

UNIVERSITY
OF OSLO

Ingrid Dæhlen

The best of the bad

Precise and robust estimation and model selection

**Thesis submitted for the degree of
Philosophiae Doctor**

Department of Mathematics
Faculty of Mathematics and Natural Sciences

Supervisors: Ingrid Hobæk Haff , Nils Lid Hjort

2024



Abstract

This thesis concerns the topic of robust and precise estimation and model selection. New ways of precisely estimating errors are derived. Model selection is studied in a specific situation and in a general context. In more detail, approximations to the bias and mean squared error of a large class of estimators are derived. This is done in a precise way by making sure that the expected error is smaller than the variance of the estimator. Furthermore, a model fitting technique combining parametric and nonparametric methods is investigated in detail, and a technique for automatically deciding on how to balance the two methods is given. Lastly, the asymptotic effect of certain model selection procedures is studied when the model selection is achieved by minimising an estimator of the risk. This risk estimator can be quite general, including approximations based on information criteria and cross validation.

Sammendrag

Denne avhandlingen handler om robust og presis estimering og modellseleksjon. Det blir utledet nye måter å estimere feil på en presis måte. Modellseleksjon blir etterforsket i både en spesiell situasjon og mer generelt. Tilnærminger til forventningsskjevheten og gjennomsnittlig kvadrert feil blir utledet for en omfattende klasse av estimatorer. Dette blir gjort på en presis måte ved å sørge for at den forventede feilen er mindre enn variansen til estimatoren. Videre studeres en modelltilpassingsmetode som kombinerer parametri og ikke-parametri i detalj, og det blir gitt en metode for å automatisk velge hvordan de to teknikkene bør balanseres. Til slutt etterforskes den asymptotiske effekten av enkelt modellseleksjonsprosedyrer når modellseleksjonen blir gjort ved å minimere en risikoestimator. Denne risikoestimatoren kan være temmelig generell, inkludert tilnærminger basert på informasjonskriterier og kryssvalidering.

Contents

1	Introduction	1
2	M- and Z-estimators	3
2.1	Z-estimators.	3
2.2	M-estimators	8
2.3	When things are not sufficiently regular	10
2.4	Two-step estimators	13
2.5	Influence functions	14
3	Estimating errors.	17
3.1	Loss and risk functions	17
3.2	Estimating risk.	23
4	Summary of papers	31
4.1	Paper I - Accurate bias estimation with applications to focused model selection	31
4.2	Paper II - Model robust hybrid likelihood	32
4.3	Paper III - The asymptotic effect of tuning parameters	33
5	Discussion	35
5.1	Can precision be precisely estimated?	35
5.2	Accounting for every little thing	41
5.3	Smoothing over the hard parts	43
5.4	Concluding remarks	50
	Bibliography	51

Paper I Accurate bias estimation with applications to focused model selection	59
--	-----------

Paper II Model robust hybrid likelihood	97
--	-----------

Paper III The asymptotic effect of tuning parameters	121
---	------------

Appendices	155
-------------------	------------

A	Supplementary material for Paper II	157
B	Supplementary material for Paper III	167

Contents

List of Figures

1	The figure shows a randomly drawn data set from an LDA model together with the decision boundaries obtained from fitted LR and LDA models.	37
2	The figure shows the empirical variance of $\sqrt{n}\hat{\beta}(\hat{\lambda})$ based on $B = 10\,000$ samples, together with the theoretical limit as a function of the sample size n . In the figure, $V_{j,k}$ refers to the covariance between the j -th and k -th component of $\sqrt{n}\hat{\beta}(\hat{\lambda})$	44
3	The figure shows the mean value of $n^\alpha \hat{\theta}_n - \theta_0 $ for $\alpha = 1/3, 1/2$ and different values of n over $B = 1000$ iterations in a simulation experiment. Here $\hat{\theta}_n$ is the maximiser of the hybrid likelihood function and θ_0 the theoretical limit of the estimator.	49

List of Figures

Preface

In preparation for writing this thesis, I skimmed through multiple other thesis handed in at UiO. In more than one of these, the authors wrote something along the lines of “This has been hard. I wanted to give up many times.”, so I know that I am not surprising anyone when I now write that this has been extremely challenging. I have tried so many things which did not work out, and my hard disc contains at least three unfinished manuscripts of 15+ pages and so, so many thousands of lines of code. I have cried from rejection letters as well as errors in a proof, but I have also excitedly shown off of my simulation results and been high on adrenaline after giving good talk. Doing a PhD has without doubt been extremely frustrating, but also fun and rewarding, and I believe that trying many things which ultimately not work out, has made me more critical and knowledgeable than I would otherwise have been. I am very grateful to the mathematics department at UiO and to BigInsight for financing this PhD, letting me devote three whole years to mathematics. Thank you also to the statistic section at UiO for all the great discussions and talks, and to the statistics and probability group at Nanterre University who made my three month long exchange in Paris such a fun and enriching experience.

No one deserves more gratefulness, however, than my two supervisors (and co-authors) Ingrid Hobæk Haff and Nils Lid Hjort. Thank you for being my academic inspirations as well as kind and warm people. At the worst times, I would have given up were it not for your supportive words. Thank you to Nils for answering my weekend emails and never failing to believe that our papers will “pass muster”. Thank you to Ingrid for reading all of my manuscripts so carefully and for being willing to sporadically discuss with me on often very short notice.

I also want to thank all of my wonderful friends. First and foremost I am grateful to my fellow PhD students at the university of Oslo. I am of course thankful for you always double checking my maths when I needed it and for impressing me with your own research, but thank you also for making me laugh during lunch time, for going swimming at conferences and for drinking hot chocolate on Fridays. Thank you to my great gym buddies. Both for forcing me to do extra repetitions and for making sure I always **belive in myself**, and thank you to my drawing group for coming with me to kroki every Wednesday. Thanks to you I have filled three sketchbooks and finished one Inktober as well as 2/3 of a comic during these past three years. Thank you also to my two role playing groups who never fail to make me laugh. Who would have thought that pretending to be an unsuccessful academic in a fantasy world would ease the stress of doing a PhD.

Preface

I also want to thank my family for being a constant support in my life. Without your patience and kindness, I do not know who I would have been. Thank you to my mother and step father for feeding me Rugull and to my dad for the countless hours we have talked on the phone. Thank you also to my family in law for boosting my self-esteem by asking for my advice, and for lowering it by failing to solve crimes faster than Sherlock Holmes together with me. Last but not least, thank you to my boyfriend Lars. You are the reason I have functioned these years, and I am sorry my stress woke you up all those nights.

Chapter 1

Introduction

“All models are wrong (...)” states Box 1979, p. 2, meaning that all statistical models are approximations to reality and cannot ever be completely accurate. He then goes on to write “(...) but some are useful”, indicating that models provide intuition and explanations making them convenient in practical situations. For this to be the case, however, one cannot use any arbitrary model. For a model to be useful, one needs it to be “good”, a requirement which is far from well-defined. A model might fit the data well overall, but work poorly when it comes to estimation of quantities of interest. On the other hand, a model might be well suited to solve a specific problem even though its ability to fully describe the data generating procedure is lacking. Is it then a good model, or is only the procedure for solving the task at hand good? To be able to find “good” models, one needs to define just what is meant by this very ambiguous term and rank all candidates accordingly. This thesis attempts to aid in the search for good models by deriving new ways for them to be evaluated and fitted, or as its subtitle says, to develop methods of “precise and robust estimation and model selection”. **Paper I** does this by deriving a new information criterion ranking models and estimators, **Paper II** investigates a way of fitting models, making them more useful in practice. **Paper III**, takes a step back and tries to evaluate just how much all of these rankings and selections affect analysis, allowing for the selection mechanisms of the previous papers to be included in the estimating procedure.

Another ill-defined theme of this thesis can also be found in its subtitle. This theme is precision, which can be interpreted as having small errors. For this definition to be sound, one needs to first make clear what is meant by “errors”. If a nonsensical method yields the correct answer by chance, is it then correct? And if an answer is the right one to an ill-posed question, is the error small or large? When Galileo Galilei said that all objects fall to the ground at the same speed, was he wrong? Or did he simply ask the wrong question by failing to account for air resistance. To be able to talk about precision, one needs to properly define what is meant by the term “error”. In **Paper I** and **II**, this is addressed by measuring error in terms of the mean squared error (MSE), while **Paper III** considers more general errors than the MSE, and in particular cross-validated estimators thereof. To properly define the term “precise”, however, it is not enough to characterise errors. One also needs to decide how small they are required to be. In an ideal

Chapter 1. Introduction

world, the error of every estimator would be zero, and precise estimators would be error-free. Sadly, obtaining estimators with zero errors is rarely possible in practice. Instead, one typically requires the errors to be “close enough” to zero for an estimator or method to be precise. This approach is taken in this thesis, and in **Paper I**, **Paper III** as well as Chapter 3 it is carefully worked out precisely how small the errors need to be.

I will start by providing some mathematical background as well as historical overview of the most central topics to this thesis (Chapter 2). Equipped with these mathematical tools, I will move on to the concept of errors, precision and model selection in Chapter 3. Afterwards brief summaries of the papers in this thesis will be given (Chapter 4), and lastly, in Chapter 5 I will discuss whether this thesis achieves its goal of studying precise and robust estimation and model selection.

Chapter 2

M- and Z-estimators

If there is one single theme connecting all the articles in this thesis, it is the use of so called M- and Z-estimators. M- and Z-estimators are maximisers of means and roots of means respectively. Although perhaps not obvious at first glance, these two, slightly overlapping, classes turn out to be extremely wide and comprising many of the most popular approximations used in statistics, e.g. means, medians, regression coefficients and maximum likelihood estimators. Because of this, M- and Z-estimators are widely studied in literature and most textbooks in statistics contain definitions and properties for both classes. Since these estimators are so crucial to this thesis, I will start by providing definitions and stating some classical results. Most of the material I will cover is a restated version of what can be found in textbooks on asymptotic statistics, but my presentation is greatly inspired by van der Vaart 2000, Ch. 5.

2.1 Z-estimators

Z-estimation has a long history. The method of moments, which seeks to estimate the parameters in a distribution by matching theoretical with empirical moments, can be seen as a special case of Z-estimation and dates back to at least Pearson 1894¹. Not too much later Ronald Fisher published his first works on the method of maximum likelihood (Fisher 1912, 1922), which also can be reformulated as a Z-estimation problem provided certain conditions hold true. Since then, the concept has been widely studied, with Durbin 1960 and Godambe 1960 giving the first proper definitions and theorems. Furthermore, multiple extensions and simplification of the theory have been made. The generalised method of moments pioneered by Hansen 1982 has for instance found huge success within the field of econometrics (see Bera and Biliias 2002) and is closely related to the theory of Z-estimators. The same can be said about the empirical likelihood method,

¹The mathematics of this paper was extremely complicated involving, among other things, solutions to a ninth-degree equation, and according to Stigler 1990, p. 332, a referee wrote “I regret that I am not competent to give a useful opinion on the soundness of Prof. Karl Pearson’s paper. But assuming its mathematics to be correct (as is most reasonable to expect) I should recommend its publication in the Transactions as a vigorous and original attempt to supply a statistical want. ”.

developed by Owen 1988, 1990, 1991, which indeed is the main topic of **Paper II** in this thesis. I will now informally present and derive some properties of Z-estimators, but I will focus on themes relevant to this thesis, and a more thorough introduction can be found in e.g. Shao 2003, Ch. 5. See also Bera et al. 2006 or the previously mentioned Bera and Bilias 2002 for excellent reviews on the history of Z-estimators and related methods.

For data $Z_1, \dots, Z_n \in \mathbb{R}^d$, a Z-estimator is defined as the root of a function on the form $\Phi_n(\theta) = \sum_{i=1}^n \varphi(Z_i, \theta)$ where φ is some fixed map from $\mathbb{R}^{d+p}$ to $\mathbb{R}^q$, called the estimating function. The equation $0 = \Phi_n(\theta)$ is sometimes called a generalised estimating equation (GEE) and Z-estimators therefore also go by the name GEE estimators. The function φ can be chosen freely, and different maps give different estimators. With, for instance, $\varphi(z, \theta) = z - \theta$, the corresponding Z-estimator is the mean of the data, while $\varphi(x, y, \beta_0, \beta) = (y - \beta_0 - x^T \beta)(1, x^T)^T$ has the ordinary least squares regression estimator as the solution to $0 = \Phi_n(\beta_0, \beta)$, when $Z_i = (Y_i, X_i)^T$. Furthermore, when n is even, the median satisfies $0 = \Phi_n(\theta)$ with $\varphi(z, \theta) = I(y \leq \theta) - 1/2$, where I is the indicator function.

If $Z_1, \dots, Z_n$ are independent and identically distributed (i.i.d.) from some distribution F , the law of large numbers (LNN) guarantees $\Phi_n(\theta) \xrightarrow{p} E \varphi(Z, \theta) = \Phi(\theta)$, where $Z \sim F$ and $\xrightarrow{p}$ denotes convergence in probability. Because of this, one would expect that for large n , $\hat{\theta}_n$, the solution to $\Phi_n(\theta) = 0$, is close to solving the limiting equation $0 = \Phi(\theta)$. If, furthermore, $\Phi(\theta) = 0$ has a unique solution $\theta_0 \in \mathbb{R}^p$, it would therefore be reasonable to assume $\hat{\theta}_n \xrightarrow{p} \theta_0$. This intuition turns out to be correct, as $\hat{\theta}_n$ is consistent for θ_0 provided φ and F are “regular enough” (see for instance van der Vaart 2000, Th. 5.9).

Consistency of $\hat{\theta}_n$ is a very important result as it guarantees that the intuitive idea of approximating the solution to the limiting equation by solving the empirical counterpart is sound. As the root of $\Phi_n(\theta)$ does converge in probability towards θ_0 , one can safely use $\hat{\theta}_n$ to estimate the zero of the limit equation. If we, however, also want to know something about how “far off” our estimator is, a little more work is required. Assume φ is twice continuously differentiable in a neighbourhood of θ_0 , then a Taylor expansion of $\Phi_n(\theta)$ around the root of the limit equation reveals

$$0 = \Phi_n(\hat{\theta}_n) \approx \Phi_n(\theta_0) + \Phi'_n(\theta_0)(\hat{\theta}_n - \theta_0), \quad (2.1)$$

and hence $\hat{\theta}_n - \theta_0 \approx -\Phi'_n(\theta_0)^{-1} \Phi_n(\theta_0)$, where $\Phi'_n(\theta_0)$ is the Jacobian matrix of Φ_n evaluated at θ_0 . By definition, $\Phi'_n(\theta_0) = \sum_{i=1}^n \partial_{\theta_0} \varphi(Z_i, \theta)$ where ∂_{θ^*} is short hand for the partial derivative with respect to θ evaluated at θ^* . Hence, $n^{-1} \Phi(\theta_0)$ is a mean which converges to $E \partial_{\theta_0} \varphi(Z, \theta) = \Phi'(\theta_0)$ by the LNN and Leibniz integral formula. Furthermore, the central limit theorem (CLT) guarantees that $\sqrt{n} \Phi_n(\theta_0) = n^{-1/2} \sum_{i=1}^n \varphi(Z_i, \theta_0)$ converges in distribution² towards a central normal variable with variance $\text{Var} \varphi(Z, \theta_0)$. Combining this with Slutsky’s lemma and (2.1) ensures $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\{0, \Phi'(\theta_0)^{-1} \text{Var} \varphi(Z, \theta_0) [\Phi'(\theta_0)^{-1}]^T\}$. I state these findings in the following theorem. For formal proofs, consult van der Vaart 2000.

²Convergence in distribution is sometimes called convergence in law. I will use both terms interchangeably and denote it by $\xrightarrow{d}$.

Theorem 1. Let $\hat{\theta}_n$ be the solution to $\sum_{i=1}^n \varphi(Z_i, \theta) = 0$ and θ_0 the root of $\Phi(\theta) = E \varphi(Z, \theta)$, when $Z, Z_1, \dots, Z_n$ are i.i.d.. Then

$$(1) \hat{\theta}_n \xrightarrow{P} \theta_0, \text{ and}$$

$$(2) \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, V(\theta_0)] \text{ where } V(\theta) = \Phi'(\theta)^{-1} \text{Var} \varphi(Z, \theta) [\Phi'(\theta)^{-1}]^T,$$

provided the regularity assumptions of van der Vaart 2000, Th. 5.9 and Th. 5.21 hold true.

With $\varphi(z, \theta) = z - \theta$, (1) of Theorem 1 reduces to the LNN and (2) to the CLT. Before I continue on with M-estimators I will go through two slightly more complicated applications of Theorem 1.

2.1.1 The OLS estimator - a sanity check

For the ordinary least squares (OLS) estimator $\hat{\beta}_n$, assume $X \in \mathbb{R}^d$ follows some distribution with mean value μ and covariance matrix Σ , and that for a pair $(Y, X) \sim F$, $Y = \alpha_0 + \alpha^T X + \epsilon$ where ϵ is a random variable with mean 0 and standard deviation σ , independent of X . Then $\hat{\beta}_n$ is a Z-estimator with $\varphi(y, x, \beta_0, \beta) = (y - \beta_0 - \beta^T x)(1, x^T)^T$ and

$$\Phi(\beta_0, \beta) = \begin{pmatrix} \alpha_0 - \beta_0 + \mu^T(\alpha - \beta) \\ (\alpha_0 - \beta_0)\mu + (\alpha - \beta)(\Sigma + \mu\mu^T) \end{pmatrix},$$

which has $\beta_0 = \alpha_0$ and $\beta = \alpha$ as a root. Hence the OLS estimator is consistent for the “true” coefficients $(\alpha_0, \alpha^T)^T$. Furthermore, direct computation shows

$$\Phi'(\alpha_0, \alpha) = - \begin{pmatrix} 1 & \mu^T \\ \mu & \Sigma + \mu\mu^T \end{pmatrix}$$

and

$$\text{Var} \varphi(Z, \alpha_0, \alpha) = \text{Var} \begin{pmatrix} \epsilon \\ \epsilon x \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \mu^T \\ \mu & \Sigma + \mu\mu^T \end{pmatrix},$$

showing that the OLS estimator is approximately normally distributed with mean $(\alpha_0, \alpha^T)^T$ and covariance $-\sigma^2 \Phi(\alpha_0, \alpha)^{-1}/n$.

In introductory statistics courses, one learns that the OLS estimator is asymptotically normal with mean equal to the true regression coefficients and variance $\sigma^2 (X^T X)^{-1}$, where X is the $n \times (d+1)$ -design matrix with $(1, X_i^T)$ as its i -th row (see e.g. Hastie et al. 2009, Se. 3.2 or any introductory textbook in statistics). By the LNN, $n^{-1} X^T X \xrightarrow{P} -\Phi'(\alpha_0, \alpha)$, and hence

$$\sigma^2 (X^T X)^{-1} = n^{-1} \sigma^2 (n^{-1} X^T X)^{-1} \approx -n^{-1} \sigma^2 \Phi'(\alpha_0, \alpha)^{-1}.$$

Hence, Theorem 1 agrees with the classic results.

It is worth noting that in order for Theorem 1 to be applicable, the data $Z_1, \dots, Z_n$ need to be i.i.d.. This forces us to deviate somewhat from the classic regression setup, where the covariates are treated as fixed and non-random

quantities. There are ways to circumvent this. For instance, one can use a CLT for non-i.i.d. sequences to show a version of Theorem 1 for independently, but not equally distributed data points.³ See e.g. Shao 2003, Ch. 5 for an overview of such results. In my opinion, however, treating the covariates as random variables is oftentimes unproblematic. Participants in a survey, people filing insurance claims or costumers at a streaming service are usually not picked out in a targeted way, but rather drawn randomly from the relevant population. I therefore fail to see any real mathematical or philosophical reason to treat the covariates as fixed in many commonly encountered situations. Furthermore, as is shown in Rosset and Tibshirani 2020, the true prediction error is larger when the covariates are random, and hence, ignoring the distribution of the X_i s can lead to over-optimistic inference if the covariates really are not fixed.

2.1.2 Maximum likelihood estimation

Although some preliminary ideas were given by other scholars e.g. Edgeworth 1908, 1909, maximum likelihood theory as we know it today was developed by Ronald Aylmer Fisher in Fisher 1912, 1922. The method is one of the perhaps most famous and influential statistical techniques and is based on analysing the likelihood function L_n defined as

$$L_n(\theta) = \prod_{i=1}^n f_\theta(Z_i),$$

for a given parametric family $\mathcal{F}_\theta = \{f_\theta \text{ for } \theta \in \Theta \subseteq \mathbb{R}^p\}$ and i.i.d. data $Z_1, \dots, Z_n$. The maximiser of L_n is called the maximum likelihood estimator (ML estimator), and typically, working with $\ell_n(\theta) = \log L_n(\theta)$ is simpler than the likelihood function itself. With these definitions, the ML estimator is equal to

$$\hat{\theta}_n = \operatorname{argmax}_{\theta} \ell_n(\theta) = \operatorname{argmax}_{\theta} \sum_{i=1}^n \log f_\theta(Z_i).$$

If $\hat{\theta}_n$ is an interior point of Θ and ℓ_n is smooth, the ML estimator satisfies $0 = \sum_{i=1}^n u(Z_i, \hat{\theta}_n)$, where $u(z, \theta) = \nabla_\theta \log f_\theta(Z_i)$ is called the score function and ∇_{θ_*} denotes the gradient evaluated at θ_* . Hence, the ML estimator is a Z-estimator with $\varphi = u$ in the notation of this chapter. Assume furthermore that F has a density function (pdf) or probability mass function (pmf) g . Then, again due to the smoothness condition, solving $0 = \Phi(\theta)$ is the same as maximising $E \log f_\theta(Z)$, or equivalently minimising

$$\text{KL}(g, f) = E_g \log g(Y)/f_\theta(Y), \quad (2.2)$$

where the subscript in the expectation indicates with respect to which measure one integrates. (2.2) is called the Kullback-Leibler divergence (KL divergence) from the true pdf/pmf g to f_θ , and will be discussed in more depth in Section 3.1.2.

³An option is the Lindeberg theorem, first shown in Lindeberg 1922. See e.g. Shao 2003, p. 67 for a more readable (and less German) presentation.

For now, however, it suffices to say that $\text{KL}(g, f)$ can be thought of as a measure of how “far off” f is from g . Hence, θ_0 , the solution to $0 = \nabla_{\theta} \mathbb{E} \log f_{\theta}(Z)$, is the parameter for which f_{θ} is “closest” to the true density g .

If the true density g is equal to f_{θ_*} for some $\theta_* \in \Theta$, g is a member of the parametric class. I will refer to this as the model being correctly specified. In this case $\theta_0 = \theta_*$ (see e.g. van der Vaart 2000, le. 5.35), which is quite natural as the “closest” member of $\mathcal{F}_{\theta}$ to f_{θ_*} should be f_{θ_*} itself. If, on the other hand, there is no $\theta \in \Theta$ such that $g = f_{\theta}$, interpretation of θ_0 becomes slightly more difficult. By the above arguments, θ_0 is the minimiser of the KL divergence from $\mathcal{F}_{\theta}$ to the true density g , and hence, θ_0 can be thought of as the parameter for which f_{θ} is as close as possible to the truth. This intuitive explanation was given by Akaike 1973, and motivated by it, I will do as Schweder and Hjort 2016 and call θ_0 the least false parameter or lf-parameter for short.

Although some theoretical properties of the ML estimator outside of model conditions were investigated by others, e.g. Huber 1967, White 1982 popularised model agnostic maximum likelihood theory by giving a unified framework for the field. Furthermore, he gave explicit and consistent estimators for the limiting variance of the ML estimator, as well as a test for checking model assumptions. In addition, White’s paper provided much needed intuition by emphasising how the model agnostic results simplify to the classic theorems of maximum likelihood theory when the parametric model fitted is correctly specified. This laid the groundwork for model-agnostic maximum likelihood inference, a topic which is extremely relevant to all papers in this thesis and in particular to the focused information criterion, which I will discuss in Section 3.2.2.

Provided everything is sufficiently regular, Theorem 1 guarantees that the ML estimator $\hat{\theta}_n$ converges in probability towards θ_0 , and furthermore, that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, V(\theta_0)]$, with V as in the theorem. For a general function h , let $H_{\theta_0}h(Z, \theta)$ denote the Hessian matrix of h at θ_0 . Then, by definition, $\partial_{\theta_0} \mathbb{E} u(Z, \theta) = H_{\theta_0} \mathbb{E} \log f_{\theta}(Z)$. The expression on the right hand side of this equation is the negative of the Fisher information matrix, which I will denote by $J(\theta_0)$. Further, let $K(\theta)$ be the variance matrix of $u(Z, \theta)$, Theorem 1 then ensures that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, J(\theta_0)^{-1}K(\theta_0)J(\theta_0)^{-1}]. \quad (2.3)$$

Due to Theorem 1, results like (2.3), where the limiting variance can be written as a matrix product of a covariance matrix “sandwiched” between two Jacobian or Hessian matrices, are very common in robust statistics, and both **Paper II** and **Paper III** derive their own counterparts to (2.3) with well-defined “least false” parameters.

