms separately. This decomposition will be used when we derive an expression for FIC for models fit with maximum hybrid likelihood.

9.2 MSE of the maximum hybrid likelihood estimator

We are interested in deriving FIC for models fit with maximum hybrid likelihood. For a fixed focus parameter, ψ, FIC estimates the mean squared error of a model's estimate of ψ. So, to derive an expression for FIC for models fit with maximum hybrid likelihood, we need to estimate the bias and variance of the maximum hybrid likelihood estimators of the focus parameter. The limit results derived in the previous section will help us do exactly this.

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables following some distribution with cumulative distribution function F and

$$\mathcal{F} = \{ f_\theta \mid \theta \in \Theta \}$$

a parametric family we wish to fit to the data using maximum hybrid likelihood with some control parameter, μ, and balance parameter, a. Assume furthermore, there is a differentiable function, g, such that $g(\theta)$ is the value of ψ in the parametric model, f_θ. The maximum hybrid likelihood estimate of ψ in the parametric model is given by

$$\widehat{\psi}_{hl} = g(\widehat{\theta}_{hl}).$$

The FIC for models fitted with hybrid likelihood is the mean square error of this quantity.

In practice, the mean squared error of $\widehat{\psi}_{hl}$ is not known. To use the focused information criterion, we will therefore have to estimate $\mathrm{MSE}(\widehat{\psi}_{hl})$. We will start by using the limit results from the previous chapter to find an expression asymptotically equivalent to $\mathrm{MSE}(\widehat{\psi}_{hl})$.

Assume the conditions of Theorem 8.1.1 and Theorem 8.2.2 hold true. Then, the maximum hybrid likelihood estimator of θ has a normal limit distribution after proper centering and scaling. By the delta method, this implies

$$\widehat{\psi}_{hl} \overset{d}{\approx} \mathrm{N}\left(g(\theta_0), \frac{\nabla g(\theta_0)^T (J^*)^{-1} K^* (J^*)^{-1} \nabla g(\theta_0)}{n} \right),$$

where θ_0 is the limit of $\widehat{\theta}_{hl}$ defined in Section 8.1 and J^* and K^* are the matrices defined in (8.9) and (8.13) respectively. Furthermore, $\widehat{\psi}_{hl}$ converges in probability to $g(\theta_0)$ by the continuous mapping theorem and consistency of $\widehat{\theta}_{hl}$ towards θ_0. As a consequence, the expected value of the maximum hybrid likelihood estimate of ψ converges to $g(\theta_0)$ if $\widehat{\psi}_{hl}$ are uniformly integrable (see e.g. Billingsley 1999, p. 31). Hence, provided the necessary conditions, the following is a asymptotically equivalent to the mean squared error of $\widehat{\psi}_{hl}$:

$$[g(\theta_0) - \psi_0]^2 + \frac{\sigma^2}{n}. \tag{9.2}$$

Here ψ_0 denotes the true value of the focus parameter and

$$\sigma^2 = \nabla g(\theta_0)^T (J^*)^{-1} K^* (J^*)^{-1} \nabla g(\theta_0).$$

9.3 Estimating the MSE

(9.2) is a theoretically convenient expression. In practice, it is, however, less useful. The quantities involved are rarely, if not never, known. We will therefore

attempt to replace the unknown values in (9.2) with estimators. This leads to the definition of the focused information criterion for the hybrid likelihood, or HFIC for short.

A first approach

We start with a simple estimator of $\mathrm{MSE}(\widehat{\psi}_{hl})$. Estimation of J^* and K^* was discussed in Section 8.3. Consistent estimators of the matrices were derived in this section, and their formulas are given in (8.17) and (8.21) respectively. If, in addition, ∇g is a continuous function, $\nabla g(\widehat{\theta}_{hl})$ converges in probability to $\nabla g(\theta_0)$. Because of this, the following is only $o_{\Pr}(1)$ away from the σ^2.

$$\widehat{\sigma}^2 = \nabla g(\widehat{\theta}_{hl})^T (\widehat{J}^*)^{-1} \widehat{K}^* (\widehat{J}^*)^{-1} \nabla g(\widehat{\theta}_{hl}).$$

Hence, the second term of (9.2), the variance of the maximum hybrid likelihood estimate, can be estimated consistently by $\widehat{\sigma}^2/n$.

Estimation of the first term, the bias, is more complicated. A naive approach is to use the estimator,

$$\widehat{b} = \left(\widehat{\psi}_{hl} - \widehat{\psi}\right)^2, \tag{9.3}$$

for some consistent estimator $\widehat{\psi}$ of ψ. By the continuous mapping theorem, $\widehat{b}$ is asymptotically equivalent to the bias of the maximum hybrid likelihood estimate. Using this formula leads to a first estimate of $\mathrm{MSE}(\widehat{\psi}_{hl})$,

$$\widehat{b} + \frac{\widehat{\sigma}^2}{n}. \tag{9.4}$$

In (9.4) we do not make any assumptions about $\widehat{\psi}$ other than it being consistent for the true value of the focus parameter. In practice, this amounts to choosing a robust estimator that is not affected by possible misspecification of the model. Examples of such non-parametric choices are the sample mean or median provided, of course, their population versions are used as focus parameter.

Correcting for the bias

Although appealing because of its simplicity, (9.4) tends to overshoot the actual value of limiting mean squared error of $\widehat{\psi}_{hl}$. This is a consequence of the variability in $\widehat{b}$. To correct for this, we will find an approximation to the variance of the bias term and subtract it from (9.3). This approach is popular in newer versions of FIC and have been used in e.g. Jullum and Hjort 2017 and Claeskens, Cunen, and Hjort 2019. The following arguments and calculations are similar to what can be found in these articles.

To correct for the variance of the bias term, κ^2/n, we need to find and estimate it. Using the formula,

$$\frac{\kappa^2}{n} = \mathrm{Var}\left(\widehat{\psi}_{hl} - \widehat{\psi}\right) = \mathrm{Var}\,\widehat{\psi}_{hl} + \mathrm{Var}\,\widehat{\psi} - \mathrm{Cov}\left(\widehat{\psi}_{hl}, \widehat{\psi}\right),$$

we notice that this amounts to understanding the variance of the maximum hybrid likelihood estimate and the consistent estimator of the focus parameter

in addition to their covariance. In practice, this can be quite complicated if additional assumptions are not made on $\widehat{\psi}$. The assumption we will make, is that there exists a function $\phi\colon \mathbb{R}^{d+1} \to \mathbb{R}$ such that the following approximation holds:

$$\sqrt{n}\Big(\widehat{\psi} - \psi_0\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\phi(Y_i, \psi_0) - \mathrm{E}\,\phi(Y, \psi_0)] + o_{\Pr}(1), \tag{9.5}$$

when $Y \sim F$. At first glance, this condition might seem strict. It does, however, hold true for many natural focus parameters. If ψ is the expected value of the data, we have

$$\sqrt{n}\big(\bar{Y}_n - \psi_0\big) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathrm{E}\,Y),$$

where $\bar{Y}_n$ denotes the sample mean of $Y_1, \ldots, Y_n$. Similarly,

$$\sqrt{n}\Big(\overline{h(Y)}_n - \psi_0\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(Y_i) - \mathrm{E}\,h(Y)],$$

for functions $h\colon \mathbb{R}^d \to \mathbb{R}$. So, with $\widehat{\psi}$ equal to the sample mean of $h(Y_1), \ldots, h(Y_n)$, (9.5) is satisfied when the expected value of $h(Y)$ is used as focus parameter. A less obvious example is the median, $F^{-1}(0.5)$. For this focus parameter,

$$\sqrt{n}\big[\mathbb{F}_n^{-1}(0.5) - F^{-1}(0.5)\big] = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{I[Y_i \le F^{-1}(0.5)] - 0.5}{f[F^{-1}(0.5)]} + o_{\Pr}(1), \tag{9.6}$$

see e.g. Vaart 1998, p. 307. Here f denotes the density function in the distribution of the data, and, as before, F is the cumulative distribution function. Furthermore, $\mathbb{F}_n$ is the empirical cumulative distribution function,