If the parametric model is specified correctly, θ_0 is the value such that $g = f_{\theta_0}$, by the previous arguments. Ignoring the tedious job of checking all regularity

conditions, one finds

$$\begin{aligned}
 -J(\theta_0) &= \partial_{\theta_0} \int_{\mathbb{R}^d} \frac{\nabla_{\theta} f_{\theta}(z)}{f_{\theta}(z)} f_{\theta_0} \, dz \\
 &= \int_{\mathbb{R}^d} \frac{H_{\theta_0} f_{\theta_0}(z) f_{\theta_0}(z) - \nabla_{\theta_0} f_{\theta_0}(z) \nabla_{\theta_0} f_{\theta_0}(z)^T}{f_{\theta_0}(z)^2} f_{\theta_0} \, dz \\
 &= \int_{\mathbb{R}^d} H_{\theta_0} f_{\theta_0}(z) \, dz - \int_{\mathbb{R}^d} \left(\frac{\nabla_{\theta_0} f_{\theta_0}(z)}{f_{\theta_0}} \right) \left(\frac{\nabla_{\theta_0} f_{\theta_0}(z)}{f_{\theta_0}} \right)^T f_{\theta_0}(z) \, dz.
 \end{aligned}$$

The first term is equal to zero by Leibniz' integral formula and the fact that $1 = \int_{\mathbb{R}^d} f_{\theta}(z) \, dz$ for all θ . The second term is the expected value of $E u(z, \theta_0) u(z, \theta_0)^T$, and by definition $E u(z, \theta_0) = 0$. Hence, the above informal calculations ensure $J(\theta_0) = K(\theta_0)$ under model conditions, showing that (2.3) reduces to the classic result $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, J(\theta_0)^{-1}]$ when $g = f_{\theta_0}$.

It can, of course, happen that the maximum likelihood estimator maximises the log-likelihood function ℓ_n without solving $\ell'_n(\hat{\theta}) = 0$. This can for instance happen if ℓ_n is not sufficiently smooth or the maximiser lands on the boundary of the parameter space. In such cases, the ML estimator is no longer a Z-estimator and the theory outlined in this section does not apply. All hope is not lost, however, as $\hat{\theta}_n$ might still be an M-estimator, for which analogue results hold true. I will now move on to discuss this class of estimators.

2.2 M-estimators

The name M-estimator was introduced in Huber 1964, and refers to estimators which can be written as maximisers of expressions on the form $\Psi_n(\theta) = \sum_{i=1}^n \psi(Z_i, \theta)$. Huber himself was motivated by robust statistics when writing his 1964 paper, and the concept has therefore gained a lot of popularity within this field. See e.g. Elsenner and van de Geer 2018 or Prasad et al. 2020 for relatively recent work building on M-estimation, or Ronchetti 2021 for a short review on robust statistics. The book Huber and Ronchetti 2009, contains theory regarding M-estimators as well as applications to robust statistics.

In Section 2.1.2, I argued that if the parametric model $\log f_{\theta}$ is sufficiently smooth, ML estimators can be seen as Z-estimators, making Theorem 1 applicable. This does, of course, hold true for other M-estimators than the ML one. If ψ is sufficiently smooth and regular, maximising $\Psi_n(\theta)$ is equivalent to solving $0 = \sum_{i=1}^n \nabla_{\theta} \psi(Z_i, \theta)$. Hence, in most cases, M-estimators are also Z-estimators and all the theory outlined in the previous section hold true by replacing φ with $\nabla_{\theta} \psi$ and Φ with Ψ . The theory in both **Paper II** and **Paper III**, hinges on such arguments, and limit results are derived by finding conditions which “force” M-estimators into being Z-estimators.

That being said, it is not always the case that the maximiser of Ψ_n will satisfy $\Psi'_n(\hat{\theta}_n) = 0$. If, for instance, Ψ_n is not differentiable, Ψ'_n is not defined, and the maximisation problem does not have a Z-estimation counterpart. Luckily, it turns out that assuming smooth differentiability is slightly too strict of an assumption. This is stated in the following theorem.

Theorem 2. Let $\hat{\theta}_n$ be the maximiser of $\Psi_n(\theta) = \sum_{i=1}^n \psi(Z_i, \theta)$ and θ_0 the maximiser of the limit function $\Psi(\theta) = \mathbb{E}\psi(Z, \theta)$ when $Z, Z_1, \dots, Z_n$ are i.i.d.. Then as long as ψ is differentiable at θ_0 and there exists a function m such that $\mathbb{E}m(Z)^2 < \infty$ and

$$|\psi(z, \theta_1) - \psi(z, \theta_2)| \leq m(z)\|\theta_1 - \theta_2\| \quad (2.4)$$

for all θ_1, θ_2 in a neighbourhood of θ_0 and almost all z ,

$$(1) \hat{\theta}_n \xrightarrow{p} \theta_0, \text{ and}$$

$$(2) \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, V(\theta_0)], \text{ where}$$

$$V(\theta_0) = [H_{\theta_0}\Psi(\theta)]^{-1} \text{Var} \nabla_{\theta_0}\psi(Z, \theta)[H_{\theta_0}\Psi(\theta)^T]^{-1}$$

provided the additional regularity assumptions of van der Vaart 2000, Th. 5.7 and Th. 5.23 hold true.

At first glance, Theorem 2 might not look too impressive. For the result to be applicable, one still needs some sort of smoothness as existence of $\nabla_{\theta_0}\psi(z, \theta)$ is assumed. Furthermore, the Lipschitz-condition in (2.4) forces this smoothness to extend somewhat around θ_0 . These two facts lead to the exclusion of multiple interesting choices of ψ s. In particular, for (2.4) to hold true, ψ needs to be continuous as a function of θ for almost all fixed z . Because of this, most ψ s based on the indicator function $I(z \leq \theta)$, which returns one if the condition within the parenthesis is true and zero otherwise, will fail to satisfy (2.4), since for any neighbourhood $\mathcal{N}$, $I(z \leq \theta)$ is discontinuous in θ when $z \in \mathcal{N}$. See Section 5.3 for more details concerning M-estimators based on the indicator function.

That being said, there are still interesting estimators which fit into the framework of Theorem 2. The median is an example. This estimator maximises $\Psi_n(\theta)$ when $\psi(z, \theta) = -|z - \theta| + |z|$, and this ψ satisfies (2.4) by the triangle inequality. Furthermore, as is shown in van der Vaart 2000, p. 55

$$\Psi(\theta) = 2 \int_0^\theta F(z) dz - \theta$$

Hence, by the fundamental theorem of calculus, as long as the cumulative distribution function F is differentiable at θ_0 with derivative $f(\theta_0)$, the same holds true for Ψ , and $\Psi''(\theta_0) = 2f(\theta_0)$. Furthermore, $\nabla\psi(z, \theta_0) = -I(z < \theta_0) + I(z > \theta_0)$ as long as $z \neq \theta_0$, with variance $F(\theta_0) + 1 - F(\theta_0) - [1 - 2F(\theta_0)]^2 = 4F(\theta_0) - 4F(\theta_0)^2 = 1$. Hence, Theorem 2 can be used to show that the empirical median converges in probability to the true median in the distribution of the data, and that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\{0, [2f(\theta_0)]^{-2}\}$.

Consistency and asymptotic normality of the median was shown by Laplace in 1818, see Stigler 1973, and hence, being able to prove these properties using Theorem 2 is perhaps not the most impressive achievement. Nevertheless, the fact that Theorem 2 is able to handle less smooth functions and formulates the estimation procedure as a maximisation problem rather than solving a set of

equations, gives the result added utility. To be able to work with really non-smooth ψ s, however, Theorem 2 is not enough, and theory with slightly more meat on the bone is required. Furthermore, the maximiser of the function Ψ_n might lie on the boundary of the parameter space, which leads to its own whole set of problems. I will now discuss approaches to and theory regarding these, in my own opinion, two largest problems with M- and Z-estimators.

2.3 When things are not sufficiently regular

2.3.1 Lack of smoothness

The literature on non-smooth estimating equations is a lot less accessible than for smooth problems. The reason for this is that not being able to differentiate, makes Taylor's theorem inapplicable, which renders the mathematics a lot more complicated. Furthermore, when φ or ψ is not smooth, one is often “forced” out of the $O_p(n^{-1/2})$ -regime,⁴ i.e. the estimator $\hat{\theta}_n$ no longer satisfies $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N$ for some distribution N , and instead $n^\alpha(\hat{\theta}_n - \theta_0)$ converges in law to some limit distribution, with a $\alpha < 1/2$.⁵ I will now briefly outline some approaches to non-smooth M-estimation problems, but as neither of the papers in this thesis have been concerned with such estimators, I will be a lot more brief than in the previous sections.

In 1990, Kim and Pollard 1990 proved a general result concerning asymptotics for estimators of cube root order, i.e. estimators satisfying $\hat{\theta}_n - \theta_0 = O_p(n^{-1/3})$. This paper shed light on earlier results, like the estimator of the mode studied in Chernoff 1964, by being the first to provide proper conditions and a unified theory for M-estimators with $O_p(n^{-1/3})$ convergence rates. The main theorem of the paper does not only give conditions for when $(\hat{\theta}_n - \theta_0) = O_p(n^{-1/3})$, but also shows that $n^{1/3}(\hat{\theta}_n - \theta_0)$ converges in distribution towards the random vector maximising a Gaussian process with continuous sample paths and expected value and covariance structure which are both given in the paper. This theorem bears strong resemblance with the results in van der Vaart and Wellner 1996, Sec. 3.2 which extend the result of Kim and Pollard 1990 to other orders than $O_p(n^{-1/3})$. What makes Kim and Pollard 1990 and van der Vaart and Wellner 1996, Sec. 3.2 relevant to this subsection of the thesis, is the fact that very few smoothness conditions are imposed on ψ directly in these works. In both sources, it is assumed that the limit function Ψ is twice differentiable, but differentiability of ψ (or Ψ_n) is not required. Instead both sources manage to replace the smoothness condition of Theorem 2 with more abstract concepts from empirical process theory. These requirements are highly technical and will be skipped for now, but intuitively, what is required is that the classes of functions $M_\delta = \{\psi(\cdot, \theta) - \psi(\cdot, \theta_0) \mid \|\theta - \theta_0\| < \delta\}$, for

⁴For a sequence of random variables $(X_n)_{n=1}^\infty$, $X_n = O_p(n^{-\alpha})$ means that $n^\alpha X_n$ is stochastically bounded or bounded in probability, see van der Vaart 2000, Ch. 2. The notation $X_n = o_p(n^{-\alpha})$ means that $n^\alpha X_n$ converges in probability to zero.

⁵This is, of course, not always the case, see e.g. Hjort et al. 2018 for an example of an estimator which maximises a non-smooth function while preserving the $O_p(n^{-1/2})$ convergence rate.

δ close to zero, are in some sense “small”, or as Kim and Pollard 1990 beautifully calls it “uniformly manageable”. In Section 5.3, I will discuss these requirements in more detail.

There has of course been made more recent advances to the theory for non-smooth M- and Z-estimation problems. For instance, Radchenko 2008 derives results for $\hat{\theta}_n = (\hat{\theta}_n^1, \hat{\theta}_n^2)$ when $\hat{\theta}_n^j$ have different convergence rates for $j = 1, 2$. Furthermore, nontrivial applications of the theory has been made in e.g. Mohammadi and van de Geer 2005 who investigate asymptotics for parametric classifiers fitted by minimising the empirical error rate, and Delsol and Van Keilegom 2020 combines non-smooth M-estimator results with theory concerning semi-parametric estimators. Nevertheless, Kim and Pollard 1990 and van der Vaart and Wellner 1996 stand firm as the two “holy grails” for non-smooth M-estimation problems despite the relatively high entry level of these texts.

2.3.2 Boundary issues

Let $\hat{\theta}_n$ be an M-estimator maximising $\Psi_n(\theta) = \sum_{i=1}^n \psi(Z_i, \theta)$. If ψ is differentiable and $\hat{\theta}_n$ lies in the interior of the parameter space Θ , $\Psi'_n(\hat{\theta}_n) = 0$, and the M-estimation problem reduces to a Z-estimation. This is, however, only the case if $\hat{\theta}_n$ is in the interior of Θ . If $\hat{\theta}_n$ lies on the boundary of Θ ($\partial\Theta$), $\Psi'_n(\hat{\theta}_n) = 0$ is not necessarily true. In many cases this is not a problem, as it suffices that $\hat{\theta}_n$ “eventually” lies in the interior of Θ , which is the case when its limit, θ_0 , is an interior point of Θ . To see this, notice that by definition of the interior there is an $\epsilon > 0$ such that the open ball centered at θ_0 and with radius ϵ , $B_\epsilon(\theta_0)$, is contained in Θ . As $\hat{\theta}_n \xrightarrow{p} \theta_0$, the probability of $\hat{\theta}_n \in B_\epsilon(\theta_0)$ tends towards one, and $\hat{\theta}_n$ being in the interior of Θ will therefore be more and more probable. Hence, the effect of $\hat{\theta}_n \in \partial\Theta$ can be asymptotically neglected. When θ_0 lies on the boundary, however, the story becomes very different. If $\theta_0 \in \partial\Theta$, the probability of $\hat{\theta}_n \in \partial\Theta$ will not disappear and this case needs to be accounted for. All hope is not lost, however, as quite a bit is still known about how maximisers behave even when Θ is not open, assuming the situation is, as one often needs to require, “sufficiently regular”.

Assume that $\partial\Theta$ can be parameterised as $\partial\Theta = \{\theta \mid h(\theta) = 0\}$ for some differentiable function $h: \mathbb{R}^p \rightarrow \mathbb{R}^q$ with component functions $h_1, \dots, h_q$. Then, the maximiser of Ψ_n at the boundary, $\hat{\theta}_n^B$, can be found by using the Lagrange method of multipliers (see any introductory textbook in analysis, e.g. Lindstrøm and Hveberg 2015). By this method,

$$\nabla_{\hat{\theta}_n^B} \Psi(\theta) = \sum_{j=1}^q \lambda_j \nabla_{\hat{\theta}_n^B} h_j(\theta)$$

for some constants $\lambda_1, \dots, \lambda_q \in \mathbb{R}$. Since $h(\theta) = 0$ is satisfied if and only if $nh(\theta) = 0$, the above equation can be reformulated to require

$$0 = \sum_{i=1}^n \begin{pmatrix} \nabla_{\hat{\theta}_n^B} \psi(Z_i, \theta) - \sum_{j=1}^q \lambda_j \nabla_{\hat{\theta}_n^B} h_j(\theta) \\ h(\theta) \end{pmatrix}. \quad (2.5)$$

To find $\hat{\theta}_n^B$, it therefore suffices to solve the above equation for $\lambda = (\lambda_1, \dots, \lambda_q)^T$ and θ . After scaling by the sample size, the right-hand side of (2.5) is a mean,

and hence, the maximiser of Ψ_n on the boundary of Θ is a component of a Z-estimator. Standard theory can therefore be used to study the limiting behaviour of $\hat{\theta}_n^B$. This approach is taken in Aitchison and Silvey 1958, where they study asymptotic properties of ML estimators subjected to constraints.

Assume furthermore that Ψ_n has a maximum at $\hat{\theta}_n^I$ in the interior of Θ . Then $\hat{\theta}_n^I$ solves $\Psi'_n(\theta) = 0$ and is also a Z-estimator. Furthermore, by definition $\hat{\theta}_n$, the maximiser of Ψ_n in Θ , can be written as

$$\hat{\theta}_n = \hat{\theta}_n^B I[\Psi_n(\hat{\theta}_n^B) > \Psi_n(\hat{\theta}_n^I)] + \hat{\theta}_n^I I[\Psi_n(\hat{\theta}_n^B) \leq \Psi_n(\hat{\theta}_n^I)]. \quad (2.6)$$

One can therefore use more or less standard theory to study the limiting behaviour of $\hat{\theta}_n$. To see this, simply, or should I rather say complicatedly, let $P = (\theta^B, \theta^I, \lambda^T, M^1, M^2)$ and notice that $\hat{P} = ((\hat{\theta}_n^B)^T, (\hat{\theta}_n^I)^T, \hat{\lambda}_n^T, \Psi_n(\hat{\theta}_n^B)^T, \Psi_n(\hat{\theta}_n^I)^T)^T$ is the solution to

$$0 = \sum_{i=1}^n \begin{pmatrix} \nabla_{\theta^B} \psi(Z_i, \theta) - \sum_{j=1}^q \lambda_j \nabla_{\theta^B} h_j(\theta) \\ h(\theta^B) \\ \nabla_{\theta^I} \psi(Z_i, \theta) \\ \psi(Z_i, \theta^B) - M_1 \\ \psi(Z_i, \theta^I) - M_2 \end{pmatrix}, \quad (2.7)$$

when $\hat{\lambda}$ is a vector solving (2.5). Again the right-hand side of the above equation is a sum of functions, ensuring that $\hat{P}$ is a Z-estimator converging in probability towards some $P_0 = ((\theta_0^B)^T, (\theta_0^I)^T, \lambda^T, M_0^1, M_0^2)^T$ and $\sqrt{n}(\hat{P} - P_0) \xrightarrow{d} N(0, V)$ for some matrix V . Now let $g(P) = \theta^B I(M_1 > M_2) + \theta^I I(M_2 \geq M_1)$. Then by (2.6), $\hat{\theta}_n = g(\hat{P}_E)$. Furthermore, as long as $E\psi(Z, \theta_0^B) = M_0^1 \neq M_0^2 = E\psi(Z, \theta_0^I)$, the delta method can be used to show that $\hat{\theta}_n$ is consistent for θ_0 the minimiser of Ψ in Θ and that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in law to a central normal distribution with variance matrix

$$g'(P_0)Vg'(P_0)^T = \begin{cases} V_{1,1}, & \Psi(\theta_0^B) > \Psi(\theta_0^I) \\ V_{2,2}, & \Psi(\theta_0^B) < \Psi(\theta_0^I) \end{cases} \quad \text{where } V = \begin{pmatrix} V_{1,1} & V_{1,2} & V_{1,3} \\ V_{2,1} & V_{2,2} & V_{2,3} \\ V_{3,1} & V_{3,2} & V_{3,3} \end{pmatrix},$$

with $V_{1,1}, V_{2,2}$ being $p \times p$ -matrices and $V_{3,3}$ a $(q + 2p) \times (q + 2p)$ -matrix. Hence, $\hat{\theta}_n$ behaves asymptotically as “the best” of the two alternatives $\hat{\theta}_n^B$ and $\hat{\theta}_n^I$. This idea is a variation on how boundary issues are dealt with in **Paper III**.

For the method presented in the above paragraph to work, so very many regularity conditions need to be met. Because of this, more elegant and powerful results exist in the literature. Moran 1971 considers ML estimation when the components are restricted to lie within bounded intervals, and Chernoff 1954 and Chant 1974 discuss how one can test if the true value of the parameter lies in $\partial\Theta$. LeCam 1970 and Geyer 1994 study this problem under quite general assumptions⁶. Andrews 1999, 2001, 2002 extend the theory even further to allow for extremum estimators of locally quadratic objective functions. For more modern applications, see e.g. Ketz 2018 or Fan and Shi 2023.

⁶Geyer 1994, p. 1995 even writes “Any closed cone is Chernoff regular at its vertex, for example, so the support of the MLE could be a fractal.”, implying that his theory can deal with highly complicated parameters sets. A fact which is nice to know should one ever find oneself in the, perhaps slightly absurd, situation of optimising a likelihood function over a fractal.

2.4 Two-step estimators

In the previous section, I discussed the estimator $\hat{\theta}_n$ defined in (2.6). This estimator is defined by first computing $\hat{\theta}_n^B$ and $\hat{\theta}_n^I$, and afterwards selecting the one with the highest associated value of $\Psi_n(\theta)$. This is an estimation procedure divided into two steps, and because of this, $\hat{\theta}_n$ is what is called a two-step estimator. In addition to $\hat{\theta}_n$ of Section 2.3.2, examples of two-stage estimators include estimators in the presence of plug-in estimators and, by the results of **Paper III**, estimation in penalised likelihood models when the size of the penalisation term is set to minimise a cross-validated estimator of the risk.

In Section 2.3.2, the two-stage process of estimating $\hat{\theta}_n$ was reduced to the “single” step of solving (2.7). Because of this, limiting properties of $\hat{\theta}_n$ could be studied using classic theory for Z-estimators. It turns out that this is the case for many two-stage estimators. A quite general case is the following. Assume that for every fixed value of α , β is found by solving $0 = \sum_{i=1}^n \varphi_1(Z_i, \alpha, \beta)$, and denote the solution it by $\hat{\beta}(\alpha)$. Suppose further that α is set to the root of $\sum_{i=1}^n \varphi_2(Z_i, \alpha)$, denoted by $\hat{\alpha}$. To study $\hat{\beta}(\hat{\alpha})$, one can utilise that $(\hat{\alpha}^T, \hat{\beta}(\hat{\alpha})^T)^T$ solves the two equations $0 = \sum_{i=1}^n \varphi_2(Z_i, \alpha)$ and $0 = \sum_{i=1}^n \varphi_1(Z_i, \alpha, \beta)$ simultaneously for α and β , a procedure which is equivalent to finding the root of

$$0 = \sum_{i=1}^n \begin{pmatrix} \varphi_1(Z_i, \alpha, \beta) \\ \varphi_2(Z_i, \alpha) \end{pmatrix}. \quad (2.8)$$

This is a Z-estimation procedure with properties outlined in Section 2.1. Although relatively simple to state, solutions to (2.8) cover a large class of two-step estimators. If we, for instance, want to estimate the standard deviation σ while using the empirical mean as a plug-in for the expected value, one can use $\varphi_1(z, \mu, \sigma) = (z - \mu)^2 - \sigma^2$ and $\varphi_2(z, \mu) = z - \mu$. Another example is in maximum likelihood estimation when the likelihood function L_n depends on two, possibly multidimensional, parameters θ_1 and θ_2 . If L_n is hard to maximise with respect to both parameters simultaneously, one can instead use some plug-in estimator $\hat{\theta}_1$ of θ_1 and maximise $\theta_2 \mapsto L_n(\hat{\theta}_1, \theta_2)$ instead. If the plug-in estimator of θ_1 is a Z-estimator, this fits into the framework of (2.8) by setting $\varphi_1(z, \theta_1, \theta_2) = \nabla_{\theta_2} \log f(z, \theta_1, \theta_2)$ and letting $\varphi_2(z, \theta_1)$ be the estimating function corresponding to $\hat{\theta}_1$. In **Paper II** this method is used to allow for plug-in estimators in the empirical likelihood function.

Although the results of Gong and Samaniego 1981 for the special case of maximum likelihood estimation predates this slightly, the above way of dealing with two-stage estimators was first presented in Newey 1984 in its full generality. This method can also be extended to cover the case where $\hat{\beta}(\alpha)$ maximises a function $\beta \mapsto \sum_{i=1}^n \psi_1(Z_i, \alpha, \beta)$ rather than solving an equation on the form $0 = \sum_{i=1}^n \varphi_1(Z_i, \alpha, \beta)$, see Pagan 1986.

I, however, was personally first introduced to the concept of two-step estimation in the paper Joe 2005, where arguments similar to those given above are used to show consistency and asymptotic normality of the parameters obtained by fitting copulas in two steps. This idea was first visited in Shih and Louis 1995 for two

dimensions, and is extended to the case of possible misspecification in Ko and Hjort 2019.

The two- or multiple-step idea is actually used quite a lot in copula theory, due to the, to some extent, sequential definition of these functions. Formally, a d -dimensional copula is a distribution function on the unit square in $\mathbb{R}^d$, and by Sklar's theorem (first given in the French Sklar 1959, see e.g. Nelsen 2006, Th. 2.3.3 for an English presentation), any multivariate distribution function $F(z_1, \dots, z_d)$ can be written as

$$F(z_1, \dots, z_d) = C[F_1(z_1), \dots, F_d(z_d)],$$

for some copula C , where F_j is the marginal distribution function of Z_j for each j . This transformation allows one to separate the dependence structure from the marginal distribution of each component, and hence, modelling can be done in two steps. In situations where the dependence structure is important, this sequential approach can simplify modelling greatly and make dependencies easier to spot, as is illustrated beautifully in Figure 1 in Genest and Rémillard 2006.⁷

2.5 Influence functions

In Section 2.1, I “showed” that for Z-estimators, $\hat{\theta}_n - \theta_0 \approx -n^{-1}\Phi(\theta_0)^{-1}\Phi_n(\theta_0)$. Writing the approximation slightly differently, one sees

$$\hat{\theta}_n - \theta_0 \approx -\Phi(\theta_0)^{-1}n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta_0), \quad (2.9)$$

which shows that after subtracting θ_0 , $\hat{\theta}_n$ is actually more or less equal to a mean. In Section 2.1, this was used to argue that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to a central normal distribution, but the result actually contains more information than that. For instance, one notices that, up to the precision of the approximation, the “effect” or “influence” data point Z_i has on the estimator $\hat{\theta}_n$ is precisely the corresponding term $-n^{-1}\Phi'(\theta_0)^{-1}\varphi(Z_i, \theta_0)$ for each i . Hence, removing this i -th data point should change the estimator $\hat{\theta}_n$ by about $n^{-1}\Phi(\theta_0)^{-1}\varphi(Z_i, \theta_0)$ and adding an observation Z_{n+1} should approximately give you $\hat{\theta}_n - n\Phi(\theta_0)^{-1}\varphi(Z_{n+1}, \theta_0)$. This is the idea behind influence functions, a concept which is present in all articles of this thesis, but of particular importance in **Paper I** and **Paper III**. The formal definition, see Hampel 1974; Hampel et al. 1986 or van der Vaart 2000, Ch. 20, of an influence

⁷This paper is a answer to Mikosch 2006, a paper which stirred heated debate in the copula community. Mikosch likens the utility of copulas to the emperors new clothes in the Danish story “The emperor’s new clothes” by H.C. Andersen, meaning that the usefulness of copulas is nothing but an illusion and that researchers are overemphasising the practicality and novelty of the concept. He even goes as far as writing “I suspect that some include the word ‘copula’ in the title of their papers not because they contribute to the theory on copulas, but because they believe that one can publish easier”. This paper was not well received by all researchers in the field, and the response in Genest and Rémillard 2006 is quite crass and amusing to read, with sentences like “While we respect his right to ‘ask some naive questions,’ we can hardly accept that either through ignorance or malice, he depicts copula theory as an unsubstantiated fad that leads to ‘a biased view of stochastic dependence.’ ”.

function is quite abstract, but I will now try to give an informal explanation of the concept, which hopefully will give some intuition. All messy details will be skipped, but either of the aforementioned sources can be consulted for more rigorous treatments.

Influence functions are defined for functions mapping probability distributions on some sample space Ω into $\mathbb{R}^p$. An example of such a function is

$$P \mapsto \int_{\Omega} x \, dP(x), \quad (2.10)$$

which, when evaluated at a point P , returns the expected value of the distribution P . When evaluated at $\mathbb{P}_n$, the empirical measure putting equal weight $1/n$ on each of $Z_1, \dots, Z_n$, one gets the mean of the data. The mean, as a statistic, can therefore be defined as a composition of mapping the sample to the empirical distribution and afterwards applying T , the map defined in (2.10). This is illustrated by the following diagram.

$$Z_1, \dots, Z_n \longrightarrow \mathbb{P}_n \xrightarrow{T} n^{-1} \sum_{i=1}^n Z_i$$



 mean

Another example of a function mapping probability distributions into $\mathbb{R}^p$ is

$$P \mapsto \text{root of } \int_{\Omega} \varphi(\omega, \theta) \, dP(\omega), \quad (2.11)$$

which returns the zero of $E\varphi(Z, \theta)$ when $Z \sim P$. At $\mathbb{P}_n$ (2.11) gives us the solution to $\sum_{i=1}^n \varphi(Z_i, \theta) = 0$, and hence, the Z-estimator which for a given data set $Z_1, \dots, Z_n$ solves $\sum_{i=1}^n \varphi(Z_i, \theta) = 0$, can be characterised as mapping the empirical distribution $\mathbb{P}_n$ associated with the data set $Z_1, \dots, Z_n$ according to (2.11).

Assume now that one has some estimation procedure which can be characterised by a function T mapping probability distributions into $\mathbb{R}^p$, e.g. T as in (2.10) for means or T as in (2.11) for Z-estimators with estimating function φ . How much each data point affects the estimation procedure is dependent on the underlying distribution of the data. If, for instance, the data is Cauchy distributed, the mean will be more volatile than if the true underlying distribution is the normal. One therefore needs to consider with respect to which distribution one is computing the influence function. Call this distribution P . In this case, one is, for each $\omega \in \Omega$, interested in evaluating how much T changes when the data point ω is removed, given that the true underlying distribution is P . Hence, the influence function at a point ω is trying to measure how much T changes in the direction of $P - \delta_\omega$, where δ_ω is the measure with point-mass 1 at $\{\omega\}$, at P . Formally, one writes

$$IF_{T,P}(\omega) = \lim_{\epsilon \rightarrow 0} \frac{T[(1 - \epsilon)P + \epsilon\delta_\omega] - T(P)}{\epsilon}.$$

or equivalently, that the influence function evaluated at ω is the derivative⁸ of T

⁸Here I do of course not mean derivative as in the traditional sense for functions defined on $\mathbb{R}^q$ for some q , but rather the Gateaux-derivative, a generalised version which can be applied to functions on normed spaces. See e.g. Lindstrøm 2017, Ch. 6 or any textbook on mathematical analysis.

at P in the direction of $P - \delta_\omega$. For the mean,

$$\begin{aligned} IF_{T,P}(\omega) &= \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_{\Omega} x \, d[(1 - \epsilon)P + \epsilon\delta_\omega](x) \\ &= \lim_{\epsilon \rightarrow 0} \frac{1 - \epsilon}{\epsilon} \int_{\Omega} x \, dP(x) + \int_{\Omega} x \, d\delta_\omega(x) \\ &= -E_P X + \omega. \end{aligned}$$

For the Z-estimator in (2.11),

$$IF_{T,P}(\omega) = -\Phi(\theta_0)^{-1} \varphi(\omega, \theta_0), \quad (2.12)$$

see van der Vaart 2000, Ex. 20.4. The right-hand side of (2.9) is recognisable as the mean of (2.12) evaluated at each data point Z_i , illustrating that the influence function really does “measure” how much the Z-estimator is affected by each data point. This is a core idea in **Paper III**, where influence functions are used to find an alternative characterisation of the leave-one-out cross-validated estimator of the risk.

Although some preliminary work was apparently done in his unpublished PhD thesis (see Hampel et al. 1986), the term “influence function” was introduced by Hampel 1974⁹. The idea of differentiating with respect distributions predates this paper, however. von Mises 1947, for instance, uses a similar idea to derive an analogue to the Taylor expansion for statistical functions mapping distributions into $\mathbb{R}^p$. Informally, what he shows is that one can more or less write

$$T(\mathbb{P}_n) = T(P) + T'(P)(\mathbb{P}_n - P) + \frac{1}{2}T^{(2)}(\mathbb{P}_n - P) + \dots + \frac{1}{k!}T^{(k)}(\mathbb{P}_n - P) + \dots, \quad (2.13)$$

where $T^{(k)}(P)$ denotes the k -th derivative of T with respect to P . The expansion in (2.13) is called the von Mises expansion of T and in **Paper I** it is used to informally to give intuition about the second order bias of estimators.

Influence functions have become popular in the field of robust statistics.¹⁰ This is perhaps not too surprising as the ability to evaluate the effect each and every data point has on the estimator $\hat{\theta}_n$ is very useful when assessing how sensitive the estimator is to outliers in the observations. In this field, one is especially interested in whether an estimator’s influence function is bounded or not. This is quite natural, as a bounded influence function can guarantee that a single outlier does not “mess up” the relevant estimator too much. Recently, influence functions have also found applications in the field of machine learning and causal inference, see e.g. Fisher and Kennedy 2020 and Hines et al. 2022.

⁹He uses the word influence curve rather than influence function in this paper.

¹⁰As can be seen from e.g. Huber and Ronchetti 2009. See also or Avella-Medina 2017 for a relatively new application, and Avella Medina and Ronchetti 2015 for a review on influence functions and other themes in robust statistics.

Chapter 3

Estimating errors

A major topic of this thesis, and indeed of statistics in general, is “errors”. By errors, I mean measures of how wrong a model or estimator is. The topic of **Paper I** is precise estimation of the bias of estimators and the expected squared distance to the true value they are aiming for. In **Paper II**, these results are used to carefully evaluate the errors made by different hybrid estimators and choose the optimal one, while **Paper III** studies how choosing models by minimising error estimators affects analysis. In addition, this paper also studies a particular error estimator, the cross-validated estimator of the risk, in detail. Since the theme of errors is so fundamental to this thesis, I will attempt to formalise what I mean by this slightly imprecise term and outline different ways of approximating “how wrong” an estimator or model is.

3.1 Loss and risk functions

The first thing I will do is to formalise the notion of “errors”. To do this, I will use the concepts of loss and risk functions from decision theory. These concepts were formalised by Abraham Wald in the book Wald 1950, based on several of his previous papers starting with Wald 1939, but some of these ideas are very natural, and many concepts therefore go a lot further back than this source, the perhaps most famous example being Pascal’s wager in his book *Pensées* from 1670, where he uses risk minimisation to evaluate whether one should or should not believe in God, see e.g. Hájek 2024. In the following, I will stay true to the presentation given by Casella and Berger 2002, Ch. 7.3 on loss and risk functions.

Assume our goal is to estimate some parameter θ with true value θ_0 . This parameter could be the expectation or the median in a distribution, the parameter corresponding to the closest parametric fit or something else entirely. Assume furthermore that there is some function L , such that for every “candidate” $\theta \in \Theta$, one “loses” $L(\theta, \theta_0)$ by using θ instead of the true value θ_0 . For one dimensional θ , the most obvious example of such a loss function is perhaps $L(\theta, \theta_0) = |\theta - \theta_0|$, the distance from the candidate to the true value. This is called the absolute loss. The function $|\cdot|$ is not smoothly differentiable, and it is therefore common to instead work with the squared error loss, $L(\theta, \theta_0) = (\theta - \theta_0)^2$, which is mathematically more convenient. Another, slightly less obvious, example of a loss function is

$L(\theta, \theta_0) = \text{KL}(f_{\theta_0}, f_{\theta})$, where KL is the Kullback-Leibler divergence discussed in Section 2.1.2. See Section 3.1.2 for more on this last loss function.

A loss function tells us how much one loses by using a candidate θ instead of the true value θ_0 . Hence, for an estimate $\hat{\theta}_n$ based on a sample $Z_1, \dots, Z_n$, the associated loss of using $\hat{\theta}_n$ instead of the true value, is $L(\hat{\theta}_n, \theta_0)$. Typically, it is hard to say much about the loss a single estimate $\hat{\theta}_n$ inflicts. Instead, one usually studies the risk of the estimator, i.e. how much using the estimating procedure which produced $\hat{\theta}_n$ loses on average. Formally, this is stated in the following way. Let T be some estimator such that for a sample of data points $Z_1, \dots, Z_n$, $T(Z_1, \dots, Z_n)$ estimates θ . Then, for a given loss function L , the associated risk of using the estimator T to estimate θ_0 is $R(\theta_0, T) = \mathbb{E} L[\theta_0, T(Z_1, \dots, Z_n)]$.

For the squared error loss function, the associated risk, $R(\theta_0, T) = \mathbb{E}[\theta_0 - T(Z_1, \dots, Z_n)]^2$, is the MSE of the estimator $T(Z_1, \dots, Z_n)$. This is the perhaps most used and well-known risk function, and famously it decomposes into the sum of the squared bias and variance of $T(Z_1, \dots, Z_n)$:

$$\text{MSE}(\theta_0, T) = [\theta_0 - \mathbb{E} T(Z_1, \dots, Z_n)]^2 + \text{Var} T(Z_1, \dots, Z_n). \quad (3.1)$$

Hence, to study the MSE of an estimator T , it suffices to consider the bias and variance separately, an idea which is central in the focused information criterion which is present in all articles and will be discussed in Section 3.2.2 of this thesis.

The decomposition in (3.1) is not as unique to the MSE as one might initially think. Indeed, let L be some loss function which is smooth in its last argument. A Taylor expansion around a point $\theta \in \Theta$ reveals

$$L[\theta_0, \hat{\theta}_n] = L(\theta_0, \theta) + \nabla_{\theta} L(\theta_0, \theta)^T (\hat{\theta}_n - \theta) + \frac{1}{2} \text{Tr}[H_{\theta} L(\theta_0, \theta) (\hat{\theta}_n - \theta_0)(\hat{\theta}_n - \theta_0)^T] + \epsilon_n,$$

where I have replace $T(Z_1, \dots, Z_n)$ with $\hat{\theta}_n$ to simplify notation somewhat, and $\epsilon_n = O(\|\hat{\theta}_n - \theta\|^3)$. Assume now that $\hat{\theta}_n$ is consistent for some value θ_{lf} , and that $n^{\alpha}(\hat{\theta}_n - \theta_{\text{lf}}) \xrightarrow{d} N[0, V(\theta_{\text{lf}})]$. This is quite often the case, e.g. for M- and Z-estimators, see Sections 2.1 and 2.2. One then finds

$$R(\theta_0, T) = L(\theta_0, \theta_{\text{lf}}) + \nabla_{\theta_{\text{lf}}} L(\theta_0, \theta)^T \mathbb{E}(\hat{\theta}_n - \theta_{\text{lf}}) + \frac{n^{2\alpha}}{2} \text{Tr}[H_{\theta_{\text{lf}}} L(\theta_0, \theta) V(\theta_{\text{lf}})] + O(n^{-3\alpha}), \quad (3.2)$$

assuming everything is sufficiently regular. In the above, Tr denotes the trace operator. The first two terms in the above expansion increase with the bias of $\hat{\theta}_n$, while the third term is large when the variance $V(\theta_{\text{lf}})$ is. Hence, all risk functions, not just the MSE, can approximately be decomposed into a bias and a variance term.

Being a bit more careful with the details in Section 2.1, one can show

$$\hat{\theta}_n = \theta_{\text{lf}} - n^{-1} \Phi'(\theta_{\text{lf}})^{-1} \sum_{i=1}^n \varphi(Z_i, \theta_{\text{lf}}) + \delta_n,$$

where $\delta_n = O_p(1/n)$. By assumption, $\mathbb{E} \varphi(Z, \theta_{\text{lf}}) = 0$, and hence, $\mathbb{E}[\hat{\theta}_n - \theta_{\text{lf}}] = \mathbb{E} \delta_n$. As $\delta_n = O_p(1/n)$, one would expect that $\mathbb{E} \delta_n = O(1/n)$. Writing $\mathbb{E} \delta_n =$

$c/n + o(1/n)$, and plugging into (3.2) shows

$$R(\theta_0, T) = L(\theta_0, \theta_{\text{lf}}) + n^{-1} \left\{ \nabla_{\theta_{\text{lf}}} L(\theta_0, \theta)^T c + \frac{1}{2} \text{Tr}[H_{\theta_{\text{lf}}} L(\theta_0, \theta) V(\theta_{\text{lf}})] \right\} + o(n^{-1}). \quad (3.3)$$

Hence, provided all regularity assumptions hold true, the limiting risk of a Z-estimator consists of three parts: A constant term $L(\theta_0, \theta_{\text{lf}})$ quantifying the loss of the limit of $\hat{\theta}_n$, a term of size $O(1/n)$ involving the variance and second order bias of $\hat{\theta}_n$ and a remainder term of smaller order. This expansion is of course not unique to Z-estimators, and analogue expressions to (3.3) can be derived in a similar fashion for most estimators with influence function IF which can be written as $\hat{\theta}_n \approx \theta_{\text{lf}} + n^{-1} \sum_{i=1}^n \text{IF}(Z_i)$, see **Paper III**.

As the sample size grows, the constant term $L(\theta_0, \theta_{\text{lf}})$ will eventually dominate the right hand side of (3.3). Because of this, the estimator with smallest associated value of $L(\theta_0, \theta_{\text{lf}})$ is the asymptotically “optimal” estimator, in the sense that such approximations minimise the limiting value of $R(\theta_0, T)$. For finite sample sizes, however, this might not be the case. Even though the second term in (3.3) will eventually be negligible compared to $L(\theta_0, \theta_{\text{lf}})$, it might still be relatively large for moderate n . Hence, for finite sample sizes, using an estimator with a lot smaller c and $V(\theta_{\text{lf}})$ might be preferable in spite of a slightly larger value of $L(\theta_0, \theta_{\text{lf}})$. This idea of balancing bias and variance is often referred to as the bias-variance trade-off in statistics and machine learning. See Hastie et al. 2009, Ch. 7 for a thorough discussion.

3.1.1 Prediction and classification

Before I move on, I will briefly go through how loss and risk function fit into the field of prediction and classification as presented in Hastie et al. 2009, Sec. 2.4. Assume one wants to predict the outcome of y , and that to every potential prediction $\hat{y}$, one can describe the loss inflicted by using $\hat{y}$ in place of y by $L(y, \hat{y})$. In prediction problems, squared error loss is the most common choice. For classification, one can for instance, use the 0-1 loss defined as $L(y, \hat{y}) = I(y \neq \hat{y})$. If one has some estimator of y , based on data $Z_1, \dots, Z_n$ which can be written as $T(Z_1, \dots, Z_n)$, the associated risk of using the procedure to estimate y is $R(y, T) = \mathbb{E} L[y, T(Z_1, \dots, Z_n)]$. This risk is defined for every y , and hence in the expectation in R , y is treated as a fixed and non-random value. An example of the above, is in linear regression, where one would have $R(y, T) = \mathbb{E}(y - \hat{\beta}_0 - \hat{\beta}^T x)^2$ for every pair (y, x) .

Typically, one is not interested in the loss of estimating a single y by $\hat{y}$. Instead, some sort of “overall” performance is used to evaluate regression and classification procedures. Because of this, one often instead considers the expected predicted loss (EPL) instead of L . EPL is defined for prediction or classification models g which return an estimated value of y for every input x , and takes the form $\text{EPL}(g) = \mathbb{E} L[Y, g(X)]$. Assume further that for each x , one has some estimator $T[x, Z_1, \dots, Z_n]$ of $g(x)$ based on data $Z_1, \dots, Z_n$. One can then define what I will

call the expected predicted risk, or EPR for short:

$$\begin{aligned} \text{EPR}(T) &= \mathbb{E}_{Z_1, \dots, Z_n} \text{EPL}[T(\cdot, Z_1, \dots, Z_n)] \\ &= \mathbb{E}_{Z_1, \dots, Z_n} \mathbb{E}_{(Y, X)} \{L[Y, T(X, Z_1, \dots, Z_n)] \mid Z_1, \dots, Z_n\}, \end{aligned}$$

where (Y, X) follows the same distribution as each Z_i , but is independent of the data set $Z_1, \dots, Z_n$. In the above I have added subscripts to the expectation to emphasise with respect to which variables I am integrating.

Efron 1975 and O'Neill 1980 studies EPR for particular classes of parametric classifiers when the 0-1 loss is used. The arguments used to show (3.3) in the previous section, are loosely based on the latter of these papers, and it should therefore not come as a surprise that an expansion similar to (3.3) can be shown for the EPR. So, assume the estimator $T(x, Z_1, \dots, Z_n)$ can be written as $g_{\hat{\theta}_n}(x)$ for some estimator $\hat{\theta}_n$. Then, similar arguments to those given to show (3.3) reveals,

$$\text{EPR}(T) = \text{EPR}(\theta_{\text{lf}}) + n^{-1} \left(\nabla_{\theta_{\text{lf}}} \text{EPR}(\theta)^T c + \frac{1}{2} \text{Tr}[H_{\theta_{\text{lf}}} \text{EPR}(\theta) V] \right) + o(1/n),$$

when $\hat{\theta}_n = \theta_{\text{lf}} + n^{-1} \sum_{i=1}^n \text{IF}(Z_i) + \delta_n$ with $n \mathbb{E} \delta_n \rightarrow c$, $V = \text{Var IF}(Z)$ and a lot of other regularity conditions hold true. This expansion is used in **Paper III** to investigate whether the cross-validated estimator of the risk precisely estimates $\text{EPR}(T)$. It will also be crucial in Section 5.1.

3.1.2 The Kullback-Leibler divergence

In Section 2.1.2, I introduced the Kullback-Leibler divergence, defined as $\text{KL}(g, f) = \mathbb{E}_g \log[g(Z)/f(Z)]$, for two pdfs/pmfs f and g , and said that $\text{KL}(g, f)$ can be considered a measure of how far f is from g . This measure of distance was first introduced in Kullback and Leibler 1951, which defined the information in f as the expected value of $\log f(Z)$ under the true distribution. With this notion, $\text{KL}(g, f)$ measures distance in terms of relative loss in information.

The appeal of thinking in terms of the KL divergence was greatly extended by Hirotugu Akaike's previously discussed 1973 paper "Information theory and an extension of the likelihood principle" (Akaike 1973). This article is extremely famous as it not only manages to properly fit the KL-divergence into the framework of loss and risk functions, but also gives arguments for why minimising KL is desirable, giving an information theoretic justification of the ML procedure. Furthermore, the very popular Akaike's information criterion (AIC) was introduced here (although the author himself does not use this name in the article). The AIC connects information theory and statistics in a beautiful way, by estimating the expected Kullback-Leibler divergence from a fitted parametric model to the true underlying density. In Section 3.2.1, I will discuss the AIC in more detail.

In Akaike 1973, the author argues that for a given parametric model $\mathcal{F}_\theta = \{f_\theta \mid \theta \in \Theta\}$, with true value θ_0 , the KL-divergence from f_θ to f_{θ_0} can be considered a loss function, and this section on divergences has been placed under the chapter of loss and risk function for this reason. Indeed, let $L(\theta_0, \theta) = \text{KL}(f_{\theta_0}, f_\theta)$, then L

can be considered a loss function where the loss is measured by the distance from f_θ to f_{θ_0} , or alternatively as the loss of information when using f_θ in place of the true underlying density.

Let $\hat{\theta}_n = T(Z_1, \dots, Z_n)$ be an estimator of θ converging in probability to θ_{lf} and with $\sqrt{n}(\hat{\theta}_n - \theta_{\text{lf}}) \xrightarrow{d} N(0, V)$. Then, the corresponding risk function is $R_{\text{KL}}(\theta_0, T) = E L_{\text{KL}}(\theta_0, \hat{\theta}_n)$. By (3.3) and the fact that θ_0 minimises $\theta \mapsto \text{KL}(f_{\theta_0}, f_\theta)$, one has

$$R_{\text{KL}}(\theta_0, T) = L(\theta_0, \theta_{\text{lf}}) + n^{-1} \frac{1}{2} \text{Tr}[H_{\theta_0} L(\theta_0, \theta_{\text{lf}}) V] + o(1/n),$$

provided everything is sufficiently regular.

Now, let $\hat{\theta}_n$ be the ML estimator of θ and assume the model is specified correctly. By the results of Section 2.1.2, $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, J(\theta_0)^{-1}]$, where $J(\theta_0)$ is the Fisher matrix. Furthermore, by definition, $H_{\theta_0} L(\theta_0, \theta) = -H_{\theta_0} E \log f_\theta(Z) = J(\theta_0)$. Hence,

$$R_{\text{KL}}(\theta_0, T) = L(\theta_0, \theta_0) + n^{-1} \frac{1}{2} \text{Tr} I_p + o(1/n) = p/2n + o(1/n), \quad (3.4)$$

where I_p is the $p \times p$ -identity matrix. This shows that in a correctly specified model, the risk incurred by using the ML estimator instead of the true value is only approximately $p/2n$, at least when loss is measured in terms of KL-divergence.

But what if the true underlying density is not a member of $\mathcal{F}_\theta$? In this case, there is no “true” value θ_0 , and how to define a fitting loss function is not as obvious. A possible way to deal with the lack of a true parameter is to deviate somewhat from the language of loss and risk functions I have used thus far. Instead of considering the loss of using some parameter θ in place of the true value, one can define the “loss” the parameter θ as the amount of information lost when using the distribution f_θ instead of the true underlying pdf/pmf g . Formally, I will say $L(\theta) = \text{KL}(g, f_\theta)$. For some estimator $\hat{\theta}_n = T(Z_1, \dots, Z_n)$ of θ , the associated risk will then be $R(T) = E L(\hat{\theta}_n)$. Here it is assumed that $Z_1, \dots, Z_n$ are i.i.d. following the true distribution g .

Assume now that $\hat{\theta}_n$ is an ML estimator in the model $\mathcal{F}_\theta$. By the results of Section 2.1.2, one has $\hat{\theta}_n \xrightarrow{p} \theta_{\text{lf}}$ and $\sqrt{n}(\hat{\theta}_n - \theta_{\text{lf}}) \xrightarrow{d} N[0, V(\theta_{\text{lf}})]$, where $V(\theta_{\text{lf}}) = J(\theta_{\text{lf}})^{-1} K(\theta_{\text{lf}}) J(\theta_{\text{lf}})^{-1}$ with $J(\theta) = -H_\theta E \log f_\theta(Z)$ and $K(\theta) = \text{Var} \nabla_\theta \log f_\theta(Z)$. Using arguments similar to those preceding (3.3), one can show

$$\begin{aligned} R(T) &= L(\theta_{\text{lf}}) + n^{-1} \frac{1}{2} \text{Tr}[J(\theta_{\text{lf}}) J(\theta_{\text{lf}})^{-1} K(\theta_{\text{lf}}) J(\theta_{\text{lf}})^{-1}] + o(1/n) \\ &= \text{KL}(g, f_{\theta_{\text{lf}}}) + \frac{1}{2n} \text{Tr}[J(\theta_{\text{lf}})^{-1} K(\theta_{\text{lf}})] + o(1/n), \end{aligned}$$

utilising the fact that θ_{lf} minimises $\theta \mapsto \text{KL}(g, f_\theta)$. This equation can be seen as a model robust counterpart to (3.4), as when the model is correctly specified and there is a θ_0 such that $g = f_{\theta_0}$, the arguments in Section 2.1.2 ensure $\theta_{\text{lf}} = \theta_0$ and $J(\theta_0) = K(\theta_0)$, ensuring that the above equation simplifies to

$$R(T) = p/2n + o(1/n),$$

where p is the dimension of θ . This fact forms the backbone of the previously mentioned AIC, and in Section 3.2.1 I will show how the above calculations can be used to derive the criterion. Firstly, however, I will discuss the perhaps most famous result concerning risk functions and optimality.

3.1.3 The Cramér-Rao inequality

In the early 1900s Ronald Fisher and Karl Pearson were the representatives of two competing methods for fitting distribution. Pearson advocated for the method of moments, which fits parametric families by matching empirical and theoretical moments, while Fisher was strongly in favour of the method of maximum likelihood.¹ Now, about a hundred years later, I believe it is safe to say that Fisher has won his debate with Pearson, and that maximum likelihood stands firm as the king of the hill. There are, of course, many explanations for this, but a key reason is the fact that maximum likelihood estimators are asymptotically optimal, in the sense that their variance approaches the lowest among all asymptotically unbiased estimators as the sample size grows to infinity.²

A predecessor of the result alluded to in the previous paragraph was first shown in Edgeworth 1909, but Fisher 1922 was the first to give a proper proof and thorough discussion. In this 1922-paper, he shows that if data follows a distribution f_{θ_0} , the lowest asymptotic variance among all estimators $\hat{\theta}_n$ satisfying $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V)$ for some V is $J(\theta_0)^{-1}$, where $J(\theta_0)$ is the Fisher matrix. By Section 2.1.2, this is the limiting variance of the ML estimator in the model f_{θ} , which ensures that the maximum likelihood estimator is asymptotically efficient, i.e. achieves the lowest possible variance as $n \rightarrow \infty$. Because of this, if one is interested in minimising the MSE, it is not possible to beat an ML estimator in a correctly specified model, at least asymptotically!