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \le x).$$

By Theorem 8.2.2,

$$\sqrt{n}\Big(\widehat{\theta}_{hl} - \theta_0\Big) = (J^*)^{-1} U_n^* + o_{\Pr}(1),$$

where

$$J^* = -\frac{\partial^2}{\partial\theta\partial\theta^T}\Big|_{\theta_0} \mathrm{E}\{(1-a)\log f_\theta(Y) - a\log[1 + \lambda(\theta)M(\theta)]\},$$

and

$$U_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\eta(Y_i, \theta_0) - \mathrm{E}\,\eta(Y, \theta_0)],$$

with

$$\eta(y,\theta) = (1-a)u(y,\theta) - a\frac{\lambda'(\theta)^T m[y,\mu(\theta)] + \frac{\partial m}{\partial\theta}[y,\mu(\theta)]^T\lambda}{1 + \lambda(\theta)m[y,\mu(\theta)]}.$$

We can now apply the central limit theorem to get

$$\sqrt{n}\begin{pmatrix}\widehat{\psi}-\psi_0\\ \widehat{\theta}_{hl}-\theta_0\end{pmatrix}\xrightarrow{d}\begin{pmatrix}1 & 0\\ 0 & (J^*)^{-1}\end{pmatrix}N(0,\Sigma),$$

where Σ is the variance matrix of the vector $(\phi(Y,\psi_0),\eta(Y,\theta_0))^T$. Applying the delta method, results in

$$\sqrt{n}\begin{pmatrix}\widehat{\psi}-\psi_0\\ \widehat{\psi}_{hl}-g(\theta_0)\end{pmatrix}\xrightarrow{d}\begin{pmatrix}1 & 0\\ 0 & \nabla g(\theta_0)^T(J^*)^{-1}\end{pmatrix}N(0,\Sigma). \tag{9.7}$$

Furthermore, Σ is a $(1+p)\times(1+p)$-matrix which can be written as the block matrix

$$\Sigma=\begin{pmatrix}\tau^2 & C\\ C^T & K^*\end{pmatrix}.$$

The limit distribution in (9.7) is therefore a central normal distribution with covariance matrix

$$\begin{pmatrix}1 & 0\\ 0 & \nabla g(\theta_0)^T(J^*)^{-1}\end{pmatrix}\begin{pmatrix}\tau^2 & C\\ C^T & K^*\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & \nabla g(\theta_0)^T(J^*)^{-1}\end{pmatrix}^T=$$
$$\begin{pmatrix}\tau^2 & C(J^*)^{-1}\nabla g(\theta_0)\\ \nabla g(\theta_0)^T(J^*)^{-1}C^T & \nabla g(\theta_0)^T(J^*)^{-1}K^*(J^*)^{-1}\nabla g(\theta_0)\end{pmatrix}.$$

.

Provided sufficient regularity of ϕ, τ^2 can be estimated by the in-sample variance of $\phi(Y_1,\widehat{\psi}),\ldots,\phi(Y_n,\widehat{\psi})$. We will denote this estimator by $\widehat{\tau}^2$. The $1\times p$ matrix C can be estimated by, $\widehat{C}$, the in-sample covariance between $\psi(Y_i,\widehat{\psi})$ and $\eta(Y_i,\widehat{\theta}_{hl})$ for $i=1,\ldots,n$. The estimators derived in Section 8.3 can be used in place of $\lambda(\theta_0)$ and $\lambda'(\theta_0)$. Consistent estimation of J^*, $\nabla g(\widehat{\theta}_{hl})$ and $\widehat{\sigma}^2$ have already been discussed. Putting all of this together, results in the following estimator of κ^2:

$$\widehat{\kappa}^2=\widehat{\sigma}^2+\widehat{\tau}^2-2\cdot\widehat{C}(\widehat{J}^*)^{-1}\nabla g(\widehat{\theta}_{hl}). \tag{9.8}$$

Using (9.8), we can finally define an asymptotically unbiased estimator of the limiting mean squared error of the maximum hybrid likelihood estimator. This is the expression we will use as the focused information criterion for models fit with maximum hybrid likelihood.

Definition 9.3.1. With the definitions given previously in this chapter, we define two focused information criteria for models fit with maximum hybrid likelihood. The first one is HFIC^u:

$$\text{HFIC}^u=\left(\widehat{\psi}_{hl}-\widehat{\psi}\right)^2+\frac{\widehat{\sigma}^2}{n}-\frac{\widehat{\kappa}^2}{n}.$$

The second criterion, HFIC, is given by

$$\text{HFIC}=\frac{\widehat{\sigma}^2}{n}+\max\left\{\left(\widehat{\psi}_{hl}-\widehat{\psi}\right)^2-\frac{\widehat{\kappa}^2}{n},0\right\}.$$

As shown in the arguments preceding Definition 9.3.1, HFIC^u is an asymptotically unbiased estimator of the limiting mean squared error of $\widehat{\psi}_{hl}$. This is, of course, only true provided the conditions outlined in this chapter. The second criterion, HFIC, corrects for negative estimates of the limiting squared bias of $\widehat{\psi}_{hl}$ by truncating such values to 0.

9.4 Choosing the balance parameter

There is one obvious question regarding maximum hybrid likelihood estimation that we have avoided thus far: How should the balance parameter be chosen? In Section 7.5, we made this choice based on a predefined accepted loss of efficiency. Under model assumptions, the maximum hybrid likelihood estimate of a parameter converges in probability to the true value. This happens regardless of balance and control parameter. Hence, under sufficient regularity, the maximum hybrid likelihood estimate of a focus parameter is asymptotically unbiased, no matter the choice of of tuning parameters. Because of this, it is sufficient to compare variances to compare the limiting mean squared errors of different maximum hybrid likelihood estimators.

The same approach is not possible when the true distribution is not a member of the parametric family fit to the data. In such situations, the maximum likelihood estimate of a focus parameter is not guaranteed to be consistent for the true value. Furthermore, as we saw in Section 8.1, the hybrid likelihood estimate converges to different values depending on a and μ. Because of this, the asymptotic bias of the different estimators are not the same. Hence, comparison of variances only is not sufficient. We need to estimate and compare limiting mean squared errors for each choice of a and μ.

For this, the derivations from the previous section will be useful. Both HFIC^u and HFIC can be used to estimate the mean squared error of the maximum hybrid likelihood estimator of the focus parameter. To choose a balance parameter, a, we therefore propose the following strategy:

(1) Choose a focus parameter with an estimator that can be written as in (9.5) for some $\phi\colon \mathbb{R}^{d+1} \to \mathbb{R}$.

(2) Choose one of the two focus criteria defined in this chapter: HFIC^u or HFIC.