This result is an asymptotic analogue of the perhaps more famous Cramér-Rao inequality, proved by Rao 1945 and Cramér 1946 within a year of each other. (More pedagogical expositions can be found in most statistics textbook, e.g. Casella and Berger 2002.) This results states that if data follows a f_{θ_0} distribution, any unbiased estimator of θ_0 will have larger variance than $[nJ(\theta_0)]^{-1}$, the inverse of the Fisher information in the sample.³ Even more generally, the theorem ensures that the variance of any unbiased estimator of $g(\theta)$ is bounded from below by $\nabla_{\theta_0} g(\theta) J(\theta_0)^{-1} \nabla_{\theta_0} g(\theta) / n$, which is exactly the limiting variance of the ML estimator of $g(\theta)$. In this thesis, I will refer to this result as well as Fisher's asymptotic analogue by “the Cramér-Rao inequality”, and this theorem

¹Most of their work from this period references the competing method. For instance, Fisher 1922 dedicates multiple sections to show that the method of moments is efficient only when data is “almost” normally distributed, while Pearson 1936 goes in detail through the calculations in this and other papers, pointing out flaws in the computations.

²Although Pearson 1936 has this to say about the fact: “It does not appear to me that anything of value as to the nature of the Method of Maximum Likelihood can be deduced from $N \rightarrow \infty$ cases.”

³In $\mathbb{R}$, larger is meant as in the classic way. In higher dimensions, a matrix $A \geq B$ if $A - B$ is positive semi-definite.

will be essential in Section 5.1, where I will discuss whether it is really possible to estimate the terms of size $O(1/n)$ in (3.3).

3.2 Estimating risk

Even though the Cramér-Rao inequality proves that, asymptotically speaking, it is not possible to beat an ML estimator in a correctly specified model, the result does not give any assurances when the parametric model fitted is not assumed to be correct. In fact, in the situation where none of the members of $\mathcal{F}_\theta$ are equal to the true underlying distribution, the ML estimator is neither guaranteed to eventually achieve the lowest possible variance, nor be asymptotically unbiased. Because of this, other approaches than those based on the Cramér-Rao inequality needs to be taken if one wishes to compare different estimators in a model agnostic fashion. A natural way is to estimate the risk of all candidate estimators directly. I will now go through a couple of methods which have inspired my own work, in particular in relation with **Paper III**. Error estimation is, however, an enormous field and the techniques presented here do by no means constitute an exhaustive list. Both Stein's unbiased risk estimator (Stein 1981) and estimation based on the bootstrap (Efron 1979) are, for instance, among the darlings I chose to cut.

3.2.1 The AIC and the TIC

The previously mentioned AIC is a very famous risk estimator. It attempts to approximate the risk of the ML estimator in a parametric model $\mathcal{F}_\theta = \{f_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^p\}$, and I will define it as

$$\text{AIC} = -\ell_n(\hat{\theta}_n) + p,$$

where ℓ_n is the log-likelihood function in the parametric model.⁴ I will now informally explain why the AIC takes this form, greatly inspired by Appendix A.7 in Schweder and Hjort 2016.

Assume that the model $\mathcal{F}_\theta$ is correctly specified so that pdf/pmf in the true underlying distribution of the data is f_{θ_0} . A Taylor expansion around θ_0 shows

$$\ell_n(\hat{\theta}) \approx \ell_n(\theta_0) + \nabla_{\theta_0} \ell_n(\theta)^T (\hat{\theta}_n - \theta_0) + \frac{1}{2} (\hat{\theta}_n - \theta_0)^T H_{\theta_0} \ell_n(\theta) (\hat{\theta}_n - \theta_0). \quad (3.5)$$

Now let $U_n(\theta) = n^{-1/2} \nabla_\theta \ell_n(\theta)$ and $J_n(\theta) = -n^{-1} H_\theta \ell_n(\theta)$, which converges to $U(\theta) \sim N[0, J(\theta)]$ and $J(\theta)$ in distribution and probability, respectively (this follows from the CLT, the computations in Section 2.1.2 and the LNN). By (2.1), $\hat{\theta}_n - \theta_0 \approx n^{-1/2} J(\theta_0)^{-1} U_n(\theta_0)$. Plugging this into (3.5) and manipulating the expression somewhat, shows that $\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)$ is approximately equal to

$$\text{Tr}[J(\theta_0)^{-1} U_n(\theta_0) U_n(\theta_0)^T] - \frac{1}{2} \text{Tr}[J_n(\theta_0) J(\theta_0)^{-2} U_n(\theta_0) U_n(\theta_0)^T],$$

⁴Slight variations of this formulation are present in the literature, e.g. $2[p - \ell_n(\hat{\theta}_n)]$ is often used.

which converges in distribution to

$$\text{Tr}[J(\theta_0)^{-1}U(\theta_0)U(\theta_0)^T] - \frac{1}{2} \text{Tr}[J(\theta_0)^{-1}U(\theta_0)U(\theta_0)^T] = \frac{1}{2} \text{Tr}[J(\theta_0)^{-1}U(\theta_0)U(\theta_0)^T],$$

by Slutsky's lemma. Because of this, one would expect that the expectations also converge, i.e. that

$$\text{E } \ell_n(\hat{\theta}_n) - \text{E } \ell_n(\theta_0) \rightarrow \frac{1}{2} \text{Tr}[J(\theta_0)^{-1}J(\theta_0)] = p/2,$$

implying

$$\text{E } \ell_n(\hat{\theta}_n) \approx n \text{E } \log f_{\theta_0} + p/2.$$

By the results of Section 3.1.2,

$$\text{E AIC} \approx -n \text{E } \log f_{\theta_0}(Z) + p/2 = -n \text{E } \log f_{\theta_0}(Z) + nR(T),$$

where $R(T)$ is the expected Kullback-Leibler divergence from $f_{\hat{\theta}_n}$ to f_{θ_0} , see Section 3.1.2. Hence, up to a constant and scaling factor, the AIC is an approximately unbiased estimator of the risk!

The above arguments work also when the model is misspecified. In this case, one finds

$$\text{E } \ell_n(\hat{\theta}_n) \approx n \text{E } \log f_{\theta_{\text{lf}}}(Z) + \frac{1}{2} \text{Tr}[J(\theta_{\text{lf}})^{-1}K(\theta_{\text{lf}})], \quad (3.6)$$

where $J(\theta) = -H_\theta \text{E } \log f_\theta(Z)$ and $K(\theta) = \text{Var } \nabla_\theta \log f_\theta(Z)$. The chain of arguments is completely analogue to the case where $g = f_{\theta_0}$ so I leave out the details. The approximation in (3.6) justifies the model robust, and sadly also a lot less known, version of the AIC called Takeuchi's information criterion (TIC).⁵ It takes the following form

$$\text{TIC} = -\ell_n(\hat{\theta}_n) + \text{Tr}[\hat{J}^{-1}\hat{K}], \quad (3.7)$$

where $\hat{J}$ and $\hat{K}$ are consistent estimators of $J(\theta_{\text{lf}})$ and $K(\theta_{\text{lf}})$ respectively. Provided sufficient regularity, the above calculations reveal

$$\text{E TIC} \approx -n \text{E } \log f_{\theta_{\text{lf}}}(Z) + \frac{1}{2} \text{Tr}[J(\theta_{\text{lf}})^{-1}K(\theta_{\text{lf}})] = nR(T) - n \text{E } \log g(Z),$$

which again is only a constant and scaling factor away from the true risk. Since the constant $-n \text{E } \log g(Z)$ is not dependent on the model fitted, the TIC misses its target by a number which is equal for all candidate models. It can therefore be safely used to compare different models in a robust fashion. A low value of KL suggest that the model fits the data well “overall”. Sometimes, however, a good model fit is not the sole goal of an analysis, and in such cases a more specialised selection criterion is needed. I will now go through one such option.

⁵A possible reason for the TIC relatively modest fame might be that it was originally published in Japanese (Takeuchi 1976). For a translated version, see Takeuchi 2020.

3.2.2 The focused information criterion (FIC)

Assume one has a data set and multiple plausible models for the data, e.g. one might be unsure if data is Weibull or Gamma distributed, or one might have a regression model where it is not clear which covariates are relevant. In such situations, one needs a way of choosing between the different candidate models. One way to do this, could be to compute the AIC or TIC discussed in the previous subsection for each available model, and choose the one with the lowest value. This procedure is sound if one wishes to find the model with the lowest KL divergence (or equivalently the candidate with the smallest information loss), i.e. the model which fit the data best “overall”. In some applications, however, overall model fit is not really the main goal of the analysis. One might instead be more interested in the median or mean in the population one is modelling, or one might fit a regression model to ascertain the effect a covariate has on a single prediction. In such situations, an option is to use the focused information criterion, or the FIC for short, an information criterion created for this very purpose.

Unlike the AIC or the TIC, the FIC does not rank model according to how well they fit the data “overall”. Instead, models which are well-suited for the application at hand are preferred. This notion of being “well-suited” is formalised through the concept of a focus parameter, a parameter which must be chosen to reflect the goal of the analysis before starting the model evaluation procedure. If a model is fitted with the goal of making inference about median in the underlying distribution, the median should be chosen as the focus parameter. If the mean is of interest, the mean should be used. In general, the focus parameter should be set by the researcher doing the analysis, and should contain the most relevant information which is to be inferred from the data.

For a given focus parameter μ , the FIC ranks models according to the estimated MSE of the corresponding estimator of μ . Hence, for a parametric model $\mathcal{F}_\theta = \{f_\theta | \theta \in \Theta\}$ for which μ takes the value $h(\theta)$, the FIC is an estimate of the MSE of $h(\hat{\theta}_n)$ where $\hat{\theta}_n$ is an estimator of the parameter θ . Typically, $\hat{\theta}_n$ is an ML estimator, but this is not always the case, see e.g. **Paper II**.

The FIC was first introduced in Claeskens and Hjort 2003 (see also Hjort and Claeskens 2003a,b), with a number of important extensions following closely after this paper. For instance, Claeskens et al. 2006 considers other risk functions than the MSE and Zhang and Liang 2011 extends the framework to cover generalised additive partial linear models. For an overview and introduction to the FIC, I refer to the textbook Claeskens and Hjort 2008, in particular to Chapter 6. In the original definition, the FIC was only formulated for nested models, and a very important step forward was made in Jullum and Hjort 2017 where the authors extend the framework to non-nested models. The formulas given in this paper are, however, not precise to a sufficiently small order, and in **Paper I** the expressions are made more sharp. I will now informally derive a version of the FIC based on arguments given in this article. For more details, consult **Paper I** in addition to the aforementioned Jullum and Hjort 2017.

Let $Z_1, \dots, Z_n$ be i.i.d. and $\hat{\mu}_1 \in \mathbb{R}$ some estimator of a parameter with true value μ_0 based on these data. Assume further that this estimator has an influence

function IF_1 , so that one can write $\hat{\mu}_1$ in the following way

$$\hat{\mu}_1 = \mu_1 + n^{-1} \sum_{i=1}^n \text{IF}_1(Z_i) + \epsilon_n$$

for some $\epsilon = O_p(1/n)$ and $\mu_1 \in \mathbb{R}$ which is not necessarily equal to μ_0 . By the arguments given in the beginning of this chapter, the MSE of $\hat{\mu}_1$ is equal to

$$\text{MSE}(\hat{\mu}_1) = (\mu_1 - \mu_0)^2 + n^{-1} \left[2(\mu_1 - \mu_0)c_1 + \frac{1}{2} \text{Tr } V \right] + o(n^{-1}), \quad (3.8)$$

where $V = \text{Var IF}(Z)$ and c_1 is the limit of $n \text{E } \epsilon_n$ (assuming it exists). The FIC works by approximating all terms in the above expression, creating an estimator which is only $o(1/n)$ away from the true value of $\text{MSE}(\hat{\mu}_1)$ in expected value.

The variance matrix V can be estimated by $\hat{V}$, the empirical variance of $\text{IF}(Z_1), \dots, \text{IF}(Z_n)$. Estimators of c_1 are also relatively easy to obtain, see **Paper I** or alternatively Firth 1993 and Kim 2016 for methods to ensure that c_1 is equal to zero. The part of (3.8) which is hard to estimate, turns out to be the term which at first glance might look like the easiest part: $b^2 = (\mu_1 - \mu_0)^2$. Since $\hat{\mu}_1$ is consistent for the possibly incorrect value μ_1 only, there is no way of knowing how far off the limit is from the true value μ_0 . To be able to quantify this, one needs to introduce a second estimator $\hat{\mu}_0$.

I will assume that $\hat{\mu}_0$ is consistent for μ_0 and that one can write

$$\hat{\mu}_0 = \mu_0 + n^{-1} \sum_{i=1}^n \text{IF}_0(Z_i) + \delta_n,$$

for some $\delta_n = O_p(1/n)$ and influence function IF_0 . Then, after employing multiple regularity conditions, direct computations show

$$\text{E}(\hat{\mu}_1 - \hat{\mu}_0)^2 = b^2 + n^{-1}(\kappa^2 + 2bc_1 - 2bc_0) + o(1/n),$$

where $\kappa^2 = \text{Var}[\text{IF}_1(Z) - \text{IF}_0(Z)]$ and $c_0 = \lim_{n \rightarrow \infty} n \text{E } \delta_n$. Hence, the following estimator is an approximately unbiased estimator for $(\mu_1 - \mu_0)^2 + 2bc_1/n$:

$$\hat{b}^2 = (\hat{\mu}_1 - \hat{\mu}_0)^2 + n^{-1}[2(\hat{\mu}_1 - \hat{\mu}_0)\hat{c}_0 - \hat{\kappa}^2],$$

where $\hat{\kappa}$ is the empirical variance of $\text{IF}_1(Z_1) - \text{IF}_0(Z_1), \dots, \text{IF}_1(Z_n) - \text{IF}_0(Z_n)$ and $\hat{c}_0$ are either of the estimators defined in **Paper I**. This ensures that the MSE of $\hat{\mu}_1$ can be estimated approximately unbiased by $\hat{b}^2 + \hat{V}/n$.

When $\hat{\mu}_1$ is an estimator stemming from a model, that models FIC score is the estimated MSE of $\hat{\mu}_1$. Hence, the FIC works by precisely estimating the MSE of the corresponding estimators of the focus parameter and ranking models accordingly. In **Paper II** the FIC is used to choose the optimal balance between parametric and non-parametric in a specific setting, and in **Paper III** the effect of tuning models with respect to the FIC is investigated.

Both the AIC, TIC and FIC are based on theoretical results and computations. When employable, these criteria are quick and theoretically justifiable to use. Not all risk functions and estimators fit into the framework of these information criteria, however, and in such situations alternative, more general methods need to be used. One option is to use the cross validation, a method based on resampling which can be employed in very general settings. I will now move on to discuss this risk estimator.

3.2.3 Cross-validation (CV)

I will now give a brief presentation of the very popular and well-known tool cross-validation, which is a major topic of **Paper III**. For a more details, consult Hastie et al. 2009, Ch. 7.

Assume $Z_1, \dots, Z_n$ are i.i.d. data points where $Z_i = (Y_i, X_i^T)^T$ for each i , and that $\hat{Y}$ is a prediction or classification model returning an estimate of the associated value of y given a co-variate x . This function $\hat{Y}$ could be the prediction in a linear regression model, e.g. $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}^T x$ for some fitted coefficients $(\hat{\beta}_0, \hat{\beta}^T)^T$, the classification in a fitted binary logistic regression model, e.g. $\hat{Y}(x) = \operatorname{argmax}_k \Pr_{\hat{\beta}}(Y = k | X = x)$, where $\Pr_{\beta}(Y = 1 | X = x) = [1 + \exp(-\beta_0 - \beta^T x)]^{-1}$, or something more complicated like the prediction returned by a fitted smoothing spline or neural net. Call the fitting procedure T so that $\hat{Y}(x) = T(x, Z_1, \dots, Z_n)$. Suppose further that one wishes to estimate how well this fitting procedure performs on “average” with respect to a loss function L . A natural way to evaluate this is the EPR discussed in Section 3.1.1 which tells us exactly how much the fitting procedure used to create $\hat{Y}$ is expected to loose when repeated on new data sets.

A simple way to estimate $\operatorname{EPR}(T)$, is to use what I will call the training error (TE), defined as

$$\operatorname{TE}(T) = n^{-1} \sum_{i=1}^n L[Y_i, \hat{Y}(X_i)].$$

This estimator averages the loss incurred by using $\hat{Y}(X_i)$ in place of Y_i over the data set, which “should” in some sense measure how much one loses on average by using $\hat{Y}$. In the above, however, the function $\hat{Y}$ is kept still, and hence, the randomness introduced by fitting this prediction or classification model is not taken into account. For this reason, $\operatorname{TE}(T)$ cannot possibly be a good way to estimate $\operatorname{EPR}(T)$.

If multiple independent data sets $\mathcal{Z}_1, \dots, \mathcal{Z}_B$ where $\mathcal{Z}_b = \{(Y_i^b, (X_i^b)^T)^T\}_{i=1}^n$ for $b = 1, \dots, B$ were available, one could use T on each data set to get B “samples” of the function $\hat{Y}(\cdot)$. This could be used to create an estimator of $\operatorname{EPR}(T)$. One does, however, need to be a bit careful. The definition of $\operatorname{EPR}(T)$ is

$$\operatorname{EPR}(T) = \mathbb{E}_{Z_1, \dots, Z_n} \mathbb{E}_{(Y, X)} \{L[Y, T(X, Z_1, \dots, Z_n)] | Z_1, \dots, Z_n\},$$

where (Y, X) is assumed to be independent of $Z_1, \dots, Z_n$. Because of this, $(Bn)^{-1} \sum_{b=1}^B \sum_{i=1}^n L[Y_i^b, T(X_i^b, \mathcal{Z}_b)]$ or similar estimators cannot be used to approximate $\operatorname{EPR}(T)$, as $T(\cdot, \mathcal{Z}_b)$ is not independent of $(Y_i^b, (X_i^b)^T)^T$. A way to solve this problem could be to set one data set aside and evaluate each fitted model on this data set. In symbols, this means using the approximation

$$(B-1)^{-1} \sum_{b=1}^{B-1} n^{-1} \sum_{i=1}^n L[Y_i^B, T(X_i^B, \mathcal{Z}_b)]. \quad (3.9)$$

(3.9) is a nested mean, and “should” converge to its expected value, $\operatorname{EPR}(T)$, in most cases.

The calculations in the previous paragraphs are all well and good, but (3.9) can only be used when B independent data sets are available. In practice, this is rarely the case, as one in such a situation would probably just merge all the $\mathcal{Z}_b$ s into one large data set and create a prediction or classification model based on this larger set. Luckily, there is a way to “cheat” and approximate (3.9) without B different data sets. This way is called cross-validation.

Cross-validation works by taking a data set $\mathcal{Z} = Z_1, \dots, Z_n$ and splitting it into subsets $\mathcal{Z}_1, \dots, \mathcal{Z}_B$ called “folds”. For each of these folds, the model is fitted to the data points not present in that subset. Afterwards, the average loss of the fitted model is computed on the fold held out. This is repeated for all $\mathcal{Z}_b$ s and the risk is estimated by the average of the result. In symbols this means

$$\text{CV}(T) = B^{-1} \sum_{b=1}^B |\mathcal{Z}_b|^{-1} \sum_{i=1}^{|\mathcal{Z}_b|} L[Y_i^b, T(X_i^b, \mathcal{Z}_{-b})], \quad (3.10)$$

where $\mathcal{Z}_{-b} = \bigcup_{b' \neq b} \mathcal{Z}_{b'}$ and $|\mathcal{Z}_b|$ denotes the number of elements in the set $|\mathcal{Z}_b|$. When $B = n$ and $\mathcal{Z}_i = \{Z_i\}$, $\text{CV}(T)$ is called the leave-on-out cross-validation (LOOCV). When $B = n/K$ for some K and $\mathcal{Z}_b = \{Z_i, |i = (b-1)K, \dots, bK\}$, $\text{CV}(T)$ is called the K-folded cross validation.

The LOOCV was introduced independently by Stone 1974 and Allen 1974, although some rudimentary versions predates these sources (see the historical remarks in Arlot and Celisse 2010). The paper Geisser 1975 is also often cited as an inventor, as other schemes for omitting observations than the LOOCV are considered here. This includes the K-folded CV discussed above, although Geisser does not use this term.

By the previous arguments, it seems reasonable to think that $\text{CV}(T)$ estimates $\text{EPR}(T)$. Taking a second look at (3.10), however, reveals that this is not so immediate at one might first think. Even though the expression is a mean, the components of the outer sum are not independent. This is because T is applied to $\mathcal{Z}_{-b}$ in each term, and these sets overlap and are therefore dependent. This makes convergence of $\text{CV}(T)$ towards $\text{EPR}(T)$ a lot less trivial than for (3.9). For this reason, it is sometimes recommended to increase the size of the folds to reduce dependence and improve the estimator $\text{CV}(T)$. This can reduce variance, but will in turn introduce bias as the predictor or classifier $\hat{Y}$ will be applied to fewer data points in each iteration and hence the expected prediction risk of T when applied to $|\mathcal{Z}_{-b}|$ rather than $Z_1, \dots, Z_n$ will be estimated. See Hastie et al. 2009, Ch. 7.10 for more discussions.

It also is worth noting that in the above I started with the EPR as the target. This is natural if one wishes to estimate the quality of an estimation procedure, but might be of less interest in specific applications if one wishes to evaluate how well a fixed and fitted model performs. By the previous informal arguments, however, this quantity is not what CV estimates! For nuances in these two goals, see Hastie et al. 2009, Ch. 7 or the excellent discussion in Bates et al. 2024.

Although I used the framework of prediction and classification in the above exposition, there is nothing in the definition of CV which requires such a setting. Indeed, let L be any function depending on z and a parameter θ , and assume

there is some estimation procedure T such that $\hat{\theta}(z) = T(z, Z_1, \dots, Z_n)$. Then, $E L[Z, \hat{\theta}(Z)]$ where $Z, Z_1, \dots, Z_n$ are i.i.d. can be estimated by

$$\text{CV}(T, L) = B^{-1} \sum_{b=1}^B |\mathcal{Z}_b|^{-1} \sum_{i=1}^{|\mathcal{Z}_b|} L[Z_i^b, T(Z_i^b, \mathcal{Z}_{-b})].$$

In **Paper III**, the above definition of CV in parametric contexts is studied, when the leave-one-out procedure is used.

Lastly, I want to remark that there exist lots and lots of variations on cross-validation, e.g. the 0.632 method of Efron and Tibshirani 1997 which combines CV with the bootstrap, or repeated learning-testing (Breiman et al. 1984) and Monte-Carlo CV (Picard and Cook 1984), where splits are repeatedly and randomly created without and with replacement respectively. The Generalized CV (GCV), introduced in Craven and Wahba 1979, is another important example, where the LOOCV of linear fitting models⁶ under squared-error loss, is approximated by an elegant expression which is easy to compute. For an overview, consult Arlot and Celisse 2010, which also contains multiple references to theoretical developments. This paper is, however, relatively old in a field which moves so quickly and therefore does not include some of the more recent methodological developments. For instance, some authors have worked with developing theory to better understand the asymptotic properties of CV and create confidence intervals for their risk based on the technique, see e.g. Bayle et al. 2020 or Bates et al. 2024, in addition to the sources given therein.

⁶As defined in Hastie et al. 2009, p. 244

Chapter 4

Summary of papers

In Geisser 1975, p. 320, the author writes the following about the paper Stone 1974:

And so it goes. Indeed the step- by-step evolution of the method as discussed by him is really a model in many ways for how research should proceed. That Stone actually advanced along these lines is doubtful, but it is conceivable, he being scholar as well as researcher.

In the previous chapters I have attempted to do as Stone and give a step-by-step evolution of the background material for this thesis, and when I now present summaries of each article in the thesis, I hope it will be clear where in the field my research fits in.

4.1 Paper I - Accurate bias estimation with applications to focused model selection

Dæhlen, I., Hjort, N. L. and Hobæk Haff, I. (2023). ‘Accurate bias estimation with applications to focused model selection’. In: *Scandinavian journal of statistics* 51.2, pp. 724–759.

In this paper, we derive formulas for the second order bias of a large class of estimators, as well as approximations thereof. We use these results to create approximations to the MSE, which are correct to the order $o(1/n)$ in expected value. Such precision is necessary to make sure that the variance is properly taken into account as this term is typically of size $O(1/n)$. Furthermore, we use the approximations to derive a new and more precise version of the FIC.

In the paper, we work with the bias of estimators on the form

$$\hat{\theta}_n = \theta_0 + n^{-1} \sum_{i=1}^n \phi(Z_i, \theta_0) + \epsilon_n,$$

for some function ϕ . This includes many popular estimators like means, medians and ML estimators. In particular, we take a second look at the remainder term ϵ_n , and give conditions for when $n E \epsilon_n$ converges towards a fixed quantity c .

In additions, we give formulas for and estimators of this c for many types of estimators, including means, quantiles, ML estimators and functions thereof. In particular, we show a result characterising c when $\hat{\theta}_n$ is an ML estimator in a possibly misspecified parametric model. This is an extension of the result of Cox and Snell 1968, which covers correctly specified models only.