(3) Fit the parametric model to data using maximum hybrid likelihood for each value of a in a grid of points between 0 and 1 and compute the chosen criterion.

(4) Choose a minimizing the chosen criterion.

In the next chapter, we will use this procedure when we revisit the examples with incomes in Oslo from Section 4.1 and number of battle deaths in wars from Section 4.3 and Section 7.5.

In practice, the control parameter will often be used as the focus parameter. That being said, there is nothing in the above requiring ψ and μ to be equal. In Section 10.2, we will go through an example where these two parameters, indeed, are different.

CHAPTER 10

Examples

In this chapter, we will present some examples illustrating the results and methods from the two previous chapters.

10.1 Theory and practice

We will start by working with simulated data and compare the empirical results to the theoretical ones in Chapter 8 and Chapter 9. We will not focus on making inference about parameters in this section. For examples related to such problems, the two remaining sections of this chapter can be consulted.

We will work with Poisson distributed data with rate parameter equal to 2.5. The parametric model will be the family of geometric distributions. This has the following parameterization:

$$p_\theta(y) = (1-\theta)^y\theta \quad \text{for } y = 0, 1, \ldots \text{ and } \theta \in (0, 1].$$

For the empirical likelihood part of the hybrid likelihood function, we will use the estimating function

$$m(y, \mu) = I(y > 2) - \mu,$$

to increase robustness of the parametric estimate of $\Pr(Y > 2)$. In the geometric distribution this probability is given by $\mu(\theta) = (1-\theta)^3$. To begin with, we will put equal weight on the empirical and parametric part of the hybrid likelihood function. This corresponds to fixing the balance parameter at $1/2$.

The form of the estimating equation allows us to find an explicit formula for $\lambda(\mu)$, the solution to the equation

$$\begin{pmatrix} \mathrm{E}\{m(Y,\mu)/[1+\lambda m(Y,\mu)]\} \\ \Pr[1+\lambda m(Y,\mu) \leq 0] \end{pmatrix} = 0.$$

Straight forward computation shows

$$\mathrm{E}\left(\frac{m(Y,\mu)}{1+\lambda m(Y,\mu)}\right) = 0$$

is solved by

$$\lambda(\mu) = \frac{p-\mu}{\mu(1-\mu)}$$

with $p = \Pr(Y > 2)$ for $Y \sim \text{Pois}(2.5)$. Using this explicit formula, Γ, defined in Section 8.1, takes the form

$$\frac{1}{2}[2.5\log(1-\theta) + \log\theta] + \frac{1}{2}\left[p\log\left(\frac{(1-\theta)^3}{p}\right) + (1-p)\log\left(\frac{1-(1-\theta)^3}{1-p}\right)\right].$$

The minimizer of this function is $\theta_0 = 0.2650$. So by the results of Section 8.1 the maximum hybrid likelihood estimator of θ should aim for this value.

In the previous paragraph we found an explicit expression for Γ. Using this we can compute the exact value of J^* as defined in (8.9). Furthermore, the function ψ defined in Theorem 8.2.2 is given by the ensuing expression

$$\begin{aligned}&\frac{1}{2}[y\log(1-\theta) + \log\theta]\\ &-\frac{1}{2}\log\left(1 + \frac{p-(1-\theta)^3}{(1-\theta)^3[1-(1-\theta)^3]}[I(y>2) - (1-\theta)^3]\right),\end{aligned}$$

and hence, the true value of K^* can also be computed exactly. In particular, this means that the exact limiting variance of $\sqrt{n}\widehat{\theta}_{hl}$ can be calculated with the formula $K^*/(J^*)^2$. For our situation the exact limiting variance of $\sqrt{n}\widehat{\theta}_{hl}$ is 0.0048. To compare this with empirical results, we simulated 100 i.i.d. data points from a Pois(2.5)-distribution 300 times. In each iteration we found the maximum hybrid likelihood estimator. The empirical mean and variance of these data points were 0.2650 and 0.0041 respectively, which are both close to the theoretical values computed above.

The previous arguments can be applied to functions of θ as well. By the delta method

$$\sqrt{n}\Big(\mu(\widehat{\theta}_{hl}) - \mu(\theta_0)\Big) \xrightarrow{d} \text{N}\left(0, \frac{\mu'(\theta_0)^2 K^*}{(J^*)^2}\right),$$

and so the limiting mean squared error of $\mu(\widehat{\theta}_{hl})$ is given by

$$[\mu(\theta_0) - 0.456187]^2 + \frac{\mu'(\theta_0)^2 K^*}{n(J^*)^2},$$

as 0.456187 is the probability of a Pois(2.5)-distributed variable being greater than 2. As we have computed the exact value of θ_0, J^* and K^* previously in this example, the true value of the limiting mean squared error of $\mu(\widehat{\theta}_{hl})$ can be calculated to be 0.005660. This analysis is of course not limited to the case of $a = 1/2$, and all limits can be calculated similarly as above for other value of the balance parameter. In Figure 10.1 we have plotted the limiting mean squared error of $\mu(\widehat{\theta}_{hl})$ as a function of a. The information criterion defined in the previous chapter attempts to estimate this quantity. We have therefore also displayed a plot of HFIC-values calculated using a fixed set of 100 i.i.d. Pois(2.5)-distributed variables, using μ as focus parameter in the same figure.

From Figure 10.1, we notice that although the estimated and the exact curves are similar, the estimate works best for balance parameters close to 0.5 in this particular case. The minimizers of the two curves were 0.78 and 0.70 for the estimated and exact curve respectively. Using the information criterion from the previous chapter, therefore results in selection of a slightly higher value of a than the true optimal balance parameter. The difference between the two curves and minimizers are, however, not tremendous and indicates that the estimates work well in this particular case.

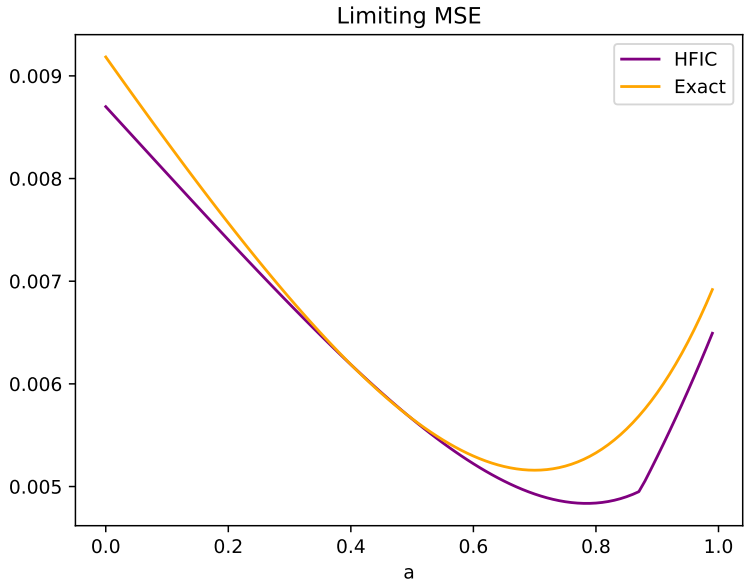


Figure 10.1: A plot of HFIC values together with the exact values of the limiting mean squared error of the maximum hybrid likelihood estimate of the control parameter. The sharp bend in the purple line is a consequence of the truncation done by HFIC.