We use the estimators of c to define approximations to the MSE of $\hat{\theta}_n$ with expected errors of order $o(1/n)$. This is done as explained in Section 3.2.2, and leads to a new version of the FIC extending that of Jullum and Hjort 2017. In the 2017 paper, the following criterion is defined

$$\text{FIC}_{\text{old}} = \max\{0, \hat{b}^2 - \hat{\kappa}/n\} + \hat{\tau}^2/n,$$

where $\hat{b}$, $\hat{\kappa}$ and $\hat{\tau}$ are estimators given in the paper. We instead propose

$$\text{FIC}_{\text{new}} = \max\{0, \hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}/n\} + \hat{\tau}^2/n,$$

where $\hat{c}$ is either of the estimators of c derived in the article. Unlike FIC_{old} , FIC_{new} has expected error of order $o(1/n)$, which is the precision required to properly take the variability of $\hat{\theta}_n$ into account. See Section 3.2.2 for more details.

4.2 Paper II - Model robust hybrid likelihood

Dæhlen, I. and Hjort, N. L. (2024). ‘Model robust hybrid likelihood’. *Submitted for publication*.

In the second paper, we study a strategy for fitting parametric models introduced in Hjort et al. 2018. This technique is called hybrid likelihood and is based on combining the classic parametric likelihood with the empirical likelihood function developed by Art B. Owen in Owen 1988, 1990, 1991, and defined in the following way. For i.i.d. data $Z_1, \dots, Z_n$ from a distribution F , the empirical likelihood function for the solution to $\text{E}m(Z, \mu) = 0$ where $m: \mathbb{R}^{d+q} \rightarrow \mathbb{R}^s$ is some function and $Z \sim F$ is defined as

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

One can show that $\text{EL}_n(\mu)$ is large when μ is close to solving the equation $n^{-1} \sum_{i=1}^n m(Z_i, \mu) = 0$. Furthermore, Owen 1988 showed that $-2 \log \text{EL}_n(\mu_0) \xrightarrow{d} \chi_q^2$ for the solution μ_0 to the equation $\text{E}m(Z, \mu) = 0$. This can be seen as an analogue to Wilks’ theorem in parametric likelihood theory, which states that $-2[\ell_n(\hat{\theta}_n) - \ell_n(\theta_0)] \xrightarrow{d} \chi_q^2$ when $\hat{\theta}_n$ is an ML estimator in a correctly specified parametric model with log-likelihood function ℓ_n and true value of θ equal to θ_0 , see e.g. Schweder and Hjort 2016, Ch. 2. For this reason the empirical likelihood function can be seen as a nonparametric counterpart to the classic parametric likelihood.

4.3. Paper III - The asymptotic effect of tuning parameters

For a given parametric model f_θ for $\theta \in \Theta \subseteq \mathbb{R}^p$ and a control parameter $\mu \in \mathbb{R}^q$ taking the value $\mu(\theta)$ in the parametric model, Hjort et al. 2018 defines the hybrid likelihood function as

$$\text{HL}_n(\theta) = \left(\prod_{i=1}^n f_\theta(Z_i)^{1-a} \right) \text{EL}_n[\mu(\theta)]^a,$$

for every fixed value of the balance parameter $a \in [0, 1)$. Furthermore, the authors suggest fitting the parametric model by maximising HL_n . The likelihood function is large at values of θ for which the model f_θ fits the data well “overall”. EL_n , on the other hand, is maximised when $\mu(\theta)$ is close to the solution to $n^{-1} \sum_{i=1}^n m(Z_i, \mu) = 0$. Hence, when a is close to zero, maximising HL_n ensures that the fitted model will fit the data well overall, while an a close to one would ensure that f_θ is fitted in a way ensuring that the corresponding model’s estimate of the control parameter is robust. For intermediate values of a , both overall model fit and model robust estimation of μ will be taken into account.

Hjort et al. 2018 studies the behaviour of the maximiser of HL_n , $\hat{\theta}_n$, when the model f_θ is assumed to be correctly specified. With this assumption, they show that $\hat{\theta}_n$ is consistent for the true value of θ , θ_0 , and that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in law towards a central normal distribution. Since the hybrid likelihood framework is created as a way to ensure model robust estimation of the control parameter, however, the assumption of correct model specification is a hard pill to swallow. Why use hybrid likelihood to ensure robustness if there is nothing wrong with the model? In **Paper II**, we address this issue by deriving results for $\hat{\theta}_n$ also in the case of possible model misspecification. We show that $\hat{\theta}_n$ is consistent for the minimiser of a well-defined divergence function as well as $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V)$, where V is a variance matrix given in the paper. In addition, our framework allows for plug-in estimators in the empirical likelihood function, i.e. that $\mu(\theta)$ can be replaced by $(\mu(\theta), \hat{P})$ where $\hat{P}$ is some Z-estimator.

For the hybrid likelihood framework to be useful in practice, one needs a way to decide on a balance parameter. Ideally, this should be done in a data-driven way, where a is set to a value close to zero if the parametric model fits the data well, and closer to one otherwise. In **Paper II**, we propose doing this by using the FIC derived in **Paper I**. Using this method, researchers can, with relative ease, choose a good hybrid model, be it purely parametric ($a = 0$) or partly parametric, partly nonparametric ($a > 0$).

4.3 Paper III - The asymptotic effect of tuning parameters

Dæhlen, I., Hjort, N. L. and Hobæk Haff, I. (2024). ‘The asymptotic effect of tuning parameters’. *Submitted for publication*.

In **Paper II**, we derived limiting results for the maximizer of the hybrid likelihood function. These theorems hold true for every fixed value of the balance parameter a . We also gave a way to estimate the optimal a from data: $\hat{a}$. By fixing a at

this optimal value $\hat{a}$, we get a new estimator $\hat{\theta}_n(\hat{a})$ to which the point-wise results of **Paper II** cannot be applied. In the third paper of this thesis, we attempt to correct this by deriving results taking the process of estimating a into account.

In general, we consider classes of estimators $\hat{\theta}_n(\lambda) \in \mathbb{R}^p$ defined as

$$\hat{\theta}_n(\lambda) = \text{root}_{\theta} \text{ of } \sum_{i=1}^n \varphi(Z_i, \theta, \lambda),$$

for every value of a tuning parameter $\lambda \in \Lambda \subseteq \mathbb{R}^q$. Furthermore, we assume that a value of λ is chosen by minimizing some approximation to the risk: $\hat{\lambda}$. What is shown in **Paper III**, is that under relatively general conditions, $\hat{\theta}(\hat{\lambda})$ is consistent for a natural quantity $\theta(\lambda_0)$, where $\theta(\lambda)$ is the solution to $0 = \mathbb{E} \varphi(Z, \theta, \lambda)$ and λ_0 is “optimal” in the limit. Furthermore, we prove that $\sqrt{n}[\hat{\theta}_n(\hat{\lambda}) - \theta(\lambda_0)]$ converges in law towards a central normal distribution. With these results one can make valid asymptotic inference about θ in a way taking the added randomness by estimating λ into account. This includes the maximiser of the hybrid likelihood function when the balance parameter is chosen as proposed in **Paper II**.

The framework described in the previous paragraph often shows up in regression and classification settings. In, for instance, ridge regression, the regression coefficients are found by minimising $n^{-1} \sum_{i=1}^n (Y_i - \beta_0 - \beta^T X_i)^2 - \lambda \|\beta\|^2$. In such settings, $\hat{\lambda}$ is often found by minimising the cross-validated estimator of the risk discussed and defined in Section 3.2.3, and for this reason, the procedure for estimating λ should allow for cross-validation procedures in order to be really useful in practice. In **Paper III** this is ensured by deriving an alternative characterisation of the LOOCV. In particular, we show that

$$\sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = \sum_{i=1}^n \psi(Z_i, \hat{\theta}) + \text{Tr}(J^{-1}C) + o_p(1), \quad (4.1)$$

for matrices J and C defined in the paper and when $\hat{\theta}$ and $\hat{\theta}_{(-j)}$ are the roots of $\sum_{i=1}^n \varphi(Z_i, \theta)$ and $\sum_{i \neq j} \varphi(Z_i, \theta)$ respectively. This expression is a generalisation of the result of Stone 1974, which shows (4.1) when φ is the score function in a correctly specified model and $\nabla_{\theta} \psi = \varphi$.

Chapter 5

Discussion

A major theme of this thesis is the concept of “precision”. **Paper I** defines precise approximations to the bias and MSE, **Paper II** investigates how the most precise in a class of estimators can be chosen, and in **Paper III**, precisely estimating the effect of tuning parameters is considered. As a mathematician, this word “precise” has a very technical meaning to me, and when I say that a result is “precise” I usually mean that a specific remainder term is of this and that order. In daily speech, however, the concept of precision is much broader than this, and to be “precise” means to be accurate and in some sense “error-free”. In this chapter I will attempt to address a couple of ways in which I believe my own research fails to be precise. Since this is in some sense a critique of my own work, I will be a lot more opinionated than previously in this thesis and give my own subjective thoughts.

5.1 Can precision be precisely estimated?

The two very first papers I read in relation to this PhD project was Efron 1975 and O’Neill 1980. In these articles, the authors investigate the performance of generative and discriminative classifiers. When Y is a categorical variable taking K different values, discriminative classification procedures are methods which estimate $\Pr(Y = k | X = x)$ for $k = 1, \dots, K$ directly from data. Generative classification methods instead estimate the full distribution of $\Pr(Y, X)$ and deduce $\Pr(Y = k | X = x)$ for $k = 1, \dots, K$ therefrom. Both classes of procedures attempt to approximate Bayes’ classification rule by classifying a response $Y(x)$ as the maximiser of $\Pr(Y = k | X = x)$ for $k = 1, \dots, K$. Furthermore, for every generative model, which is fit by maximising the full log-likelihood $\ell_G(\theta) = \sum_{i=1}^n \log f_\theta(Y_i, X_i)$, there is a corresponding discriminative procedure, which maximises the discriminative likelihood $\ell_D(\theta) = \sum_{i=1}^n \log \{ [\sum_{k=1}^K I(Y_i = k) f_\theta(k, X_i)] / [\sum_{k=1}^K f_\theta(k, X_i)] \}$.

Efron 1975 and O’Neill 1980 find that using a correctly specified generative model grossly outperforms the discriminative counterpart. Building on this, the first idea for this PhD project was to create estimators of the error rate and use these approximations to choose between generative and discriminative models in a data-driven way. As one might guess from the fact that there is no paper

in this thesis titled “Precise selection of generative and discriminative models”, however, this idea did not really go anywhere. Creating estimators for the expected risk is not too complicated. Using the approximations in a meaningful way, however, turned out to be more difficult than first anticipated. To see this, assume $Z_1, \dots, Z_n$, are i.i.d. following some distribution g , with $Z_i = (Y_i, X_i^T)^T$ for each i where Y_i is a binary response variable and X_i a vector of covariates. Furthermore, consider parametric classifiers $\hat{Y}_\theta$, and let L be the 0-1- loss function $L(y, \hat{y}) = I(y \neq \hat{y})$. By Section 3.1.1,

$$\text{EPR}(T) = \text{EPR}(\theta_{\text{lf}}) + n^{-1} \left(\nabla \text{EPR}(\theta)^T c + \frac{1}{2} \text{Tr}[H_{\theta_{\text{lf}}} \text{EPR}(\theta) V] \right) + o(1/n), \quad (5.1)$$

when $\hat{\theta}_n = T(Z_1, \dots, Z_n)$ is some estimator of θ consistent for θ_{lf} and with the property $\sqrt{n}(\hat{\theta}_n - \theta_{\text{lf}}) \xrightarrow{d} N(0, V)$. A special case of this expansion was shown in O’Neill 1980 when $\hat{\theta}_n$ is an ML estimator resulting from fitting either a correctly specified generative model or a corresponding discriminative one. Among other things O’Neill shows that when the model is specified correctly, the ML estimator of θ in the generative model, $\hat{\theta}_n^G$, as well as the estimator obtained from the discriminative counterpart, $\hat{\theta}_n^D$, converge in probability towards the value θ_0 , which makes the corresponding classifiers into Bayes’ classification rule, the optimal solution to the classification problem, see e.g. Hastie et al. 2009, Se. 2.4. Because of this, the first terms in (5.1) are equal for both $\hat{\theta}_n^G$ and $\hat{\theta}_n^D$. Furthermore, as both the generative and logistic regression classifier is Bayes’ classification rule at the limit, the EPR attains its minimum at θ_0 in a correctly specified model. For this reason, the first term within the parenthesis in (5.1) disappears, and the limiting risk of the two methods differ only in the $n^{-1}(1/2) \text{Tr}[H_{\theta_{\text{lf}}} \text{EPR}(\theta) V]$ -term. The two methods can therefore be compared by computing this value for both procedures. Doing this, O’Neill finds that, in his words (O’Neill 1980, p. 157),

The poor performance for situations in which good discrimination is possible casts doubt on the use of logistic regression discrimination and suggests that the maximum likelihood estimator of the optimal discriminant rule for the specific distributions at hand should be used whenever possible.

Here the ML estimator refers to $\hat{\theta}_n^G$ while by “logistic regression discrimination”, O’Neill means the corresponding discriminative model. Efron 1975 who studies the special case of linear discriminant analysis (LDA) vs the alternative of classic logistic regression (LR), arrives at a similar conclusion.

But how different are really these two methods? By simply fitting LR and LDA models to some simulated data, one quickly finds that the two procedures give very similar results. This is illustrated in Figure 1, where I have plotted the decision boundaries in fitted LDA and LR models based on 50 simulated data points.¹ Looking at the plot, one notices that there is virtually no difference

¹Fifty data points $(Y_1, X_1^T)^T, \dots, (Y_{50}, X_{50}^T)^T$ were simulated from the following model. $Y \sim \text{Bernoulli}(1/2)$ and $X|Y = j \sim N(\mu_j, \Sigma)$, where $\mu_0 \approx (1.237, 1.237)$, $\mu_1 = -\mu_0$ and Σ is equal to the identity matrix in $\mathbb{R}^{2 \times 2}$. These parameters were chosen to be able to compare the results with the table in Efron 1975

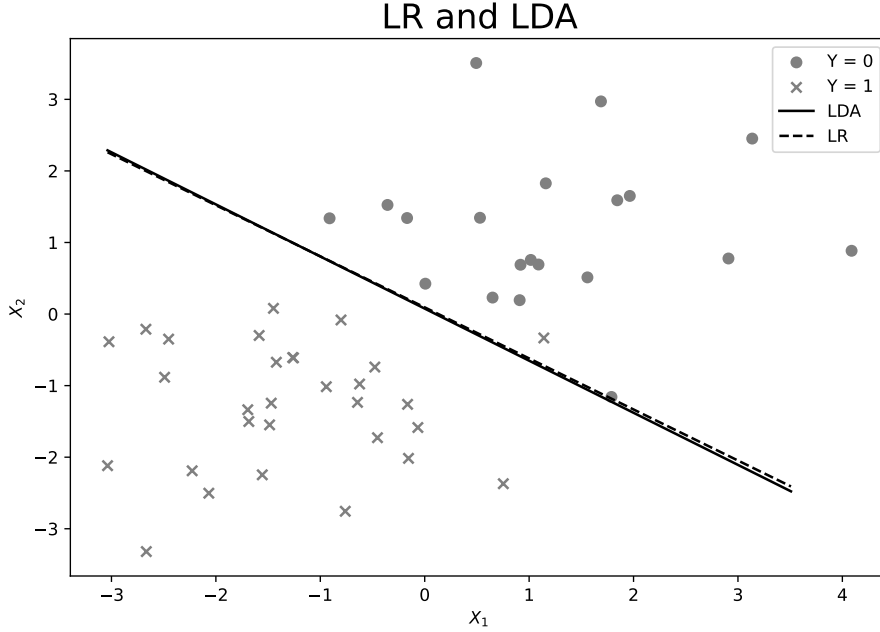


Figure 1: The figure shows a randomly drawn data set from an LDA model together with the decision boundaries obtained from fitted LR and LDA models.

between the two methods in this case. According to Efron 1975 and O’Neill 1980, however, LDA should be a little more than twice as efficient as logistic regression for this simulation setting. So what is happening here? Well, Efron and O’Neill are comparing the models in terms of relative efficiency, which means they are computing

$$\frac{\text{EPR}_{\text{LDA}} - \text{EPR}(\theta_0)}{\text{EPR}_{\text{LR}} - \text{EPR}(\theta_0)} = \frac{\text{Tr}[BV_{\text{LDA}}]}{\text{Tr}[BV_{\text{LR}}]}, \quad (5.2)$$

where EPR_{LR} and EPR_{LDA} are the EPRs when θ is estimated using LR and LDA respectively, $B = H_{\theta_0} \text{EPR}(\theta)$ and V_{LR} and V_{LDA} are the limiting variances of $\hat{\theta}_n^G$ and $\hat{\theta}_n^D$ scaled by $1/n$. Computing (5.2) is in some sense natural if one wishes to calculate the limiting efficiency loss, but the shifting and implicit scaling in (5.2), makes the comparison quite misleading. It is true that for the simulated setting in Figure 1, the relative efficiency loss as defined in (5.2) is about 0.486, which is what Efron 1975 finds. That being said, the true values of EPR_{LR} and EPR_{LDA} are 0.048 and 0.044 respectively, numbers which are very close and almost indistinguishable in most settings. I therefore believe that the results of Efron 1975 and O’Neill 1980, although mathematically correct, are presented in a way which puts generative models in a much better light than what is really the case.

The lack of actual difference between generative and discriminative models is, however, not the only reason why this attempt at a PhD project did not revolutionise the academic community. By the calculations of O’Neill 1980, using a generative model is always preferable to the discriminative counterpart when the model is specified correctly. Because of this, estimating EPR and using the approximation to choose between the two approaches is really only necessary

when one is not willing to make this assumption. In this case, however, precisely estimating the EPR becomes a lot more difficult than in the case of correct model-specification. This is because, when the model is not correctly specified, the first term in (5.1) needs to be estimated.

Looking at (5.1), one notices that the expansion consists of a constant term $\text{EPR}(\theta_{\text{f}})$, a term of size $O(1/n)$ and a remainder term of smaller order. By the previous arguments, the difference between generative models and the discriminative counterpart is of size $O(1/n)$ when the model is correctly specified. For this reason, it is necessarily to approximate the EPR with errors of at most order $o(1/n)$ for the estimator to be able to work well in all settings and be truly model-agnostic. This is relatively easy to achieve for the $O(1/n)$ -term in (5.1). Since this term is divided by n , it suffices to find consistent estimators of all quantities inside the parenthesis and plug them into the expression in (5.1). This is of course not trivial, but in principle relatively easy since the n^{-1} factor in the $O(1/n)$ -term in (5.1) will ensure that the resulting estimator has an error of size $o_p(1/n)$. This is sufficiently precise. When it comes to estimation of the first term in (5.1), however, one runs into problems.

The estimator of $\text{EPR}(\theta_{\text{f}})$ needs to be precise to the order of $o(1/n)$. In Section 3.2.2, a similar property needed to hold for the estimator of the limiting bias in construction of the FIC. There, this was achieved by creating an approximation which had expected error of size $o(1/n)$. An idea could therefore be to proceed in a similar manner and try to define an approximately unbiased estimator of $\text{EPR}(\theta_{\text{f}})$. My first attempt at this involved the cross-validation procedure described and discussed in Section 3.2.3. The LOOCV estimator of the EPR would in this case be

$$\text{LOOCV} = n^{-1} \sum_{i=1}^n I[Y_i \neq \hat{Y}_{-i}(X_i)],$$

where $\hat{Y}_{-i}(x) = T_{n-1}(x, Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$ is the classifier one gets after fitting the model to all but the i -th data point. Since the Z_i s are i.i.d. we get

$$\begin{aligned} \text{ELOOCV} &= n^{-1} \sum_{i=1}^n \Pr[Y \neq \hat{Y}_{-i}(X) \mid \hat{Y}_{-i}] \\ &= \mathbb{E} \Pr[Y \neq \hat{T}_{n-1}(X, Z_1, \dots, Z_{n-1}) \mid Z_1, \dots, Z_{n-1}] \\ &= \mathbb{E}_{Z_1, \dots, Z_{n-1}} \mathbb{E}_Z[I(Y \neq T_{n-1}(X, Z_1, \dots, Z_{n-1}) \mid Z)] \\ &= \text{EPR}(T_{n-1}) \end{aligned}$$

where $Z, Z_1, \dots, Z_{n-1}$ are i.i.d.. Now if the classifiers $\hat{Y}_n$ and $\hat{Y}_{n-1}$ are fit by estimating some parameter θ and setting $\hat{Y}_n(x) = g(x, \hat{\theta}_n)$ and $\hat{Y}_{n-1}(x) = g(x, \hat{\theta}_{n-1})$, where $\hat{\theta}_n$ and $\hat{\theta}_{n-1}$ are estimated based on all n data points and only the first $n-1$ data points respectively, one sees $\text{EPR}(T_{n-1}) = \text{EPR}(T_n) + o(1/n)$ by plugging everything into (5.1). Hence, the LOOCV estimator of $\text{EPR}(T_n)$ misses its target by only a term of size $o(1/n)$ in expectation. Because of this, the LOOCV is approximately unbiased for what it estimates! There is not even really a need to estimate the term of size $O(1/n)$ in (5.1). The LOOCV does it for free. If one

5.1. Can precision be precisely estimated?

instead considers K-folded cross-validation (CV_K), completely analogue arguments show

$$\begin{aligned} E CV_K &= EPR(T_{n(K-1)/K}) \\ &= EPR(T_n) - \frac{K}{n} \left(\nabla EPR(\theta)^T c + \frac{1}{2} \text{Tr}[H_{\theta_{\text{lr}}} EPR(\theta) V] \right) + o(1/n). \end{aligned}$$

Because of this, CV_K needs to be slightly corrected in order to estimate $EPR(T_n)$ to the desired level of precision. It is therefore necessary to either use the LOOCV or a modified version of the K-folded CV if one wishes to capture the $O(1/n)$ effects of the EPR. This point is, however, sometimes overlooked in the literature, see e.g. Bouchard 2007. In my opinion, this is a huge oversight in works where one wishes to seriously discuss and capture the difference between generative and discriminative procedures.

Based on the arguments above, one would expect that using the LOOCV to choose between generative and discriminative counterparts outperforms using the CV_K . Surprisingly though, this is not always the case, something which can be seen by repeating the simple simulation used to create Figure 1 a number of times, using both LOOCV and CV_5 to estimate whether a generative or discriminative model is preferable in each iteration. By the results of Efron 1975 and O’Neill 1980, one would expect the generative model to be the best choice in this case. I did this simulation experiment $B = 1000$ times, and found that the LOOCV chose the generative model about 84% of the time. For the the 5-folded procedure, the asymptotically optimal model was found to be preferable in around 89% of the iterations. Moreover, in each iteration, I registered which model both LOOCV and CV_5 chose and the error rate of the preferred model as computed on an independent test set of $N = 10000$ data points. After averaging the results over the B iterations, I found that the procedure which selects either LDA or LR based on the LOOCV had a mean error rate of 0.046, while choosing model based on the CV_5 gave an average error rate of 0.045. From this simulation, it seems like the CV_5 actually works better than the LOOCV for capturing the $1/n$ -effects! This could of course be an artefact of the previously discussed fact that LDA and LR have extremely similar performances in this case, but I believe there is more to it than that, and that a main reason for this strange behaviour is the fact the neither the LOOCV nor CV_5 manage to precisely estimate the EPR.

The previous arguments show that LOOCV is only $o(1/n)$ away from the EPR in expected value. This is a good first step and indicate that the LOOCV “should” be a reasonable estimator for the EPR, as the bias is of a sufficiently small order. In Section 3.2.2, I used the same argument to justify why minimising the FIC is desirable when one wishes to get a small value of the MSE. The simulations described in the previous paragraph, however, indicate that the LOOCV does not perform as well as one might expect. I believe this is partially due to the fact the even though the LOOCV (and the FIC) misses its target by a term of order $o(1/n)$ in expectation, the actual estimators are much more imprecise. To show this, I will change the setting slightly and work with

$$CV = n^{-1} \sum_{i=1}^n g(Z_i, \hat{\theta}_{n-1,(-i)}),$$

where g is a smooth function and $\hat{\theta}_{n-1,(-i)}$ is some Z-estimator based on $Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n$. In **Paper III**, it is shown that

$$n^{-1} \sum_{i=1}^n g(Z_i, \hat{\theta}_{n-1,(-i)}) = n^{-1} \sum_{i=1}^n g(Z_i, \hat{\theta}_n) - (2n)^{-1} \text{Tr}[\Phi'(\theta_{\text{lf}})^{-1}C] + o_p(1/n),$$

where $\hat{\theta}_n$ is the Z-estimator based on all data, and C is the covariance between $\varphi(Z, \theta_{\text{lf}})$ and $\nabla_{\theta_{\text{lf}}}g(Z, \theta)$. Taylor expanding this equation around θ_{lf} , reveals

$$\text{CV} = n^{-1} \sum_{i=1}^n g(Z_i, \theta_{\text{lf}}) + n^{-1} V_n^T U_n + n^{-1} U_n^T W_n U_n - (2n)^{-1} \text{Tr}[\Phi'(\theta_{\text{lf}})^{-1}C] + o_p(1/n),$$

where $U_n = \sqrt{n}(\hat{\theta}_n - \theta_{\text{lf}})$, $V_n = n^{-1/2} \sum_{i=1}^n \nabla_{\theta_{\text{lf}}}g(Z_i, \theta)$ and $W_n = n^{-1} \sum_{i=1}^n \nabla_{\theta_{\text{lf}}}^2 g(Z_i, \theta)$. Assuming everything is sufficiently regular, U_n and V_n both converge in law towards central normal distributions, while W_n converges in probability towards a fixed number by the law of large numbers. Hence,

$$\text{CV} = n^{-1} \sum_{i=1}^n g(Z_i, \theta_{\text{lf}}) + O_p(1/n).$$

The first term on the right-hand side of this equality is a mean, and by the CLT one therefore has $\sqrt{n}[\text{CV} - \mathbb{E} g(Z, \theta_{\text{lf}})] \xrightarrow{d} N[0, \text{Var} g(Z, \theta_{\text{lf}})]$.² A similar argument can be made to show that also $\sqrt{n}[\text{FIC} - (\mu_1 - \mu_0)^2]$ converges in law to a central normal distribution, and similar results for the AIC and TIC.

At first glance, the observation that CV and the information criteria are $\sqrt{n}$ -consistent for natural quantities might look like a good thing. After all the result can be used to make inference about the estimators and create confidence intervals for the limiting risks. The limiting behaviour does, however, have another, quite dire implication. Since CV and the information criteria are $\sqrt{n}$ -consistent for...something, their error is of size $O_p(1/\sqrt{n})$, a term which is larger than $O(1/n)$, and since these extra precise estimators are used to properly capture effects of size $O(1/n)$, the CV and the information criteria do not achieve their goals outside of the expected value!³ In **Paper III**, this is made very clear by the fact that the terms of size $O(1/n)$ have no effect on the limiting distribution of $\hat{\theta}(\hat{\lambda})$ when λ is set to the minimiser of either of the risk estimators.

The above might lead one to conclude that the FIC, CV, AIC and TIC are simply “bad” estimators and that one should sit down and come up with better alternatives. Sadly, I doubt this is an easy task. More precisely, in all but a very few cases, I believe it to be impossible to estimate precisely enough to capture $O(1/n)$ -effects.

To see this, assume the true underlying distribution is equal to f_{θ_0} for some θ_0 and that one is trying to estimate a parameter τ with true value $\tau_0 = g(\theta_0)$ for some smooth function g . For simplicity, let τ and θ be one dimensional. Assume furthermore that some estimator $\hat{\tau}$ has the property $\mathbb{E}_{\theta} \hat{\tau} = g(\theta) + \delta_n(\theta)$, where

²This fact has been noted before, see e.g. Wager 2020. This paper also claims Yang 2007 implicitly shows the result.

³In the context of linear regression, this was pointed out for the CV by Shao 1993.

$\delta_n(\theta) = o(1/n)$, $\delta'_n(\theta_0) = o(1/n)$ and the subscript in the expectation indicate that integration is being done with respect to the measure induced by f_θ . By the Crámer-Rao inequality discussed in Section 3.1.3,

$$\text{Var } \hat{\tau} \geq \frac{[g'(\theta_0) + o(1/n)]^2}{nJ(\theta_0)},$$

which shows that the variance of $\hat{\tau}$ is at least $C/n + o(1/n)$ for some $C > 0$. In particular, this ensures that the variance of $n(\hat{\tau} - \tau_0)$ is bounded from below by $nC + o(n)$, which grows arbitrarily large as $n \rightarrow \infty$. The fact that the variance of $n(\hat{\tau} - \tau)$ diverges, proves that $n(\hat{\tau} - \tau)$ does not converge towards zero in the L^2 -norm.⁴ This type of convergence is, however, stronger than convergence in probability, and the above therefore does not prove $n(\hat{\tau} - \tau) \xrightarrow{P} 0$. For this to be implied, one needs the additional assumption that $n^2(\hat{\tau} - \tau)^2$ be uniformly integrable, i.e. that $\lim_{\alpha \rightarrow 0} \sup_n E I(|X_n| > \alpha) |X_n| = 0$ where $X_n = n^2(\hat{\tau} - \tau)^2$ which follows when $\sup_n E|X_n|^{1+\epsilon} < \infty$ for some $\epsilon > 0$ (Billingsley 1995, p. 338). If this property holds, standard theory ensures that $n(\hat{\tau} - \tau) \xrightarrow{P} 0$ implies $n^2 E(\hat{\tau} - \tau)^2 \rightarrow 0$, see e.g. Billingsley 1995, Th. 25.12. This is equivalent to, $\text{Var } \hat{\tau} = o(1/n^2)$, but $\text{Var } \hat{\tau} \geq C/n + o(1/n)$, by the Cramér-Rao inequality, leading to a contradiction and showing that $\hat{\tau}$ can only be $o(1/n)$ -consistent for τ when $n(\hat{\tau} - \tau)$ is relatively “irregular”. Because of this, I am unsure if approximating functions on the form of (5.1) with errors smaller than $O(1/n)$ is really doable, at least if one also wishes to estimate the first term in the expansion, and want the estimator to be relatively well-behaved.

5.2 Accounting for every little thing

In **Paper I**, a new version of the FIC is derived, and it is proposed to use this criterion to choose between different parametric models or estimators. In **Paper II**, the FIC is used to choose the balance between parametric and nonparametric in the hybrid likelihood framework. In both cases, a whole new step in the estimation procedure is created, and this step requires the estimation of multiple quantities. Because of this, a lot of variance could potentially be introduced, which is not accounted for by the results in **Paper I** and **II**. This makes the results “imprecise”. I will now address this issue, but as parts of the following discussion are covered by **Paper III**, I will allow myself to be slightly brief. I also remark that the issue of extra variability introduced by adding a model selection step into the estimation procedure is not unique to the FIC. The same problem arises if one uses the AIC, the TIC, CV or any other model selection tool.

First, I will introduce some notation. For each value of a parameter λ , I will assume there is a corresponding estimator $\hat{\theta}(\lambda)$. The parameter λ can be continuous like the balance parameter in the hybrid likelihood framework, or discrete which covers the problem of model selection in a finite class of models $M_1, \dots, M_K$ by letting $\hat{\theta}(k)$ be the estimated parameter in model k for $k = 1, \dots, K$. I will also assume there exists some procedure which estimates

⁴ X_n converges to X in L^2 -norm if $E(X_n - X)^2 \rightarrow 0$.

the optimal value of λ , $\hat{\lambda}$. One usually more or less knows how $\hat{\theta}(\lambda)$ behaves for each fixed value of λ . Typically, one can show results like $\hat{\theta}(\lambda) \xrightarrow{p} \theta(\lambda)$ and $\sqrt{n}[\hat{\theta}(\lambda) - \theta(\lambda)] \rightarrow N[0, V(\lambda)]$ for each λ . I will refer to this as point-wise results, and since these results only hold true for each fixed value of λ , they cannot be used to make inference about $\hat{\theta}(\hat{\lambda})$. In practice, however, this point is often ignored, and λ is treated as fixed at $\hat{\lambda}$, leading to the approximation $\hat{\theta}(\hat{\lambda}) \approx \theta(\hat{\lambda})$ and that $\hat{\theta}(\hat{\lambda})$ is approximately $N[0, V(\hat{\lambda})]$ -distributed. This conclusion is, however, not correct a-priori, as the added randomness of estimating λ is not taken into account.

In **Paper III**, the issue described in the previous paragraph is addressed when λ is continuous, i.e. when λ takes values in an open subset of a Euclidian space, and precise limit results are derived for $\hat{\theta}(\hat{\lambda})$ when λ is set to the minimiser of either of the risk estimators discussed in this thesis. This covers the case of models fitted with hybrid likelihood when the balance parameter is set to the minimiser of the FIC, which was indeed the motivation for **Paper III**, but since λ is required to be continuous, the extremely common situation of choosing between $K \in \mathbb{N}$ candidate models, is not covered. The reason why this situation is left out from the article is that the result is simply... uninteresting, as one often gets that estimating λ has no effect on the limiting distribution of $\hat{\theta}(\hat{\lambda})$.

To see that the effect of estimating $\hat{\lambda}$ vanishes when λ is discrete, assume that $R(\theta)$ is some function which quantifies the risk of using θ , see Section 3.1. Furthermore, let $\hat{R}(\lambda)$ be some estimator of the expected value of $R[\hat{\theta}(\lambda)]$ for each λ , this covers the case of the FIC, the AIC, the TIC and also the CV. Typically, one has $\hat{R}(\lambda) \xrightarrow{p} R[\theta(\lambda)]$ for each λ . If the limiting risk $R[\theta(\lambda)]$ takes unique values for all λ s, there is some λ_0 which is asymptotically optimal in the sense that it minimises $R[\theta(\lambda)]$. Furthermore, since $R[\theta(\lambda)]$ is assumed to be unique for all values of λ , the function $(x, y) \mapsto I(x < y)$ is continuous at every pair $(R[\theta(\lambda_1)], R[\theta(\lambda_2)])$. Hence, the continuous mapping theorem guarantees that

$$\hat{\theta}(\hat{\lambda}) = \sum_{k=1}^K \hat{\theta}(\lambda_k) \prod_{j \neq k} I[\hat{R}(\lambda_k) < \hat{R}(\lambda_j)] \xrightarrow{p} \sum_{k=1}^K \theta(\lambda_k) \prod_{j \neq k} I\{R[\theta(\lambda_k)] < R[\theta(\lambda_j)]\} = \theta(\lambda_0)$$

proving that $\hat{\theta}(\hat{\lambda})$ is consistent for $\theta(\lambda_0)$. Furthermore, if $\sqrt{n}[\hat{\theta}(\lambda) - \theta(\lambda)] \xrightarrow{d} N[0, V(\lambda)]$ for each $\lambda = 1, \dots, K$,

$$\sqrt{n}[\hat{\theta}(\hat{\lambda}) - \theta(\hat{\lambda})] = \sum_{k=1}^K \sqrt{n}[\hat{\theta}(\lambda_k) - \theta(\lambda_k)] \prod_{j \neq k} I[\hat{R}(\lambda_k) < \hat{R}(\lambda_j)] \xrightarrow{d} N[0, V(\lambda_0)],$$

by Slutsky's lemma and the continuous mapping theorem once again. This shows that when $R(\lambda)$ takes different values for every λ , the effect estimating λ has on the limiting distribution of $\hat{\theta}(\hat{\lambda})$ disappears as the sample size grows!

The above argument only holds when $R[\theta(\lambda)]$ takes different values for all λ s. When $\theta(\lambda)$ is constant with respect to λ , none of the arguments above can be applied. This is the case for instance when the models considered are nested and the smallest model is the correct one, or when choosing between discriminative and generative classifiers in correctly specified models, see Section 5.1. In these cases, however, the limiting risk differs only in a term of size $O(1/n)$, by (3.3).

Because of this, one needs to be able to estimate effects of this size in order to properly choose a value for λ . As discussed in the previous section, however, this comes with its own whole set of problems, and I will therefore not consider this situation in the present section. See, however, Appendix B in **Paper III** for more discussions.

Even though the above arguments show that the effect of estimating discrete λ disappears in the limit, the theory says nothing about what happens for finite sample sizes, and in practice, the model selection step can introduce quite a lot of variance for finite n . To illustrate this, I fitted linear regression models in one and two dimensions to some simulated data, and used the AIC to choose one of the two models.⁵ This procedure was repeated $B = 10\,000$ times for a selection of sample sizes. Figure 2 shows the empirical variance of $\sqrt{n}\hat{\beta}(\hat{\lambda})$ in the B samples as a function of n together with the theoretical limit. From the plot, one sees clearly that although the variance of $\sqrt{n}\hat{\beta}(\hat{\lambda})$ approaches what the theory suggests as $n \rightarrow \infty$, the convergence takes some time to “kick in”. Because of this, methods which account for the added randomness of the model selection step are often desirable to use, even though they are not necessary, asymptotically speaking.

The field which attempts to make inference about $\hat{\theta}(\hat{\lambda})$ in a way taking the added randomness of estimating λ into account is called post-selection inference and has been popular in recent years, see Kuchibhotla et al. 2022 for an excellent review and overview of methods. I do, however, have some reservations against these techniques. By the previous arguments, the effect of estimating λ is too small to affect the limiting distribution of $\hat{\theta}(\hat{\lambda})$, and using point-wise results is therefore really asymptotically unproblematic. Moreover, most of classic inference is based on asymptotic arguments, results which require infinite sample sizes no less than the arguments above. For this reason, it feels a bit arbitrary to me not to accept the above calculations and use, oftentimes quite complicated, techniques to account for an effect which is no less negligible than the error one makes by approximating the distribution of $\sqrt{n}[\hat{\theta}(\lambda) - \theta(\lambda)]$ by $N[0, V(\lambda)]$.

5.3 Smoothing over the hard parts

In all papers in this thesis, it has been required that everything is “sufficiently regular”. When this assumption is made, the conditions are usually relatively unproblematic and standard, typically requiring that means of certain orders exist or that a quantity is uniformly integrable. There is, however, one assumption which is required in all articles and which has pretty big consequences on the theory. This is the assumption of smoothness discussed in Section 2.3.1.

In my opinion, the results in **Paper II** and **Paper III** are severely weakened by the smoothness conditions imposed. In **Paper II**, one needs the estimating function used in the construction of the empirical and hybrid likelihood function to be smooth. This renders the theory inapplicable when $m(y, \mu) = I(y \leq \mu) - p$, which makes it impossible to use quantiles as control parameters in the hybrid

⁵I simulated independent variables $X \sim N(0, I_2)$ where I_2 is the 2×2 identity matrix and $\epsilon \sim N(0, 1)$, and set $Y = \beta_0 + \beta^T X + \epsilon$ where $\beta = (\beta_1, \beta_2)^T$ and $\beta_j = 0.1$ for $j = 0, 1, 2$.

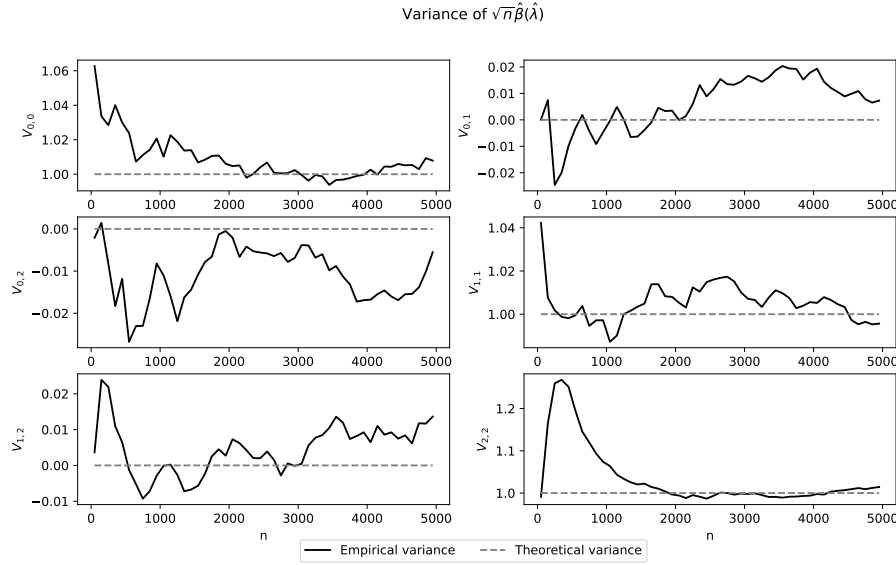


Figure 2: The figure shows the empirical variance of $\sqrt{n}\hat{\beta}(\hat{\lambda})$ based on $B = 10\,000$ samples, together with the theoretical limit as a function of the sample size n . In the figure, $V_{j,k}$ refers to the covariance between the j -th and k -th component of $\sqrt{n}\hat{\beta}(\hat{\lambda})$

likelihood function while using the theorems of **Paper II**. In **Paper III**, it is required that the tuning parameter λ is set to the minimiser of $\sum_{i=1}^n \psi[(Z_i, \hat{\theta}(\lambda))]$ where ψ is smooth in its second argument. This excludes the very common situation where λ is tuned with respect to the error rate, i.e. when $\psi(y, \theta) = I[y \neq \hat{Y}(\theta)]$. Furthermore, it is in this paper assumed that for each fixed value of λ , $\hat{\theta}(\lambda)$ is a Z-estimator, which also can be seen as imposing a smoothness condition as M-estimators which maximise non-smooth maps cannot be rewritten as Z-estimators, see Section 2.2. For these reasons, it is impossible to use the results of **Paper II** and **Paper III** to make precise inference in many situations.

The problem in both **Paper II** and **Paper III** can be reduced down to the fact that standard theory is not able to cover M-estimators on the form

$$\hat{\theta}_n = \operatorname{argmax}_{\theta} \left(n^{-1} \sum_{i=1}^n g(\theta) I(Z_i \leq \theta) + h(\theta) \right), \quad (5.3)$$

where g and h are functions. This is because the map $(y, \theta) \mapsto I(y \leq \theta)$ is discontinuous in θ for all y , which implies that the Lipschitz-condition of Theorem 2 cannot possibly hold. Because of this, the asymptotic behaviour of $\hat{\theta}_n$ defined as in (5.3), cannot be studied with standard theory. At first, I thought this problem could be avoided by replacing the indicator function with a smooth approximation to the map. This would allow all the theory to go through as if the Lipschitz-condition in Theorem 2 was satisfied, and with a close enough approximation, replacing I with a smooth approximation would make virtually no difference to the analysis. I quickly discovered, however, that this worked poorly in practice, something I believe stems from the fact that the lack of smoothness in I makes $\hat{\theta}_n$ of (5.3) behave very differently from classic M-estimators in smooth problems. To show this, I will now informally derive some results for $\hat{\theta}_n$. All of my arguments

are based on van der Vaart 2000, Ch. 5, in particular Example 5.54.

I will start by showing consistency of $\hat{\theta}_n$ towards

$$\theta_0 = \operatorname{argmax}_{\theta} [g(\theta)F(\theta) + h(\theta)],$$

where F is the cdf in the true underlying distribution of the data. By van der Vaart 2000, Th. 5.7, it suffices to show that $\Phi_n(\theta) = n^{-1} \sum_{i=1}^n g(\theta)I(Z_i \leq \theta) + h(\theta)$ converges uniformly towards $\Phi(\theta) = g(\theta)F(\theta) + h(\theta)$ in probability, and that Φ does not “flatten out” around θ_0 . The latter property is ensured by assuming $\Phi'(\theta_0) \neq 0$, and since

$$|\Phi_n(\theta) - \Phi(\theta)| = g(\theta)|\mathbb{F}_n(\theta) - F(\theta)|,$$

where $\mathbb{F}_n$ is the empirical distribution function based on $Z_1, \dots, Z_n$ and F is the cdf in the true underlying distribution, the first property follows more or less directly from the Glivenko-Cantelli theorem (see e.g. van der Vaart 2000, Th. 19.1).⁶ Hence, $\hat{\theta}_n \xrightarrow{p} \theta_0$ under relatively weak conditions.

The next goal is to show that $n^\alpha(\hat{\theta}_n - \theta_0)$ converges in law to some distribution. To do this, one first needs to establish that $n^\alpha(\hat{\theta}_n - \theta_0) = O_p(1)$ for some α . This can be achieved with relative ease by applying 5.52 in van der Vaart 2000. To use this result, however, one needs to be familiar with bracketing numbers and related quantities. I will now give a very informal introduction to this topic, but as the concepts I will discuss are quite abstract, a formal introduction is beyond the scope of this section. For more details, consult van der Vaart 2000, Ch. 19, which is also the main inspiration for my presentation.

As alluded to in Section 2.3.1, one needs to figure out “the size” of the class of functions $\mathcal{M} = \{(y, \theta) \mapsto g(\theta)I(y \leq \theta) + h(\theta) - g(\theta_0)I(y \leq \theta_0) - h(\theta_0) \mid \theta \in \Theta\}$ in order to find the rate of convergence for $\hat{\theta}_n - \theta_0$. One such way of measuring “size” is in term of bracketing numbers. The bracketing number of $\mathcal{M}$, $N_{[]}[\epsilon, \mathcal{M}, L_r(F)]$, is the minimum number of ϵ -brackets in $L_r(F)$ ⁷ needed to cover the class $\mathcal{M}$. A bracket is defined in terms of two functions: a lower bound l and an upper bound u . For l and u , the $L_r(F)$ -bracket $[l, u]$ is the set of all functions in $L_r(F)$ which are point-wisely bounded from below by l and from above by u , i.e. $[l, u] = \{f \mid f \in L_r(F) \text{ and } l(z) \leq f(z) \leq u(z)\}$. The size of $[l, u]$ is the $L_r(F)$ norm of $z \mapsto u(z) - l(z)$, i.e. $E[u(Z) - l(Z)]^{1/r}$. Hence, $N_{[]}[\epsilon, \mathcal{M}, L_r(F)]$ measures how “large” $\mathcal{M}$ is by counting how many brackets one needs to cover the set. One can think of this as a functional analogue of measuring the size of a set in $\mathbb{R}^p$ by counting how many balls with diameter ϵ one needs to cover it.

For a fixed ϵ , $N_{[]}[\epsilon, \mathcal{M}, L_r(F)]$ measures how large $\mathcal{M}$ is in terms of ϵ -brackets in $L_r(F)$. As $\epsilon \rightarrow 0$, $N_{[]}[\epsilon, \mathcal{M}, L_r(F)]$ will therefore usually grow large. To be able to properly evaluate the size of $\mathcal{M}$, one therefore needs to ensure that this growth is not too explosive. This is measured in terms of the bracketing integral defined

⁶For the argument to go through, one needs g to be bounded. This can be ensured by assuming g to be continuous and the parameter space compact.

⁷Here $L_r(F)$ refers to the vector space of functions with bounded $L_r(F)$ -norm, i.e. all F -measurable functions f with $\|f\|_{L_r(F)} = (E f(Z)^r)^{1/3} < \infty$.

in the following way

$$J_{[]}[\delta, \mathcal{M}, L_2(F)] = \int_0^\delta \sqrt{\log N_{[]}[\epsilon, \mathcal{M}, L_2(F)]} d\epsilon.$$

Notice that the above is only defined for brackets in $L_2(F)$. This is intentional, as this definition is really all that is needed.

With the definition of the bracketing integral in our arsenal, the order of convergence of $\hat{\theta}_n - \theta_0$ can be computed. By van der Vaart 2000, Th. 5.2, $n^{2(\alpha-\beta)}|\theta - \theta_0| = O_p(1)$ if the following two conditions are satisfied for sufficiently small δ and some $C > 0$:

$$\sup_{\delta/2 \geq |\theta - \theta_0| < \delta} \Phi(\theta) - \Phi(\theta_0) \leq -C\delta^\alpha \text{ and} \quad (5.4)$$

$$E\left(\sup_{|\theta - \theta_0| < \delta} \sqrt{n}|\Phi_n(\theta) - \Phi_n(\theta_0) - \Phi(\theta) + \Phi(\theta_0)|\right) \leq C\delta^\beta. \quad (5.5)$$

Since θ_0 maximises $\Phi(\theta)$ by definition, one usually has $\Phi'(\theta_0) = 0$, and hence a two term Taylor expansion around θ_0 ensures that the first property holds true with $\alpha = 2$ assuming Φ is twice continuously differentiable in a neighbourhood of θ_0 . For the second condition, it follows from van der Vaart 2000, Co. 19.35 that

$$E\left(\sup_{|\theta - \theta_0| < \delta} \sqrt{n}|\Phi_n(\theta) - \Phi_n(\theta_0) - \Phi(\theta) + \Phi(\theta_0)|\right) \leq J_{[]}[\|\text{env}\|_{F,2}, \mathcal{M}_\delta, L_2(F)], \quad (5.6)$$

where env is an envelop function for the class $\mathcal{M}_\delta = \{(z, \theta) \mapsto g(\theta)I(z \leq \theta) + h(\theta) - g(\theta_0)I(z \leq \theta_0) - h(\theta_0) \mid |\theta - \theta_0| < \delta\}$, i.e. $f(z) \leq \text{env}(z)$ for every $f \in \mathcal{M}_\delta$. For $\mathcal{M}_\delta$, the function which is constantly equal to

$$K_\delta = \sup_{\theta_0 - \delta \leq \theta \leq \theta_0 + \delta} \{|g(\theta)| + |h(\theta)|\}$$

is an envelop with $\|\text{env}\|_{F,2} = K_\delta$ which, by the extreme value theorem and mean value theorem, is bounded by $K_1 + \delta K_2$ for constants K_1, K_2 if g and h are continuously differentiable in $[\theta_0 - \delta, \theta_0 + \delta]$.

To be able to compute the right hand side of (5.6), the ϵ bracketing numbers of the class $\mathcal{M}_\delta$ for each δ need to be found. If the Lipschitz property in (2.4) had held true for $\psi(z, \theta) = g(\theta)I(z \leq \theta) + h(\theta)$, this would have been a relatively easy feat, as this property ensures that the functions in $\mathcal{M}_\delta$ are in some sense as close as the θ s in $[\theta_0 - \delta, \theta_0 + \delta]$ are, and hence, that the class is small, see van der Vaart 2000, Ex. 19.7. The Lipschitz property does, however, not hold true for ψ . Because of this, one needs to proceed differently when computing the bracketing numbers for this class of functions than in the case where the Lipschitz-property holds true.