10.2 Revisiting the deadly example yet again

In Section 7.5, we fitted inverse Burr models to two data sets, the number of battle deaths in 60 wars before the Korean war and 35 struggles after this conflict. In this section, we will use the methods developed in Chapter 8 and Chapter 9 to ensure the conclusions are more robust against misspecification of the model. We will use the same parametric model and control parameters as in Section 7.5. Many of the arguments will be similar to those given earlier in the thesis, so some details will be left out.

In Section 7.5, we used the procedure described in Section 7.3 to choose the balance parameter, a, used in construction of the hybrid likelihood function. In this section, we will no longer assume the data really is inverse Burr distributed, and hence, this method can no longer be applied. Instead, we will compute HFIC, as defined Definition 9.3.1, for each value of a in a grid of points between 0 and 1 and choose the balance parameter minimizing the information criterion. We chose to use the median as focus parameter. In this case, (9.6) must be used to compute HFIC. Since the true density of the data is unknown, we estimated $f[F^{-1}(0.5)]$ with a nonparametric kernel estimator. In Figure 10.2, we have plotted the result.

The minimizers found were $a_1 = 0.49$ and $a_2 = 0.46$ for the older and newer conflicts respectively. For neither of the data sets $a = 0$ was chosen. This indicates the inverse-Burr fit might not be ideal for estimation of the median.

Now that we have chosen balance parameters, we can find the maximum

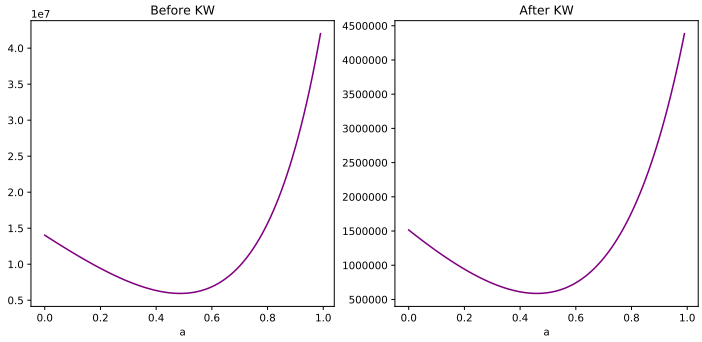


Figure 10.2: The plot on the left-hand side shows HFIC as a function of the balance parameter for the data set with battle deaths in wars before the Korean War. To the right, a corresponding plot for the newer conflicts is displayed. The minimum is $a_1 = 0.48$ for the first and $a_2 = 0.34$ for the second data set.

hybrid likelihood estimates of the median in the two models. For the older wars, the resulting estimate was 11702. For the newer conflicts, the estimate was 4930. Both these numbers are closer to their empircal estimates, 11375 and 5240 respectively, than the corresponding values obtained with maximum likelihood, 10399 and 4749. We can also compute approximate confidence curves using Theorem 8.2.2 and the delta method. This is very similar to what was done in Section 7.5. Details are therefore omitted, but a plot can be found in Figure 10.3. In place of J^* and K^* we used the estimators defined in Section 8.3.

When computing the variance of the maximum likelihood estimates, we used the matrix $J^{-1}KJ^{-1}$ in place of J^{-1}. Because of this, the dotted curves in Figure 10.3 are approximate confidence curves for the true medians in the distributions if the model is specified correctly. If this is not the case, however, they are approximate confidence curves for the minimizer of the Kullback-Leibler divergence by the results of White 1983. A similar idea applies to the fully drawn lines. If the data really follows an inverse-Burr distribution, they are approximate confidence curves for the median in the two distributions. If this is not the case, they can be used to make inference about $g(\theta_0)$ where θ_0 is the minimizer of the distance function described in Section 8.1 and g returns the median in the inverse Burr distribution with parameter θ for each θ.

Looking at Figure 10.3 we notice the confidence curves constructed using maximum hybrid likelihood theory are narrower than the corresponding curves obtained using maximum likelihood. At first glance, this might seem slightly surprising, as we are used to thinking of maximum likelihood estimators as asymptotically most efficient. This is, however, only true under model conditions. If the underlying distribution is not a member of the parametric model fit to the data, there is no guarantee that maximum likelihood estimators are consistent. Furthermore, the maximum hybrid likelihood estimators will typically aim for yet another different value. Hence, there is no theoretical result ensuring the variances of one should be greater than that of the other.

We can now use Theorem 8.2.3 to make inference about, ψ, the ratio of medians in the two data sets. After slightly modifying of the arguments and

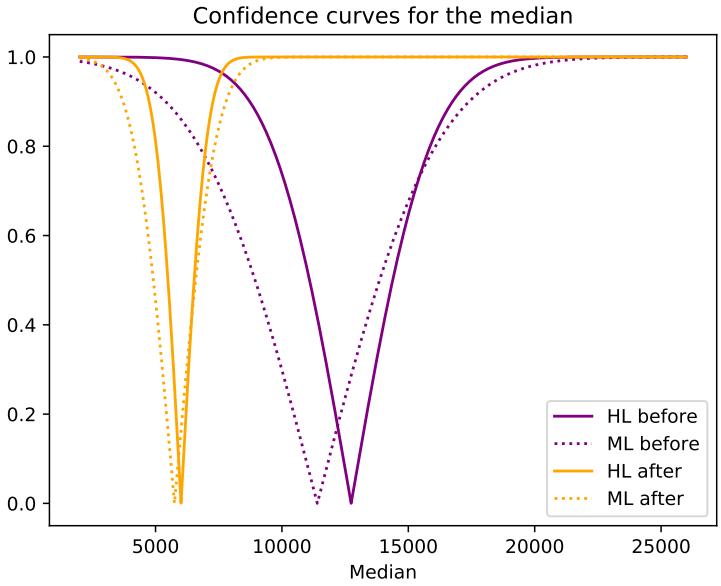


Figure 10.3: Approximate confidence curves for $g(\theta_0)$ where θ_0 is the minimizer of the distance function from Section 8.1 and $g(\theta)$ is the median in an Inverse-Burr distribution with parameters θ.

matrices, this is done similarly as in Section 7.5. We will therefore leave out the details, but a confidence curve for the focus parameter is given in Figure 10.4. The curve based on profile likelihood was construed using a version of Wilks theorem for misspecified models, see e.g. appendix A.5 of Schweder and Hjort 2016.

Figure 10.4 looks quite different from the corresponding plot in Section 7.5. The approximate confidence curve obtained using profile hybrid likelihood is much narrower than the one based on maximum likelihood. As in Figure 10.3, this is a consequence of a lacking model fit.

We can read confidence intervals and p-values off Figure 10.4. Using the curve corresponding to profile likelihood, we find that $[1.06, 4.50]$ is an approximate 90% confidence interval. For the profile hybrid likelihood the corresponding set is $[1.54, 3.59]$. Raising the level to 95%, results in the intervals $[0.92, 5.17]$ and $[1.42, 3.90]$ respectively. The p-value for testing $\psi = 1$ were 7.9% for profile likelihood and 0.1% for profile hybrid likelihood. These results do, however, not have as clear an interpretation as those of Section 7.5. If the model is specified correctly, these are indeed approximate confidence intervals for the true ratio of medians. If the true underlying distribution is not a member of the parametric family, however, they can be used to make inference about quantities expressed as functions of minimizers of the Kullback-Leibler divergence and $d_{a,\mu}$ from Section 8.1.

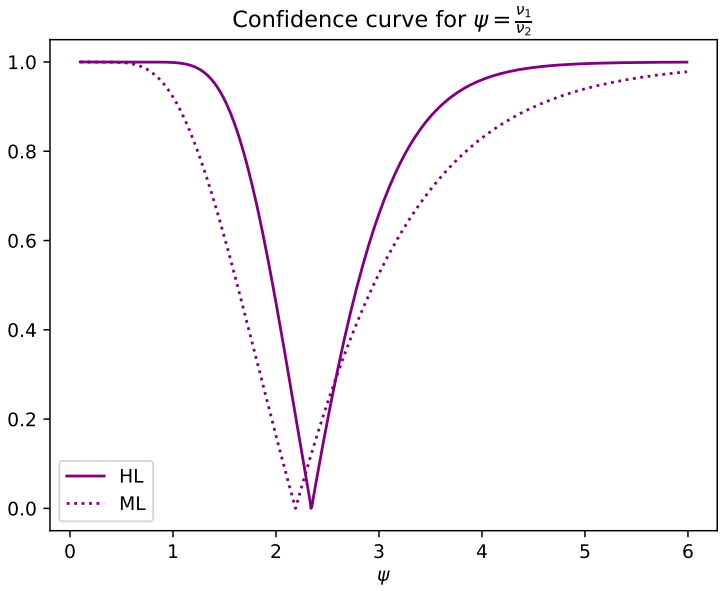


Figure 10.4: Approximate confidence curves for the ratio of medians in the two distributions. The full drawn line is constructed using Theorem 8.2.3, while the dotted line is based on a version of Wilks theorem outside of model conditions.

10.3 Modeling income in Oslo

In this section, we will revisit the example of Section 4.1 with yearly income in Oslo and fit different models to the data set using maximum hybrid likelihood. We will use the concepts from Chapter 9 to choose, not only the balance parameter, but also which of the models to use. Afterwards, we will apply the limit results of Chapter 8 to make inference about the mean yearly income in Oslo.

To apply the hybrid likelihood theory, we need to decide on a parametric family of densities to fit to the data. In this section, we will fit and evaluate four different models: the family of log-normal, Weibull, gamma and Dagum densities. These are all commonly used when modeling income distributions. For an overview of literature regarding this and arguments for using, and a definition of, the Dagum distribution see Dagum 1977. To get a general idea of how well the different models work, we fitted them all using maximum likelihood. The resulting densities can be found in Figure 10.5 together with a histogram of the observations.

Looking at Figure 10.5, it is not clear what model is the best one. All seem to fit the data well enough and could be used for further analysis. In this example, however, we are only interested in investigating the mean yearly income in Oslo. The quality of this estimate is therefore more important to us than that of the overall model fit. Furthermore, each of the maximum likelihood estimates in the different models can be made more robust against model misspecification by using the hybrid likelihood theory developed in this

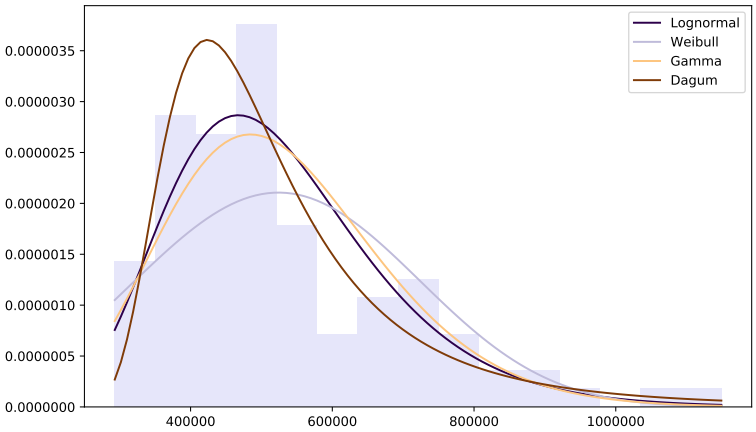


Figure 10.5: A histogram of the mean yearly income in the different sub-districts of Oslo together with estimated densities in four different models.

thesis. Different choices of models, control parameters and balance parameters will all result in different estimates and confidence intervals for the mean yearly income in Oslo. To choose one of them, we will use the focused information criterion for hybrid likelihood defined in the previous chapter.

We want to estimate the mean in the distribution of yearly income in Oslo. This quantity is therefore a natural choice for both control and focus parameter and was used for all the models in this example. To choose what model and balance parameter to use, we computed HFIC as defined in Definition 9.3.1 for each model and each value of a in a grid of points between 0 and 1. The results can be found in Figure 10.6.

Looking at Figure 10.6, we notice some interesting trends. For both the Weibull and Dagum model, we see that HFIC decreases as a function of a. In both these cases, we are therefore better off using pure empirical likelihood, rather than the more complicated hybrid likelihood strategy, to estimate the mean yearly income. The reason for this is that the maximum likelihood estimates of the mean in these two models, 527925 and 534479 respectively, are quite far off from the empirical mean: 529490. This indicates that the bias of the estimators are large, and hence, estimation would benefit greatly from the increased robustness of the empirical likelihood part of the hybrid likelihood function. On the other hand, we notice the opposite trend for the gamma model. The HFIC-curve is increasing as a function of a. Because of this, a standard maximum likelihood approach results in lower asymptotic mean squared error than a hybrid one. There is a good reason for this. As explained in Section 4.2, one of the entries in the score function in the gamma distribution is

$$y - \frac{\alpha}{\beta}$$

where α is the shape and β the rate parameter in the distribution. Hence, the

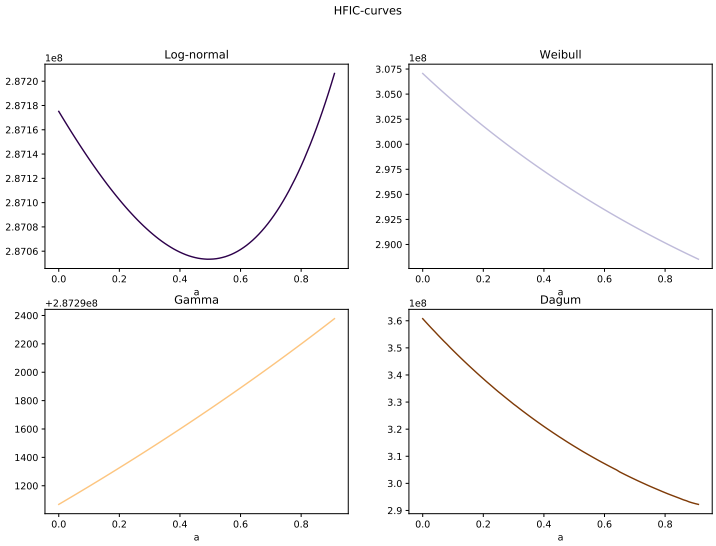


Figure 10.6: HFIC as functions of the balance parameter in the four different models fitted with maximum hybrid likelihood, using the mean as control parameter. The expected yearly income was used as focus parameter in computation of HFIC.

maximum likelihood estimator in a gamma model aims for (α_0, β_0), satisfying

$$0 = \mathrm{E}\left[Y - \frac{\alpha_0}{\beta_0}\right] \Leftrightarrow \frac{\alpha_0}{\beta_0} = \mathrm{E}\, Y.$$

The mean in a gamma distribution with shape parameter α and rate parameter β is α/β. So, because of the above, the maximum likelihood estimate of the mean in a gamma model is always consistent for the true expectation. Because of this, there is no need for additional robustness when using the gamma model to estimate means, and addition of a non-parametric part to the likelihood function does nothing but increase the variance of the maximizer.