Assume first that $g(\theta_0) \neq 0$. Then for sufficiently small δ one has either $g(\theta) < 0$ or $g(\theta) > 0$ for all $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$, provided g is continuous. I will without loss of generality assume $g(\theta_0) > 0$. The other case is similar. To construct suitable brackets, start by picking $\theta_1 < \theta_2 < \dots < \theta_N$ where $\theta_{j+1} - \theta_j \leq \delta_\theta$. For each j , let g_j^l and g_j^u denote the maximum and minimum value of $g(\theta)$ in $[\theta_j, \theta_{j+1}]$. By the extreme value theorem, these values are attained, so by the mean value theorem

and the extreme value theorem yet again $g_j^u - g_j^l \leq G\delta_\theta$ for some $G > 0$ provided g is continuously differentiable. Let h_k^l and h_j^u be defined similarly. Again one has, $h_j^u - h_j^l \leq H\delta_\theta$ for some H if h is continuously differentiable. With these definitions, define the following brackets $[l_1, u_1], \dots, [l_N, u_N]$ where

$$\begin{aligned} l_j(z) &= g_j^l I(z \leq \theta_j) + h_j^l - g(\theta_0)I(z \leq \theta_0) - h(\theta_0) \text{ and} \\ u_j(z) &= g_j^u I(z \leq \theta_{j+1}) + h_j^u - g(\theta_0)I(z \leq \theta_0) - h(\theta_0). \end{aligned}$$

By construction, these brackets cover $\mathcal{M}_\delta$. Furthermore, very extensive and tedious computations show

$$\mathbb{E}(u_j - l_j)^2 \leq (g_j^u)^2 M \delta_\theta + (G + H + 2g_j^u G M + 2g_j^k H M + 2GH) \delta_\theta^2,$$

if F is continuously differentiable with $|F'(z)| \leq M$ in $[\theta_0 - \delta, \theta_0 + \delta]$. Hence, to ensure that the brackets $[l_j, u_j]$ for $j = 1, \dots, N$ have size at least ϵ , one needs to choose the points $\theta_1, \dots, \theta_N$ in a way such that the right hand side is bounded by ϵ^2 . For small ϵ , this amounts to choosing the points so that $\delta_\theta \propto \epsilon^2$, which requires a number of points proportional to $2\delta/\epsilon^2$ since $[\theta_0 - \delta, \theta_0 + \delta]$ has diameter 2δ . Because of this, $N_{[]}[\epsilon, \mathcal{M}_\delta, L_2(F)]$ is bounded by a number proportional to δ/ϵ^2 for small ϵ . For larger ϵ , one gets $N_{[]}[\epsilon, \mathcal{M}_\delta, L_2(F)] \leq K\delta/\epsilon$ for some K .

Everything is now ready for the bracketing integral to be computed. For sufficiently small δ , $J_{[]}[\delta, \mathcal{M}, L_2(F)]$ is bounded from above by

$$\int_0^{K_1+K_2\delta} \sqrt{\log \delta - 2 \log \epsilon + C_1} d\epsilon + \int_0^{K_1+K_2\delta} \sqrt{\log \delta - \log \epsilon + C_2} d\epsilon,$$

where C_1 and C_2 are the logarithm of the proportionality constants in the relations derived in the previous paragraph. By performing change of variables, one can show that the above is bounded by $C'_1 \delta^{1/2} + C'_2 \delta$ for some constants $C'_1, C'_2 > 0$. Hence, (5.5) holds true with $\beta = 1/2$, which yields a rate of convergence of $n^{1/(2 \cdot 2 - 2 \cdot 1)} = n^{1/3}$, i.e. $\hat{\theta}_n - \theta_0 = O_p(n^{-1/3})$. This is a slower convergence than the classic $O_p(1/\sqrt{n})$ -behaviour M-estimators often shows.

In the above, I assumed $g(\theta_0) \neq 0$, so it is natural to ask what happens in the case of $g(\theta_0) = 0$. Well, in this case the brackets need to be constructed slightly differently as whether $z \mapsto g(\theta)I(z \leq \theta)$ is increasing or decreasing depends on the sign of $g(\theta)$ which can be both positive and negative. However, since $g(\theta_0) = 0$, the following functions can be used to construct brackets which cover $\mathcal{M}_\delta$

$$\begin{aligned} l_j(z) &= -G\delta + h_j^l - g(\theta_0)I(z \leq \theta_0) - h(\theta_0) \text{ and} \\ u_j(z) &= G\delta + h_j^u - g(\theta_0)I(z \leq \theta_0) - h(\theta_0). \end{aligned}$$

By similar computations as previously, one gets that $\mathbb{E}[l_j(Z) - u_j(Z)]^2$ is bounded by a term proportional to δ_θ^2 , which yields a bracketing number proportional to δ/ϵ and a bracketing integral, $J_{[]}[\delta, \mathcal{M}, L_2(F)]$ bounded by $C\delta$ for some C . This shows that when $g(\theta_0) = 0$, $\hat{\theta}_n - \theta_0 = O_p(1/\sqrt{n})$ which is a quicker rate of convergence than when $g(\theta_0) \neq 0$.

The above arguments indicate that the estimator $\hat{\theta}_n$ has the slightly weird behaviour of converging quicker when $g(\theta_0) = 0$ than when $g(\theta_0) \neq 0$. This

makes estimators on the form (5.3) pretty hard to use in practice, as one cannot know the rate of convergence of $\hat{\theta}_n$ without also knowing the true value of the parameter one is trying to estimate. This problem shows up in the context of hybrid likelihood estimation. It is possible to show, using arguments similar to those given in Section 2 of **Paper II**, that when one uses a quantile as the control parameter in the construction of the empirical likelihood function, the hybrid log-likelihood function is asymptotically equivalent to an expression on the form $\sum_{i=1}^n g(\theta)I(Z_i \leq \theta) + h(\theta)$. Furthermore, one also has that in this case $g(\theta_0) = 0$, when the parametric model used in construction of the hybrid likelihood function is correctly specified and $g(\theta_0) \neq 0$ otherwise. Hence, the maximiser of the hybrid likelihood function converges quicker if the model is correctly specified than in the case of misspecification! This makes it extremely hard, if not impossible, to create truly model agnostic results for this estimator.

Although the above shows that $\hat{\theta}_n - \theta_0 = O_p(n^{-1/3})$ when $g(\theta_0) \neq 0$, this does not strictly speaking prove that one does not also have $\hat{\theta}_n - \theta_0 = O_p(1/\sqrt{n})$. A sequence which converge to zero in probability is stochastically bounded, so any $O_p(1/\sqrt{n})$ -consistent estimator is also $O_p(n^{-1/3})$ -consistent. In order to show that $\hat{\theta}_n - \theta_0 \neq O_p(1/\sqrt{n})$, one first needs to also prove that $n^{1/3}(\hat{\theta}_n - \theta_0)$ converges to a non-trivial limit. This does, however, require quite a lot of additional efforts, which is slightly demoralising when the result one is trying to prove is a negative one. The belief that the order of convergence is different when $g(\theta_0) = 0$ and $g(\theta_0) \neq 0$ can, however, be strengthened by performing a small simulation experiment. For a selection of sample sizes $n = 100, 200, \dots, 5000$, I computed the maximiser of the hybrid likelihood function when the median is used as the control parameter, both in a setting where the parametric model used in construction of the hybrid likelihood function was correctly and incorrectly specified. This procedure was repeated $B = 1000$ times, and the mean value of $\sqrt{n}|\hat{\theta}_n - \theta_0|$ and $n^{1/3}|\hat{\theta}_n - \theta_0|$ for each n are displayed in Figure 3. From this plot one sees quite clearly that when the model is specified correctly, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ looks bounded, while the same quantity seem to diverge in the case of incorrect model-specification. This is a good indication that the speed of the convergence of $\hat{\theta}_n$ does indeed differ within and outside of model conditions.

Although not as explicit as for the maximiser of the hybrid likelihood function, this slower speed of convergence is also what makes it hard to generalise the results in **Paper III** to allow for tuning with respect to non-smooth error estimators. In this paper, the limiting behaviour of $\hat{\theta}(\hat{\lambda})$ is studied by reformulating the fitting procedure as a two-stage estimation procedure, and standard two-stage arguments do not work in the case of mixed convergence rates. To work with non-smooth risk estimators, one therefore needs more advance theory than what is used in **Paper III**.

The above is not the only reason why extending the results of **Paper III** to also cover tuning with respect to non-smooth error estimators becomes complex. One of the main results of the paper is that the LOOCV estimator of the risk, $CV = n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}_{-i})$ is not too far away from the training error, $TE = \sum_{i=1}^n \psi(Z_i, \hat{\theta})$. In the paper, this was shown by Taylor expanding ψ around $\hat{\theta}$, an argument which requires some degree of smoothness from ψ . If one were to

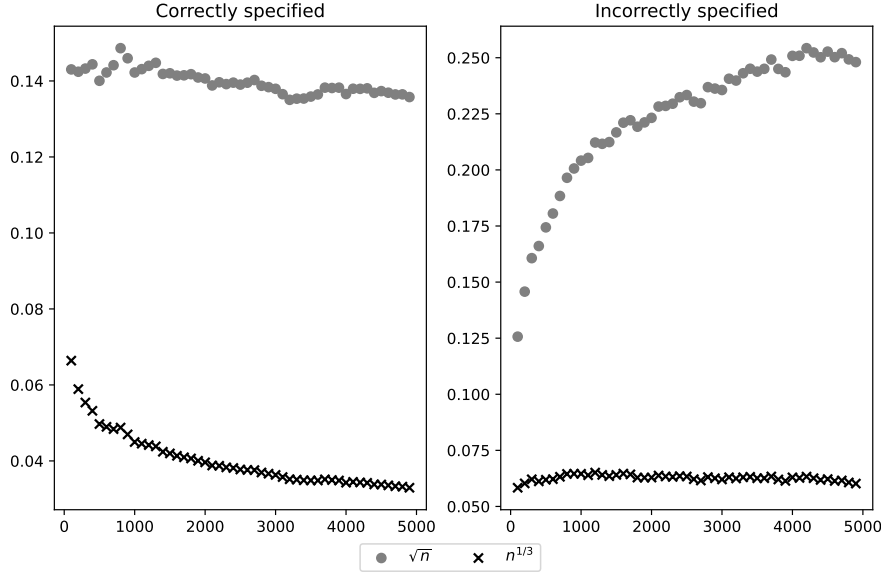


Figure 3: The figure shows the mean value of $n^\alpha |\hat{\theta}_n - \theta_0|$ for $\alpha = 1/3, 1/2$ and different values of n over $B = 1000$ iterations in a simulation experiment. Here $\hat{\theta}_n$ is the maximiser of the hybrid likelihood function and θ_0 the theoretical limit of the estimator.

try and extend the results of **Paper III** to allow for non-smooth functions ψ , one therefore needs some alternative way of computing how far off CV is from TE. For the case of the error rate, i.e. $\psi(y, x, \theta) = I[y \neq \hat{y}_\theta(x)]$, an option could be to use Stute 1984. A special case of Theorem 0.1 in this paper ensures that for any $\alpha < 1$ the following quantity converges to 1 almost surely

$$\sup_{c_1 \leq s \leq c_2} \frac{|\mathbb{F}_n(\theta_0) - F(\theta_0) - \mathbb{F}_n(\theta_0 + sn^{-\alpha}) + F(\theta_0 + sn^{-\alpha})|}{n^{1/2-\alpha/2} \sqrt{2s\alpha \log n}}$$

for any fixed $-\infty < c_1 < c_2 < \infty$. In particular, this in combination with the definition of the derivative shows

$$\begin{aligned} \sup_{c_1 \leq s \leq c_2} [\mathbb{F}_n(\theta_0) - \mathbb{F}_n(\theta_0 + sn^{-\alpha})] &= O_p\left(n^{1/2-\alpha/2} \sqrt{\log n}\right) + F(\theta_0 + sn^{-\alpha}) - F(\theta_0) \\ &= O_p\left(n^{1/2-\alpha/2} \sqrt{\log n}\right) + f(\theta_0)sn^\alpha, \end{aligned}$$

where f is the pmf or pdf in the distribution F , assuming it exists. And hence,

$$n^{-\beta} \mathbb{F}_n(\theta_0 + sn^{-\alpha}) = n^{-\beta} \mathbb{F}_n(\theta_0) + f(\theta_0)sn^{\alpha-\beta} + O_p\left(n^{1/2-\alpha/2-\beta} \sqrt{\log n}\right), \quad (5.7)$$

which can be seen as a type of Taylor expansion of $\mathbb{F}_n$ around θ_0 . In particular, with $\alpha = \beta = 1/2$,

$$n^{-1/2} \mathbb{F}_n(\theta_0 + sn^{-1/2}) = n^{-1/2} \mathbb{F}_n(\theta_0) + f(\theta_0)s + O_p\left(n^{-1/4} \sqrt{\log n}\right).$$

Although the above expansion is beautiful and provides a whole lot of intuition about how $\mathbb{F}_n$ behaves, it is not precise enough to sharply characterise the distance

from CV to TE. The reason for this is that, after dividing by $n^{1-\beta}$, (5.7) says

$$n^{-1}\mathbb{F}_n(\theta_0 + sn^{-\alpha}) = n^{-1}\mathbb{F}_n(\theta_0) + f(\theta_0)sn^{\alpha-1} + O_p\left(n^{-1/2-\alpha/2}\sqrt{\log n}\right).$$

For the two-stage arguments to be applicable, one needs the remainder term in the above to be of order $o_p(1/n)$, which would require $\alpha > 1$, but by Stute 1984, the result is only applicable when $\alpha < 1$. For this reason, I believe the sharpest result one can obtain using Stute 1984 is an expansion of the type $CV = TE + Cn^{-1+\epsilon} + o_p(n^{-1+\delta})$, where $\epsilon, \delta > 0$, which is not precise enough for the setup used in **Paper III**.

5.4 Concluding remarks

The main goal of this thesis can be found in its subtitle “precise and robust estimation and model selection”. Yet the conclusions of the previous sections do in some sense give the impression that this goal has not been achieved, and that “precision” is not really something which one can or should strive for. To ensure a more nuanced discussion, I will give some concluding remarks.

Although the conclusion of Section 5.1 is that estimation of $O(1/n)$ effects is impossible in all but a few cases, no amount of theoretical calculations can ruin the fact that the concepts critiqued in this section work very well in practice. The AIC and CV are obvious examples thereof, but also the FIC seems to work well in practice, as is indicated by the simulation studies in both **Paper I** and **II**. Because of this, concluding that estimating $O(1/n)$ effects is something which simply should not be done, is way too negative a conclusion. The calculations of Section 5.1 indicate that the $O(1/n)$ effects are not properly being taken into account. That being said, there is nothing in the computation proving that estimating these terms somehow makes the risk estimator worse than it otherwise would be. In practice, the effect of the estimated terms of size $O(1/n)$ often serves as a way to punish very complex methods, which can be beneficial in many situations. So, although I still believe that neither the AIC, the CV nor the FIC achieves the goal of properly estimating the $O(1/n)$ effects in the risk, I cannot deny the fact that they are useful and practical tools which perform well in applications and ensure that estimation procedures become extra precise.

In Section 5.3, I showed that when the indicator function is involved, estimators often exhibit more complicated behaviour than what one usually finds with smooth M-estimators. Again, the conclusion drawn is that creating proper model-agnostic results is not really doable, and that real precision is therefore not possible in this case. Furthermore, since the calculations of this section show that involving the indicator function decreases the rate of convergence, one might be lead to believe that one should avoid this map altogether. This is problematic if one wishes to make precise inference, as the indicator function is crucial in the definition of many famously robust estimators, e.g. the mode or the median. As before, I believe this to be too negative a conclusion to draw. If robustness is really important in an application, a slower rate of convergence may very well be “a price worth paying”.

It is just important to know what one is giving up by involving the indicator function and to keep the bias-variance trade-off in mind.

All in all, I will say that unlike its meaning, the word “precise” is a somewhat ambiguous term. The answer to whether or not this thesis achieves its goal of performing estimation and model selection in a “precise” manner, is therefore perhaps neither yes nor no. I do, however, believe that addressing the issue of precise estimation is done as well as can be, and that although the results derived in the thesis are not perfect, they might be, within their own small sub-fields, the best of the bad.

Bibliography

- Aitchison, J., & Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, 29(3), 813–828.
- Akaike, H. (1973). Information theory and an extension of the likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the second international symposium of information theory*. Akademiai Kiado.
- Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1), 125–127.
- Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67, 1341–1383.
- Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69, 683–734.
- Andrews, D. W. K. (2002). Generalized method of moments estimation when a parameter is on a boundary. *Journal of Business & Economic Statistics*, 20(4), 530–544.
- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–70.
- Avella Medina, M., & Ronchetti, E. M. (2015). Robust statistics: A selective overview and new directions. *WIREs Computational Statistics*, 7(6), 372–393.
- Avella-Medina, M. (2017). Influence functions for penalized m-estimators. *Bernoulli*, 23, 3178–3196.
- Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546), 1434–1445.
- Bayle, P., Bayle, A., Janson, L., & Mackey, L. (2020). Cross-validation confidence intervals for test error. *Conference on Neural Information Processing Systems*.
- Bera, A. K., & Biliyas, Y. (2002). The mm, me, ml, el, ef and gmm approaches to estimation: A synthesis. *Journal of Econometrics*, 107(1-2), 51–86.
- Bera, A. K., Biliyas, Y., & Simlai, P. (2006). Estimating functions and equations: An essay on historical developments with applications to econometrics. In T. C. Mills & K. Patterson (Eds.), *Palgrave handbook of econometrics : Vol. 1 : Econometric theory* (pp. 427–476). Palgrave Macmillan.
- Billingsley, P. (1995). *Probability and measure*. John Wiley & Sons Inc.

Bibliography

- Bouchard, G. (2007). Bias-variance tradeoff in hybrid generative-discriminative models. *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, 124–129.
- Box, G. E. P. (1979). *Robustness in the strategy of scientific model building* (tech. rep.). University of Wisconsin-Madison Mathematics Research Center.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth StatisticsProbability Series. Wadsworth Advanced Books; Software.
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.
- Chant, D. (1974). On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika*, 61, 291–298.
- Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25, 573–578.
- Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16, 31–41.
- Claeskens, G., Croux, C., & van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, 62, 972–979.
- Claeskens, G., & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464), 900–916.
- Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.
- Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 30(2), 248–275.
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.
- Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4), 377–403.
- Delsol, L., & Van Keilegom, I. (2020). Semiparametric m-estimation with non-smooth criterion functions. *Annals of the Institute of Statistical Mathematics*, 72, 577–605.
- Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 22(1), 139–153.
- Edgeworth, F. Y. (1908). On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, 71(2), 381–397.
- Edgeworth, F. Y. (1909). Addendum on "probable errors of frequency-constants". *Journal of the Royal Statistical Society*, 72(1), 81–90.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352), 892–898.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.
- Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438), 548–560.

- Elsener, A., & van de Geer, S. (2018). Robust low-rank matrix estimation. *The Annals of Statistics*, 46, 3481–3509.
- Fan, Y., & Shi, X. (2023). Wald, qlr, and score tests when parameters are subject to linear inequality constraints. *Journal of Econometrics*, 235, 2005–2026.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38.
- Fisher, A., & Kennedy, E. H. (2020). Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2), 162–172.
- Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of mathematics*, 41, 155–156.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222(594-604), 309–368.
- Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350), 320–328.
- Genest, C., & Rémillard, B. (2006). Discussion of "copulas: Tales and facts", by thomas mikosch. *Extremes*, 9(1), 27–36.
- Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of statistics*, 22, 1993–2010.
- Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4), 1208–1211.
- Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 9(4), 861–869.
- Hájek, A. (2024). Pascal's wager (E. N. Zalta & U. Nodelman, Eds.) [(Summer 2024 edition)]. *The Stanford Encyclopedia of Philosophy*. Retrieved August 28, 2024, from <https://plato.stanford.edu/archives/sum2024/entries/pascal-wager/>
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346), 383–393.
- Hampel, F. R., Ronchetti, E. M., Rousseuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. John Wiley & Sons.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 1029–1054.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.).
- Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3), 292–304.
- Hjort, N. L., & Claeskens, G. (2003a). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464), 879–899.
- Hjort, N. L., & Claeskens, G. (2003b). Rejoinder to the discussion of "frequentist model average estimators" and "the focused information criterion". *Journal of the American Statistical Association*, 98, 98:938–945.
- Hjort, N. L., McKeague, I. W., & Van Keilegom, I. (2018). Hybrid combinations of parametric and empirical likelihoods. *Statistica Sinica*, 28(4), 2389–2407.

Bibliography

- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73–101.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*.
- Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Wiley.
- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94, 401–419.
- Jullum, M., & Hjort, N. L. (2017). Parametric or nonparametric: The fic approach. *Statistica Sinica*, 27(3), 951–981.
- Ketz, P. (2018). Subvector inference when the true parameter vector may be near or at the boundary. *Journal of Econometrics*, 207, 285–306.
- Kim, J., & Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18, 191–219.
- Kim, K. I. (2016). Higher order bias correcting moment equation for m-estimation and its higher order efficiency. *Econometrics*, 4(4), 48.
- Ko, V., & Hjort, N. L. (2019). Model robust inference with two-stage maximum likelihood estimation for copulas. *Journal of Multivariate Analysis*, 171, 362–381.
- Kuchibhotla, A. K., Kolassa, J. E., & Kuffner, T. A. (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, 9(1), 505–527.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 24(2), 79–86.
- LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41, 802–828.
- Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1), 211–225.
- Lindstrøm, T., & Hveberg, K. (2015). *Flervariabel analyse med lineær algebra* (2nd ed.). Gyldendal akademisk.
- Lindstrøm, T. L. (2017). *Spaces: An introduction to real analysis* (Vol. 29). American Mathematical Soc.
- Mikosch, T. (2006). Copulas: Tales and facts. *Extremes*, 9, 3–20.
- Mohammadi, L., & van de Geer, S. (2005). Asymptotics in empirical risk minimization. *Journal of Machine Learning Research*, 6, 2027–2047.
- Moran, P. A. P. (1971). Maximum-likelihood estimation in non-standard conditions. *Mathematical Proceedings of the Cambridge Philosophical Society*, 70, 441–450.
- Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science.
- Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14, 201–206.
- O'Neill, T. J. (1980). The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. *Journal of the American Statistical Association*, 75(369), 154–160.
- Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), 237–249.

- Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1), 90–120.
- Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 1725–1747.
- Pagan, A. (1986). Two stage and related estimators and their applications. *The Review of Economic Studies*, 53(4), 517–538.
- Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185, 71–110.
- Pearson, K. (1936). Method of moments and method of maximum likelihood. *Biometrika*, 28(1/2), 34–59.
- Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387), 575–583.
- Prasad, A., Suggala, A. S., Balakrishnan, S., & Ravikumar, P. (2020). Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 601–627.
- Radchenko, P. (2008). Mixed-rates asymptotics. *The Annals of Statistics*, 36, 287–309.
- Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37, 81–91.
- Ronchetti, E. M. (2021). The main contributions of robust statistics to statistical science and a new challenge. *Metron*, 79, 127–135.
- Rosset, S., & Tibshirani, R. (2020). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 115, 138–151.
- Schweder, T., & Hjort, N. L. (2016). *Confidence, likelihood, probability*. Cambridge University Press.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 486–494.
- Shao, J. (2003). *Mathematical statistics*. Springer Science & Business Media.
- Shih, J. H., & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4), 1384–1399.
- Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP*, 8(3), 229–231.
- Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6), 1135–1151.
- Stigler, S. M. (1973). Studies in the history of probability and statistics. xxxii: Laplace, fisher, and the discovery of the concept of sufficiency. *Biometrika*, 60, 439–445.
- Stigler, S. M. (1990). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36(2), 111–147.

Bibliography

- Stute, W. (1984). The oscillation behavior of empirical processes: The multivariate case. *The Annals of Probability*, 12(2), 361–379.
- Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting [(in Japanese)]. *Suri-Kagaku*, 153, 12–18.
- Takeuchi, K. (2020). *Contributions on theory of mathematical statistics*. Springer.
- van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.
- van der Vaart, A. W., & Wellner, J. (1996). *Weak convergence and empirical processes: With applications in statistics*. Springer.
- von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3), 309–348.
- Wager, S. (2020). Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani. *Journal of the American Statistical Association*, 115(529), 157–160.
- Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4), 299–326.
- Wald, A. (1950). *Statistical decision functions*. Wiley.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.
- Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, (6), 2450–2473.
- Zhang, X., & Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39(1), 174–200.

Paper I

Accurate bias estimation with applications to focused model selection

Accurate bias estimation with applications to focused model selection

Ingrid Dæhlen  | Nils Lid Hjort | Ingrid Hobæk Haff

Department of mathematics, University of Oslo, Oslo, Norway

Correspondence

Ingrid Dæhlen, Department of mathematics, University of Oslo, Oslo, Norway.

Email: ingrdae@math.uio.no

Funding information

Norges Forskningsråd, Grant/Award Number: 237718

Abstract

We derive approximations to the bias and squared bias with errors of order $o(1/n)$ where n is the sample size. Our results hold for a large class of estimators, including quantiles, transformations of unbiased estimators, maximum likelihood estimators in (possibly) incorrectly specified models, and functions thereof. Furthermore, we use the approximations to derive estimators of the mean squared error (MSE) which are correct to order $o(1/n)$. Since the variance of many estimators is of order $O(1/n)$, this level of precision is needed for the MSE estimator to properly take the variance into account. We also formulate a new focused information criterion (FIC) for model selection based on the estimators of the squared bias. Lastly, we illustrate the methods on data containing the number of battle deaths in all major inter-state wars between 1823 and the present day. The application illustrates the potentially large impact of using a less-accurate estimator of the squared bias.

KEYWORDS

asymptotic theory, bias estimation, focused information criterion, misspecified models, model selection, MSE estimation

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Scandinavian Journal of Statistics* published by John Wiley & Sons Ltd on behalf of The Board of the Foundation of the Scandinavian Journal of Statistics.