For a close to 1, the corresponding HFIC-values in the different models are all quite similar. This is in agreement with theory as the hybrid likelihood function is almost equal to the empirical likelihood function for high values of the balance parameter. Hence, the resulting maximum hybrid likelihood estimates of the mean will be quite close to the maximizer of the empirical likelihood function when a is close to 1. The maximizer of the empirical likelihood function is the same for all models, and so estimates and confidence intervals for the mean in the different models will be similar when a is large. In particular, this ensures that the corresponding HFIC scores are almost identical.

Figure 10.6 provides one additional insight. The HFIC-curves corresponding to the log-normal model is completely dominated by those of the other families of distributions. Furthermore, it is minimized at $a = 0.49$. This indicates that, for estimation of the mean yearly income in Oslo, a log-normal model, fitted with hybrid likelihood using the mean as control and 0.49 as balance

parameter, should be chosen. It is important to remember that this choice is only optimal for estimation of the mean yearly income. There is nothing in the above analysis guaranteeing that the resulting log-normal model will be a good overall fit to the data, nor that the maximum hybrid likelihood estimate of other focus parameters will have low asymptotic mean squared error. In this example, however, our goal is to make inference about the mean yearly income in Oslo. The above analysis therefore suffices.

Now that we have decided on a parametric model, balance parameter and control parameter, we are ready to fit the log-normal model to the data using maximum hybrid likelihood. We used the ensuing parameterization of the log-normal model,

$$f_{\mu,\sigma}(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\, y - \mu)^2}{2\sigma^2}\right),$$

and obtained the following hybrid likelihood estimate of (μ, σ): $(13.1374, 0.2861)$. Using standard maximum likelihood, we got $(13.1364, 0.2859)$. The maximum hybrid likelihood estimate of the mean yearly income was computed to be 528810. This is slightly closer to the empirical mean, 529490, than the maximum likelihood estimate, 528248.

We can use the results of Chapter 8 to make inference about the mean yearly income in Oslo. Combining Theorem 8.2.2 with the delta method allows us to find the limit distribution of the maximum hybrid likelihood estimator of the mean yearly income. This can be used to construct confidence curves and intervals for the quantity. The procedure is very similar to what was done in Section 7.5 and will not be repeated here, but a plot of an approximate confidence curve can be found in Figure 10.7.

As in Section 10.2 the confidence curve obtained with maximum likelihood theory is constructed using the results of White 1983. Because of this, the curve is an approximate confidence curve for the true mean yearly income in Oslo when the model is specified correctly. If the data is not really log-normally distributed, however, the curve is an approximate confidence curve for $g(\mu_0, \sigma_0)$ where $g(\mu, \sigma)$ is the mean in a log-normal distribution with parameters μ and σ and (μ_0, σ_0) is the minimizer of the Kullback-Leibler divergence. The curve constructed using maximum hybrid likelihood can be interpreted similarly. If the model is specified correctly, confidence interval of the mean yearly income can be read off it. Otherwise the same holds true for $g(\mu_1, \sigma_1)$ where (μ_1, σ_1) is the minimizer of the distance function defined in Section 8.1.

The two curves in Figure 10.7 are very similar. This is a consequence of the quality of the maximum likelihood estimate of the mean yearly income. Since this is close to the empirical mean, the the effect of the empirical likelihood part of the hybrid likelihood function is minor. Because of this, there is not much difference between using maximum hybrid or standard likelihood to fit the log-normal model. This results in similar conclusions and, in particular, confidence curves that are almost indistinguishable.

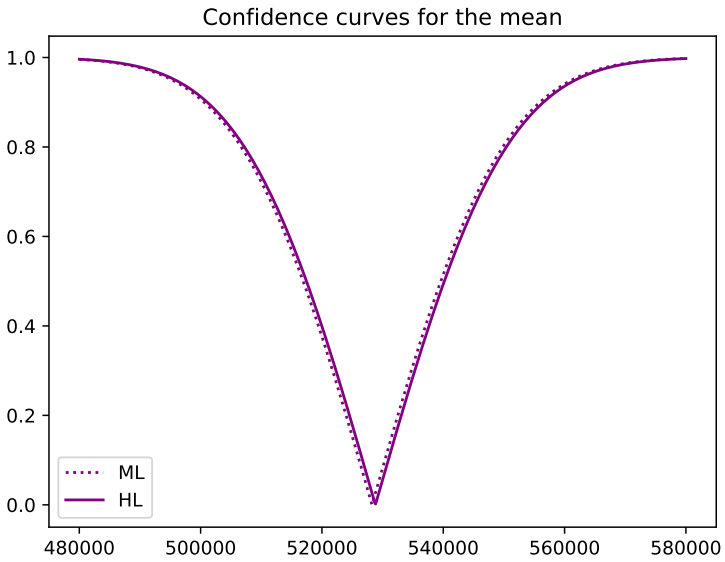


Figure 10.7: Approximate confidence curves for the mean yearly income in Oslo. The full drawn line is constructed with maximum hybrid likelihood theory. The dotted graph is based on maximum likelihood theory.

CHAPTER 11

Concluding remarks

A lot of the research regarding empirical likelihood has been focused on applications and adjustments of the concept to specific situations. This has not been the topic of this thesis. Instead, we have developed theory providing intuition about, as well as extensions of, the original concepts in Owen 1991. The results have also been applied in a hybrid setting, combining parametric and non-parametric methods.

The first achievement of the thesis was the statement and proof of Theorem 3.0.5. The result guarantees that the profile empirical likelihood function has a chi-squared limit at the true parameter. This can be used to construct approximate confidence intervals for a broader class of parameters than those described by Owen 1991 or Qin and Lawless 1994. In Chapter 4 we illustrated how the result can be applied with three examples.

In Chapter 5, we expanded on the ideas of Molanes Lopez, Van Keilegom, and Veraverbeke 2009 to develop an alternative characterization of the empirical likelihood function. The results of this chapter give insight to what goes on behind the scenes and more intuition about how the empirical likelihood function behaves. In particular, the alternative characterization reveals that the logarithm of the empirical likelihood function, divided by the sample size, is close to the mean of a certain function. In Chapter 6 this was used to show the results of Qin and Lawless 1994, involving consistency and asymptotic normality of the maximum empirical likelihood estimator, in a new way.

In the second part of the thesis, we left the realm of purely non-parametric statistics, and the remaining chapters were all dedicated hybrid likelihood. In Chapter 7, we summarized the already existing theory regarding this concept. In addition, we proved a profiling result and gave some examples of how hybrid likelihood theory can be applied.

Most of the results in Hjort, I. McKeague, and Van Keilegom 2018 can only be used when the true underlying distribution is a member of the parametric family fit to the data. In many cases, this is not the case or at least an assumption we do not want to make. In Chapter 8, we discarded the condition and investigated what happens under possible misspecification of the parametric model. The results of this chapter are useful in practice, as we can make model robust inference with them, but they also serve another purpose. The proofs in Chapter 8 were largely based on what was done in Chapter 5 and Chapter 6, and the analysis would not have been possible, had we not developed the alternative characterization of the empirical likelihood function in the first part of the

thesis. Hence, Chapter 8 can also be seen as an example of how the theoretical results of Chapter 5 and Chapter 6 can be used.

When working with the hybrid likelihood function, one needs to choose how much weight is put on the parametric and non-parametric part of the map. This is reflected by the choice of balance parameter, a, used in construction of the hybrid likelihood function. What the ideal value of a is, will depend on the situation and what we want to infer from the data. To use the hybrid likelihood function in practice, we need to develop a general way of choosing the balance parameter. In Chapter 9 we did just this. We found a fitting information criterion and derived a formula for computing this value in the context of hybrid likelihood estimation. We called the quantity HFIC, and in Section 9.4 we proposed choosing the balance parameter in a way minimizing the information criterion. Lastly, Chapter 10 was dedicated to examples illustrating and applying the results of Chapter 8 and Chapter 9.

In this thesis, we have only considered i.i.d. data. Such an assumption is convenient as it simplifies certain arguments, but many of the results can be lifted to non-i.i.d. situations with the sufficient mathematical efforts. In the remarks following the proof of Theorem 3.0.5 we discussed how this could be done for this particular theorem, but extensions of the other results in the thesis to e.g. regression settings, is certainly something that can be explored.

In Chapter 6 and Chapter 8 we assumed the estimating function was one dimensional. This was mathematically convenient, but it should be possible to extend the results to hold for multidimensional estimating functions as well. Rigorous arguments would requite more advanced matrix calculus than what we have considered in this thesis, but is something that deserves further investigation.

In addition to the above, the remaining assumptions of the theorems in Chapter 6 and Chapter 8 can probably be relaxed somewhat. Our proofs were largely based on theorems in Vaart 1998, but there are other ways to proceed. For instance, Hjort and Pollard 1993 can be used to show asymptotic normality of the maximum empirical and hybrid likelihood estimator when the corresponding function is concave. Furthermore, we showed convergence of the stochastic processes

$$A_n(s) = -2 \log \mathrm{EL}_n\left(\theta_0 + \frac{s}{\sqrt{n}}\right)$$

and

$$B_n(s) = h_n\left(\theta_0 + \frac{s}{\sqrt{n}}\right) - h_n(\theta_0)$$

to certain limits in this thesis. An alternative way of proving Theorem 6.3.2 and Theorem 8.2.2 could be to work with A_n and B_n directly, and use them to derive limit distributions of maximum empirical and hybrid likelihood estimators. This might lead to alternative conditions for convergence of the maximizers. For instance, limiting normality of maximum empirical likelihood estimators, when the estimating function is on the form $m(y, \mu) = I(y \leq \mu) - q$ for $q \in (0, 1)$, can be shown with arguments like these.

When proving limiting normality of the maximizer of the empirical and hybrid likelihood function, we only dealt with the case where the estimating

function was differentiable and the convergence of order $O_{\Pr}(1/\sqrt{n})$. Extensions of the results to non-smooth situations or other orders of convergence, could be interesting to investigate further. In Kim and Pollard 1990 and Vaart and Wellner 1996 the authors prove limiting normality of M-estimators in such situations. Similar arguments might be applicable for maximizers of empirical and hybrid likelihood functions. In particular, we believe the case with $m(y, \mu) = I(y \leq \mu) - q$, for $q \in (0, 1)$, could be interesting to investigate further.

There were ideas we had to let go due to time constraints. One such topic was mentioned in the remarks following the proof of Theorem 3.0.5. The convergence of the stochastic process A_n to the limit A, can be used for other purposes than what we have done in this thesis. One such example is, as mentioned above, to give yet another proof of asymptotic normality of the maximum empirical likelihood estimator. Profiling results for the sum of independent processes can also be proved rigorously using this fact and derivations analogous to those in the proof of Theorem 3.0.5. This was mentioned in Section 4.3 and Section 7.5, but a rigorous argument was not given.

In Section 5.4, we proposed approximating distributions with truncated versions to deal with the lack of solution to the equation

$$\begin{pmatrix} \mathrm{E}\{m(Y, \mu)/[1 + \lambda^T m(Y, \mu)]\} \\ \Pr[1 + \lambda^T m(Y, \mu) \leq 0] \end{pmatrix} = 0 \tag{11.1}$$

in the case where $m(Y, \mu)$ has unbounded support. In most applications, such an approximation is unproblematic, but the solution is not theoretically satisfying. One possible research topic could therefore be to work with a better way of dealing with the lack of solution to (11.1).

We have proved several limit results in this thesis and showed how they can be used to construct approximate confidence intervals and curves. As the number of observations grows to infinity, the results will be more and more accurate, but for small sample sizes, the approximations might work quite badly. In parametric likelihood theory, Barlett corrections are a way to deal with this. As mentioned before, Barlett correctability of the empirical likelihood function has been showed by DiCiccio, Hall, and Romano 1991, but we believe that similar corrections can be made for both the profile empirical likelihood function and the profile hybrid likelihood function, allowing for better approximation of their distributions. Arguing along the lines of chapter 7 in Schweder and Hjort 2016 is a possible approach.

Hybrid likelihood is not the only way to increase robustness of maximum likelihood estimates. There exists multiple robust parametric methods, and for a general overview, we refer to section 2.7 in Schweder and Hjort 2016. One approach that has particularly many similarities with hybrid likelihood is proposed by Basu et al. 1998. This article considers fitting a parametric model, f_θ for $\theta \in \Theta$, to data, following a distribution with a density, f, in a way that bears some resemblance to hybrid likelihood. The proposed estimator of θ aims at the minimizer of the divergence

$$d_a(f, f_\theta) = \int f_\theta(y)^{1+a} - \left(1 + \frac{1}{a}\right) f(y) f_\theta(y)^a + \frac{1}{a} f(y)^{1+a}\, dy.$$

As a goes to 0, d_a approaches the Kullback-Leibler divergence. For $a = 1$, d_a is L_2-loss. Hence, this method can be seen as a robust extension of maximum

likelihood estimation, with the parameter a deciding how much we should trust the parametric model. This idea is very similar to that of hybrid likelihood. In Section 8.1, we showed that the maximum hybrid likelihood estimator aims at the minimizer of the distance function

$$d_{a,\mu}(f, f_\theta) = (1-a)\mathrm{KL}(f, f_\theta) + a\mathrm{E}\,\log\{1 + \lambda[\mu(\theta)]m[Y, \mu(\theta)]\},$$

where, just as in Basu et al. 1998, a decides the trade-off between robustness and efficiency and $d_{0,\mu}$ equals the Kullback-Leibler divergence. It would be interesting to see how hybrid likelihood compares to the methods of Basu et al. 1998 and similar approaches. Studies like those of Jones et al. 2001 could be made.

In Chapter 9, we derived an asymptotically unbiased estimator of the mean squared error of maximum hybrid likelihood estimators of focus parameters. We chose to work with this particular loss function as it is mathematically easy to work with, in addition to being one of the most popular loss functions. The ideas presented in this chapter are, however, not limited to mean squared error, and generalizations of the information criterion to other loss functions is something that can be investigated further.

Additionally, the procedure we used in Section 10.3 to choose between different parametric models could use some additional investigation. Assume we have N candidate models, $M_1, \ldots, M_N$, and want to fit them by maximizing the corresponding hybrid likelihood functions. To choose what model and balance parameter to use, one could use HFIC. To do this, choose a focus parameter and compute the value of the corresponding information criterion for each model and each a in a grid of points between 0 and 1. Let HFIC_j denote the minimal value of HFIC for model j. To choose between $M_1, \ldots, M_N$, we can compare the values HFIC_j for $j = 1, \ldots, N$ and choose the model (and corresponding balance parameter) for which this estimate is the smallest. This procedure is easily implemented and makes intuitively sense, but warrants some additional investigation. The minimal values HFIC_j for $j = 1, \ldots, N$ are themselves stochastic, and hence, investigation of how well this method works in practice is certainly a topic to explore.

Bibliography

Anaconda Inc. (2020). *Anaconda Software Distribution*. Version 4.7.12. URL: https://docs.anaconda.com/.

Arcones, M. A. (1998). "Weak convergence of convex stochastic processes". In: *Statistics and Probability Letters* vol. 37, pp. 171–182.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). "Robust and efficient estimation by minimising a density power divergence". In: *Biometrika* vol. 85, pp. 549–559.

Billingsley, P. (1999). *Convergence of probability measures*. 2nd ed. New York: Wiley.

Campano, F. (2006). *Income distribution*. Oxford: Oxford University Press.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2006). "Variable Selection for Logistic Regression Using a Prediction-Focused Information Criterion". In: *Biometrics* vol. 62, pp. 972–979.

— (2007). "Prediction-focused model selection for autoregressive models". In: *Australian & New Zealand Journal of Statistics* vol. 49, pp. 359–379.

Claeskens, G., Cunen, C., and Hjort, N. L. (2019). "Model Selection via Focused Information Criteria for Complex Data in Ecology and Evolution". In: *Frontiers in Ecology and Evolution* vol. 7, p. 425.

Claeskens, G. and Hjort, N. L. (2003). "The Focused Information Criterion". In: *Journal of the American Statistical Association* vol. 98, pp. 900–916.

— (2008a). "Minimizing average risk in regression models". In: *Econometric Theory* vol. 24, pp. 493–527.

— (2008b). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Cunen, C., Hjort, N. L., and Nygård, H. M. (2020). "Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time". In: *Journal of Peace Research* vol. 57, pp. 221–234.

Dagum, C. (1977). "A New Model for Personal Income Distribution: Specificationand Estimation". In: *Economie Appliquée* vol. 30, pp. 413–437.

DiCiccio, T., Hall, P., and Romano, J. (1991). "Empirical likelihood is Bartlett-correctable". In: *The Annals of Statistics* vol. 19, pp. 1053–1061.

Eerkens, J. W. and Bettinger, R. L. (2001). "Techniques for Assessing Standardization in Artifact Assemblages: Can We Scale Material Variability?" In: *American Antiquity* vol. 66, pp. 493–504.

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

Guolo, A. and Adimari, G. (2010). "A note on the asymptotic behaviour of empirical likelihood statistics". In: *Statistical Methods & Applications* vol. 19, pp. 463–476.

Hjort, N. L. and Glad, I. K. (1995). "Nonparametric Density Estimation with a Parametric Start". In: *The Annals of Statistics* vol. 23, pp. 882–904.

Hjort, N. L., McKeague, I., and Van Keilegom, I. (2018). "Hybrid combinations of parametric and empirical likelihoods". In: *Statistica Sinica* vol. 28, pp. 2389–2407.

Hjort, N. L., McKeague, I. W., and Keilegom, I. V. (2009). "Extending the Scope of Empirical Likelihood". In: *The Annals of statistics* vol. 37, pp. 1079–1111.

Hjort, N. L. and Pollard, D. (1993). *Asymptotics for minimisers of convex processes*. Technical report, University of Oslo.

Huber, P. J. (2009). *Robust statistics*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons.

Jones, M. C., Hjort, N. L., Harris, I. R., and Basu, A. (2001). "A comparison of related density-based minimum divergence estimators". In: *Biometrika* vol. 88, pp. 865–873.

Jullum, M. and Hjort, N. L. (2017). "Parametric or nonparametric: The FIC approach". In: *Statistica Sinica* vol. 27, pp. 951–981.

Kim, J. and Pollard, D. (1990). "Cube root asymptotics". In: *The Annals of statistics* vol. 18, pp. 191–219.

Kitamura, Y. (1997). "Empirical Likelihood Methods with Weakly Dependent Processes". In: *The Annals of Statistics* vol. 25, pp. 2084–2102.

Kolaczyk, E. D. (1994). "Empirical likelihood for generalized linear models". In: *Statistica Sinica* vol. 4, pp. 199–218.

Kumagai, S. (1980). "An Implicit Function Theorem: Comment". In: *Journal of Optimization Theory and Applications* vol. 31, pp. 285–288.

Lindstrøm, T. L. (2017). *Spaces : An Introduction to Real Analysis*. Providence, R.I: American Mathematical Society.

McDonald, J. N. and Weiss, A. N. (2013). *A Course in Real Analysis*. 2nd ed. Amsterdam: Academic Press.

Molanes Lopez, E. M., Van Keilegom, I., and Veraverbeke, N. (2009). "Empirical Likelihood for Non-Smooth Criterion Functions". In: *Scandinavian journal of statistics* vol. 36, pp. 413–432.

Mykland, P. A. (1995). "Dual likelihood". In: *The Annals of Statistics* vol. 23, pp. 396–421.

Olkin, I. and Spiegelman, C. H. (1987). "A Semiparametric Approach to Density Estimation". In: *Journal of the American Statistical Association* vol. 82, pp. 858–865.

Owen, A. B. (1988). "Empirical likelihood ratio confidence intervals for a single functional". In: *Biometrika* vol. 75, pp. 237–249.

— (1991). "Empirical likelihood for linear models". In: *The Annals of Statistics* vol. 19, pp. 1725–1747.

— (2001). *Empirical Likelihood*. London: Chapman & Hall/CRC.

Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Toronto: Viking.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Qin, J. and Lawless, J. (1994). "Empirical likelihood and general estimating equations". In: *The Annals of Statistics* vol. 22, pp. 300–325.

Qin, J. (1994). "Semi-empirical likelihood ratio confidence intervals for the difference of two sample means". In: *Annals of the Institute of Statistical Mathematics* vol. 46, pp. 117–126.

— (2000). "Combining parametric and empirical likelihoods". In: *Biometrika* vol. 87, pp. 484–490.

Qin, J. and Wong, A. (1996). "Empirical Likelihood in a Semi-Parametric Model". In: *Scandinavian Journal of Statistics* vol. 23, pp. 209–219.

Sarkees, M. R. and Wayman, F. (2010). *Resort to War: 1816 - 2007*. Washington DC: CQ Press.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability : Statistical Inference with Confidence Distributions*. Cambridge: Cambridge University Press.

Shumway, R. H. and Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. 4th ed. Cham: Springer International Publishing AG.

Statistikkbanken (2020). *Gjennomsnittsinntekt etter delbydel, alder og kjønn (D)*. URL: https://statistikkbanken.oslo.kommune.no/webview/ (visited on 11/10/2020).

Stute, W. (1982). "The oscillation behavior of empirical processes". In: *The Annals of Probability* vol. 10, pp. 86–107.

Vaart, A. v. d. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Vaart, A. v. d. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes : With Applications to Statistics*. New York: Springer.

White, H. (1983). "Maximum Likelihood Estimation of Misspecified Models". In: *Econometrica* vol. 51, pp. 513–513.

Zhang, X. and Liang, H. (2011). "Focused information criterion and model averaging for generalized additive partial linear models". In: *The Annals of Statistics* vol. 39, pp. 174–200.