1 | INTRODUCTION

Bias estimation has a long tradition and an extensive list of techniques. Cox and Snell (1968) give expressions of the bias of maximum likelihood estimators under the assumption of a correctly specified model. The jackknife, a way of estimating bias based on resampling, was introduced in Quenouille (1949), with important extensions in Quenouille (1956) and Tukey (1958). Firth (1993) takes a different approach and proposes techniques for reducing the bias, and hence removing the need for its estimation. In the present article, we build on classic theory by deriving approximations to the bias of a large class of estimators. Our goal will be to give comprehensive formulas throughout the article, and special attention should be given to Section 2.3. Here we derive model-robust expressions for the bias of maximum likelihood estimators. We believe the main theorem of this section serves as a modern extension of the results in Cox and Snell (1968) and is a novel addition to the field of bias estimation and correction.

Our main motivation for deriving precise bias approximations is estimation of mean squared error, or MSE for short. This risk function measures the expected squared L^2 -distance from an estimator to the value it is aiming for. In symbols, this means $\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta_0)^2$, where $\hat{\theta}$ is an estimator of a parameter θ_0 . Famously, the MSE decomposes into the sum of the squared bias and the variance of the estimator evaluated. For many popular and well used estimators, for example, means, medians, and maximum likelihood estimators, the variance component is relatively easy to estimate. The bias component of the MSE, on the other hand, is often more complicated as the true value θ_0 shows up in the expression. Furthermore, the variance of many estimators, for example, quantiles, maximum likelihood estimators, and means, is of order $O(1/n)$, where n is the sample size. Because, of this, estimators of the squared bias must have expected errors of order $o(1/n)$ for the variability to be given sufficient weight in the estimated MSE. To achieve this, the precise bias approximations is crucial.

Recently, estimation of bias and MSE has been developed in great detail within the field of model selection. This is mainly due to the information criterion introduced in Claeskens and Hjort (2003) and later extended by multiple authors. See for instance Zhang and Liang (2011), Claeskens et al. (2006), or Claeskens and Hjort (2008). The criterion is called the focused information criterion, or FIC for short, and ranks models by the quality of their estimate of a prespecified parameter of interest. Quality is measured by MSE, and as a consequence, the formulation of the FIC requires sophisticated estimators of the MSE and squared bias. Because of this, many advanced techniques for estimation of the squared bias have been developed in the literature concerning the FIC. The original versions of the criterion were created for situations where models are nested within a true parametric alternative, with the degree of misspecification decreasing by the order of $O(1/\sqrt{n})$. In Jullum and Hjort (2017), however, a new version of the criterion is given, which gives a general way of estimating the MSE of estimators.

The approximation given in Jullum and Hjort (2017) avoids the assumption of asymptotically correctly specified models. Because of this, the estimators in this article have found use in multiple situations, see for example, Jullum and Hjort (2019), Ko et al. (2019), Claeskens et al. (2019), and Cunen, Walløe, and Hjort (2020). The approximations can, however, be shown to have an error term of order $O(1/n)$ in cases where some or all of the quantities involved are asymptotically biased. This makes the criterion too imprecise to give proper weight to the variance component of the MSE. As a result the variability of estimators can potentially be greatly downplayed in the estimator of the MSE developed in Jullum and Hjort (2017). In the present article, we will build on

the estimators given in Jullum and Hjort (2017) and derive an approximation to the MSE which has expected error order $o(1/n)$.

We will start by considering precise bias estimation for a wide class of estimators. Afterwards, we will discuss MSE estimation and the FIC. In particular, we will consider when and why the expressions given in Jullum and Hjort (2017) need to be corrected. In Section 4.1, we will give a new version of the FIC, extending that of Jullum and Hjort (2017). Lastly, we will illustrate the techniques developed in this article by estimating the MSE of a selection of estimators of the difference in the median number of battle deaths in major inter-state wars before and after the Korean war.

2 | PRECISE BIAS ESTIMATION

Our main topic is precise estimation of the bias. We will work with estimators on the form

$$\hat{\mu} = \mu + \frac{1}{n} \sum_{i=1}^n v(Y_i) + \epsilon_n. \quad (1)$$

where $Y_1, \dots, Y_n$ are i.i.d. from a distribution F and $E v(Y) = 0$ when $Y \sim F$. At first glance, considering estimators on the form of (1) only might look slightly restrictive. In practice, however, this condition holds for many commonly used estimators. If, for instance, $\hat{\mu}_1$ is a mean, the above equation holds trivially with $v_1(y) = y - \mu_1$. For the p quantile μ_p , $v_1(y) = [p - I(y \leq \mu_p)]/f(\mu_p)$ can be used when f , the density in the distribution F , exists. When $\hat{\mu}_1$ is a maximum likelihood estimator of μ in some parametric model f_θ , (1) holds with $v_1(y) = \nabla g(\theta_{\text{lf}})^T J(\theta_{\text{lf}})^{-1} u(y, \theta_{\text{lf}})$ where θ_{lf} is the minimizer of the Kullback–Leibler divergence from the true distribution to the parametric family, u is the score function, $J(\theta_{\text{lf}})$ is the Fisher matrix in the model evaluated at θ_{lf} and g is a function mapping θ to the value of μ in the parametric model, see Section 2.3 for details. In addition, expressions for v_1 when $\hat{\mu}_1$ is a smooth function of the above estimators, can be derived by the delta method. For conditions ensuring that (1) is satisfied for the influence function, see for example, Thm. 5.5 in Shao (2003).

The main goal of this section is to find expressions for and estimators of

$$c = \lim_{n \rightarrow \infty} n E(\hat{\mu} - \mu) = \lim_{n \rightarrow \infty} n E \epsilon_n.$$

Before we can begin with estimation of the quantity, however, we need to discuss when $n E \epsilon_n$ converges and c is well-defined. We start with an informal argument which hopefully will provide some intuition about what c is really is and when it exists.

Let F_n be the empirical distribution function of $Y_1, \dots, Y_n \in \mathbb{R}$ and F the true cumulative distribution function. Assume that $\hat{\mu}$ can be represented as a functional of F_n , that is, that there exists a function T such that $\hat{\mu} = T(F_n)$. If T is sufficiently smooth, a Taylor expansion around F reveals

$$T(F_n) - T(F) = D_{F-F_n}^{(1)} T(F) + (1/2) D_{F-F_n}^{(2)} T(F) + O(\|F_n - F\|_\infty^3), \quad (2)$$

where $D_{F-F_n}^{(k)} T$ is the k th order directional derivative of T in direction $F_n - F$. The expression above is called the von Mises expansion, see for example, chap. 20 of Van der Vaart (1998) for more details.

The Donsker theorem, see for example Van der Vaart (1998, p. 266), ensures $\|F_n - F\|_\infty = O_{pr}(1/\sqrt{n})$. Hence, (2) implies

$$T(F_n) - T(F) = D_{F-F_n}^{(1)} T(F) + (1/2)D_{F-F_n}^{(2)} T(F) + o_{pr}(1/n).$$

The first term on the right-hand side is a sum of influence functions of T and has expectation zero. Under sufficient regularity, we therefore have

$$E\{n[T(F_n) - T(F)]\} = (n/2)ED^2T_{F-F_n}(F) + o(1). \quad (3)$$

Since $D^2T_{F-F_n}(F)$ is the quadratic term in a Taylor expansion, we would expect it to be of order $O(\|F_n - F\|^2) = O_{pr}(1/n)$, ensuring that the expression in (3) converges and c exists. A formal statement, proof and precise set of conditions showing that this intuition is indeed correct is given in Thm. 3.1 of Shao (1991).

In principle, we can always estimate c by approximating the right-hand side of (3). This method is quite general and can, in theory, be used for most estimators. In practice, however, this procedure is far from straightforward and, when done in full generality, results in incomprehensible formulas requiring fine-tuning and extensive calculations in each new application. We will therefore instead focus on some situations where precise formulas and/or estimation strategies for c can be found.

We will present four ways to estimate c . First, we will work with functions of unbiased estimators and quantiles. In Section 2.3 we will work with higher-order Taylor expansions of the log-likelihood function to give expressions for the bias of maximum likelihood estimators. Afterwards, we will discuss how c can be estimated when $\hat{\mu}$ is a function of estimators for which approximations to the bias exists. Lastly, we will discuss resampling techniques.

Remark 1. In the following, we will derive multiple approximations to c . Most of the formulas will be obtained by finding estimators with sufficiently small remainder terms, δ_n . Typically these will be of size $o_{pr}(1/n)$, and we will take this to imply that their expected values are of order $o(1/n)$. For this to hold true, a sufficient condition is that $\{n\delta_n\}_n$ is uniformly integrable. See, for example, Billingsley (1999, p. 31) for a proof and details concerning uniform integrability or Shao and Wu (1989) for an alternative set of regularity conditions. To make arguments more transparent, we will not focus on these technical details in the proofs, but take $E o_{pr}(1/n) = o(1/n)$ for granted. All required regularity conditions will, however, be stated in the theorems and full proofs dealing with all technicalities may be found in the Appendix B.

2.1 | Functions of unbiased estimators

The first situation we will consider is when $\hat{\mu}$ is a function of unbiased estimators. The following result gives a formula for c in this case.

Theorem 1. *Let $\hat{a} \in \mathbb{R}^p$ be an unbiased estimator of $a \in \mathbb{R}^p$ such that $\sqrt{n}(\hat{a} - a)$ converges in distribution to a $N(0, \Sigma)$ -distributed variable and all components of $\sqrt{n}(\hat{a} - a)$ to the power of three are uniformly integrable. Assume $h : \mathbb{R}^p \rightarrow \mathbb{R}$ is a function for which all partial derivatives up to and including order three exist and are continuous in*

a neighbourhood of a . Then, if $\mu = h(a)$ and $\hat{\mu} = h(\hat{a})$,

$$c + o(1) = nE(\hat{\mu} - \mu) = (1/2)\text{Tr}[Hh(a)\Sigma] + o(1), \quad (4)$$

where $Hh(a)$ is the Hessian matrix of h evaluated at the point a . Furthermore, c can be estimated consistently by

$$\hat{c} = (1/2)\text{Tr}(Hh(\hat{a})\hat{\Sigma}), \quad (5)$$

where $\hat{\Sigma}$ is some consistent estimator of Σ .

Proof. Taylor expanding h around a_0 shows

$$\hat{\mu} - \mu = (\hat{a} - a)^T \nabla h(a) + (1/2)(\hat{a} - a)^T Hh(a)(\hat{a} - a) + \epsilon_n. \quad (6)$$

for some remainder term ϵ_n . Typically, ϵ_n will be of order $O_{pr}(\|\hat{a} - a\|^3)$ which ensures $E\epsilon_n = O(n^{-3/2})$ under uniform integrability. Under the conditions of the theorem, this is indeed the case. Consult the Appendix B for details. Hence,

$$E(\hat{\mu} - \mu) = 0 + (1/2)E[(\hat{a} - a)^T Hh(a)(\hat{a} - a)] + o(1/n).$$

As is shown in the Appendix B, the conditions in the theorem ensures that $\sqrt{n}(\hat{a} - a) \xrightarrow{d} N(0, \Sigma)$ implies $n\text{Var}\hat{a} \rightarrow \Sigma$. Hence, by properties of the trace operator, (4) holds true.

Lastly, since $\hat{a}$ and $\hat{\Sigma}$ are consistent estimators of Σ and a , respectively, $\hat{c}$ defined in (5) converges in probability to c by the continuous mapping theorem. This concludes the proof. ■

The simplest examples of functions of unbiased estimators are unbiased estimators themselves. For such estimators h is linear and its Hessian zero. Hence, $c = 0$ and needs not be estimated. Example include linear function of means and ordinary least squares regression coefficients.

Another important example is maximum likelihood estimators in exponential families. To see this, let f be a member of the exponential family. Then f is on the form

$$f(x) = A(x)B(\theta) \exp[\omega(\theta)^T T(x)], \quad (7)$$

for some functions A, B, ω , and T and a parameter θ . The maximum likelihood estimator based on a sample $Y_1, \dots, Y_n$, is the maximizer of the log-likelihood function,

$$\ell_n(\theta) = \sum_{i=1}^n \log A(Y_i) + n \log B(\theta) + \omega(\theta)^T \sum_{i=1}^n T(Y_i),$$

which is the solution to the following equation

$$0 = \frac{nB'(\theta)}{B(\theta)} + \omega'(\theta)^T \sum_{i=1}^n T(Y_i).$$

Solving the equation shows

$$\hat{\theta} = g^{-1} \left(\frac{1}{n} \sum_{i=1}^n T(Y_i) \right),$$

where g is the function

$$g(\theta) = -\omega'(\theta)^{-1} \frac{\partial}{\partial \theta} \log B(\theta) = -\omega'(\theta)^{-1} \frac{B'(\theta)}{B(\theta)}.$$

Now assume the focus parameter takes the value $k(\theta)$ in the family parameterized by (7) and let $\hat{\mu} = k(\hat{\theta})$ be the maximum likelihood estimator of this quantity. Then

$$\hat{\mu} = h \left[\frac{1}{n} \sum_{i=1}^n T(Y_i) \right],$$

where $h(x) = k[g^{-1}(x)]$. This shows that the maximum likelihood estimators in exponential families are functions of means. By Theorem 1, we get

$$c = (1/2) \text{Tr} \{ Hh[ET(Y)] \text{Var}[T(Y)] \}. \quad (8)$$

It is worth noting that (8) makes no assumptions about the parametric model being correctly specified. Hence, the expression in (8) holds true even when the true distribution of the data is not a member of the parametric family f_θ .

2.2 | Quantiles

We will now consider the case where $\hat{\mu}$ is a quantile. More specifically, for $Y_1, \dots, Y_n \sim F$ i.i.d. and some $p \in (0, 1)$, we will work with $(1 - \gamma)Y_{(j)} + \gamma Y_{(j+1)}$, where $Y_{(k)}$ is the k th order statistic and j and γ are functions of p and n . All of the standard quantile estimators in R and Python are on this form, with j and γ given in Table 1.

We start by considering the case when $Y_1, \dots, Y_n$ are uniformly distributed. The results are stated in the following lemma.

Lemma 1. *Let $U_1, \dots, U_n$ be i.i.d. uniformly distributed variables and set $\hat{\mu} = (1 - \gamma)U_{(j)} + \gamma U_{(j+1)}$ with γ and j defined by any of the rows in Table 1. Then, c takes the following form*

$$c + o(1) = nE(\hat{\mu} - p) = n[(j + \gamma)/(n + 1) - p], \quad (9)$$

as p is the p -quantile in the uniform distribution.

Proof. By standard results, the j th-order statistic follows a $\text{Beta}(j, n - j + 1)$ -distribution. The expression in (9) follows. ■

In many cases, the left-hand side of (9) converges as n grows to infinity. For the quantile versions in R with the `type` argument equal to 4, 5, ..., 9, this is the case, and the corresponding limits

are shown in Table 1. When `type` is equal to 1, 2, or 3, however, the expression to the left in (9) does not converge unless very specific conditions on how n changes from iteration to iteration is assumed. In applications, we recommend reading the limit of (9) off Table 1 when using a quantile estimator corresponding to one in the default `quantile` function in R with `type` equal to 4, ..., 9. When `type` equal to 1, 2, or 3 is used, the left-hand side of (9) can be estimated directly using j and γ from Table 1. The latter approach can, of course, also be used when `type` equals 4, ..., 9, but using the limit in Table 1 leads to relatively simpler formulas.

We will now use Lemma 1 to find an expression for c for more general distributions than the uniform.

Theorem 2. Let $Y_1, \dots, Y_n$ be i.i.d. from a distribution F with p -quantile μ for some fixed $p \in (0, 1)$. Assume that F is invertible and that there exists a neighborhood around μ on which F is thrice continuously differentiable. Let f be the density in the distribution. Then, when $\hat{\mu} = (1 - \gamma)Y_{(j)} + \gamma Y_{(j+1)}$, with j and γ defined by either of the rows in Table 1,

$$c + o(1) = nE(\hat{\mu} - \mu) = \frac{n}{f(\mu)} \left(\frac{j + \gamma}{n + 1} - p \right) - \frac{f'(\mu)p(1 - p)}{2f(\mu)^3} + O(n^{-1/2}). \quad (10)$$

provided $n^{3/2}(\hat{\mu} - \mu)^3$ is uniformly integrable. Furthermore, c can be estimated consistently by

$$\frac{n}{\widehat{f(\mu)}} \left(\frac{j + \gamma}{n + 1} - p \right) - \frac{\widehat{f'(\mu)}p(1 - p)}{2\widehat{f(\mu)}^3}, \quad (11)$$

where $\widehat{f(\mu)}$ and $\widehat{f'(\mu)}$ are some estimators going in probability to $f(\mu)$ and $f'(\mu)$, respectively.

Proof. We sketch the proof here. More details can be found in the Appendix B.

By the probability integral transform, $U_i = F(Y_i)$ is uniformly distributed. Furthermore, $Y_{(i)} = F^{-1}(U_{(i)})$ as F is increasing and preserves ordering of variables.

TABLE 1 Each row in the table describes the nonparametric quantile estimator $(1 - \gamma)Y_{(j)} + \gamma Y_{(j+1)}$.

Type	j	γ	Limit
1	$\lfloor pn \rfloor$	$I(np \neq \lfloor pn \rfloor)$	—
2	$\lfloor pn \rfloor$	$(1/2)I(np = \lfloor pn \rfloor) + I(np \neq \lfloor pn \rfloor)$	—
3	$\lfloor pn - 1/2 \rfloor$	$I(np \neq \lfloor pn \rfloor \text{ or } j \text{ even})$	—
4	$\lfloor pn \rfloor$	$np - \lfloor pn \rfloor$	$-p$
5	$\lfloor pn + 1/2 \rfloor$	$np + 1/2 - \lfloor pn + 1/2 \rfloor$	$1/2 - p$
6	$\lfloor pn + p \rfloor$	$np + p - \lfloor pn + p \rfloor$	0
7	$\lfloor pn + 1 - p \rfloor$	$np + 1 - p - \lfloor pn + 1 - p \rfloor$	$1 - 2p$
8	$\lfloor pn + (p + 1)/3 \rfloor$	$np + (p + 1)/3 - \lfloor pn + (p + 1)/3 \rfloor$	$1/3 - (2/3)p$
9	$\lfloor pn + p/4 \rfloor$	$np + p/4 - \lfloor pn + p/4 \rfloor$	$-(3/4)p$

Notes: The column *Type* gives the `type` argument that should be given to `quantile` in R to get the corresponding estimator. To the far right we have displayed c , the limits of (9) when they exist. The symbol $\lfloor \cdot \rfloor$ denotes truncation.

Hence,

$$\hat{\mu} = (1 - \gamma)F^{-1}(U_{(j)}) + \gamma F^{-1}(U_{(j+1)}). \quad (12)$$

A Taylor expansion around p shows

$$F^{-1}(U_{(j)}) = \mu + (F^{-1})'(p)(U_{(j)} - p) + (1/2)(F^{-1})''(p)(U_{(j)} - p)^2 + \delta'_n. \quad (13)$$

By arguments that can be found in the Appendix B, $\delta'_n = O_{pr}[(U_{(j)} - p)^3]$. All the standard versions of quantile estimators, have j chosen in such a way that $U_{(j)} - p = O_{pr}(1/\sqrt{n})$. Hence, $\delta'_n = O_{pr}(n^{-3/2})$. A similar result holds for $F^{-1}(U_{(j+1)})$.

By the inverse function theorem $(F^{-1})'(p) = f(\mu)^{-1}$ and $(F^{-1})''(p) = -f'(\mu)f(\mu)^{-3}$. Combining this with (12) and (13), shows

$$\hat{\mu} = \mu + f(\mu)^{-1}(\hat{\mu}_U - p) - f'(\mu)f(\mu)^{-3}[(1 - \gamma)(U_{(j)} - p)^2 + \gamma(U_{(j+1)} - p)^2] + \delta_n. \quad (14)$$

where $\delta_n = O_{pr}(n^{-3/2})$ and $\hat{\mu}_U = (1 - \gamma)U_{(j)} + \gamma U_{(j+1)}$.

Direct computations reveal

$$nE[(1 - \gamma)(U_{(j)} - p)^2 + \gamma(U_{(j+1)} - p)^2] = p(1 - p) + o(1).$$

Furthermore, the arguments in the Appendix B ensure that $\delta_n = o_{pr}(n^{-1})$ implies $E(\delta_n) = o(n^{-1})$ under the present conditions. Plugging this into (14), shows

$$nE(\hat{\mu} - \mu) = nE(\hat{\mu}_U - p)/f(\mu) - f'(\mu)p(1 - p)/2f(\mu)^3 + o(1).$$

Combining the above equation with Lemma 1, shows (10). By the continuous mapping theorem (11) is consistent for c . This concludes the proof ■

If a parametric model is known, we can estimate $f(\mu)$ and $f'(\mu)$ by their maximum likelihood estimate. When working with for instance nested models, one may then fit larger models by maximum likelihood and use the fit to estimate $f(\mu)$ and $f'(\mu)$. Otherwise, kernel density estimation can be used. In the Appendix A, we discuss the latter approach and give suggestion for bandwidths minimizing the MSE of the estimates of $f(\mu)$ and $f'(\mu)$.

2.3 | Maximum likelihood estimators

In some situations, we either know or at least strongly believe, that the true underlying distribution of the data belongs to some parametric family. When this is the case, the maximum likelihood estimator of the focus parameter will be both consistent and asymptotically unbiased for the true value. Hence, it can be used as the consistent estimator, $\hat{\mu}_0$, from Section 3.2. Because of this, we need expressions for c when $\hat{\mu}$ is a maximum likelihood estimator. Such expressions exist already. Cox and Snell (1968) derived formulas for the limit of $nE(\hat{\theta} - \theta_0)$ when $\hat{\theta}$ is a maximum likelihood estimator in a correctly specified model and θ_0 is the true value of this parameter. In most situations, however, we do not know which, if any, parametric model the distribution of the data belongs to. Using the results of Cox and Snell (1968) will therefore be too optimistic. In this

section, we will derive model robust expressions for c . These can be applied even when we do not believe that the chosen model is the true one.

Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. random variables from some underlying distribution with density or probability mass function, g , and assume we fit a parametric model, f_θ , to these data using maximum likelihood. We will let $\hat{\theta} \in \mathbb{R}^p$ denote the maximum likelihood estimator of θ , that is, the maximizer of $\ell_n(\theta) = \sum_{i=1}^n \log f_\theta(Y_i)$. Under the assumptions stated in White (1982), $\hat{\theta}$ converges in probability to the minimizer of the Kullback-Leibler divergence from g to f_θ ,

$$\text{KL}(g, f_\theta) = \mathbb{E}_g[\log g(Y) - \log f_\theta(Y)].$$

The minimizer of this expression is in some sense, the “least false” parameter and is the quantity we have denoted by θ_{lf} . By White (1982), we also know that the speed of convergence is $O_{pr}(1/\sqrt{n})$ and that

$$\sqrt{n}(\hat{\theta} - \theta_{\text{lf}}) \xrightarrow{d} N(0, J^{-1} K J^{-1}),$$

where

$$J = \mathbb{E}_g \left(-\frac{\partial^2}{\partial \theta \partial \theta^T} \Big|_{\theta=\theta_{\text{lf}}} \log f_\theta(Y) \right) \quad \text{and} \quad K = \text{Var}_g \left(\frac{\partial}{\partial \theta} \Big|_{\theta=\theta_{\text{lf}}} \log f_\theta(Y) \right). \quad (15)$$

In the above, g is added as a subscript to emphasize that we are taking expectations with respect to this distribution.

The normal limit of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ is proved by finding the root of the one term Taylor expansion of $\nabla \ell_n$ and showing that this is only $o_{pr}(1/\sqrt{n})$ away from $\hat{\theta}$. This allows us to study the behavior of an explicit expression rather than the, in general, only implicitly defined $\hat{\theta}$. For our purposes, however, a linear approximation is not sufficient, as we want to estimate the bias with an error of order $o(1/n)$ and therefore need higher precision than $o_{pr}(1/\sqrt{n})$. This is the main idea in the proof of the following proposition.

Theorem 3. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from a distribution F and let f_θ be some parametric family of densities or probability point mass functions indexed by an open set $\Theta \subseteq \mathbb{R}^p$. Furthermore, let $\hat{\theta}$ be the maximum likelihood estimator when fitting this family to the data and θ_{lf} the minimizer of the Kullback–Leibler divergence from F to f_θ . Then, the bias of $\hat{\theta}$ is given by the following formula,*

$$c + o(1) = n\mathbb{E}(\hat{\theta} - \theta_{\text{lf}}) = J^{-1} \left[\begin{pmatrix} \text{Tr}(J^{-1}V_1) \\ \vdots \\ \text{Tr}(J^{-1}V_p) \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \text{Tr}(W_1 J^{-1} K J^{-1}) \\ \vdots \\ \text{Tr}(W_p J^{-1} K J^{-1}) \end{pmatrix} \right] + o(1), \quad (16)$$

where J and K are defined in (15) and

$$V_j = \text{Cov} \left(\frac{\partial}{\partial \theta} \Big|_{\theta_{\text{lf}}} \log f_\theta(Y), \frac{\partial^2}{\partial \theta \partial \theta_j} \Big|_{\theta_{\text{lf}}} \log f_\theta(Y) \right) \quad \text{and} \quad W_j = \mathbb{E} \left(\frac{\partial^3}{\partial \theta \partial \theta^T \partial \theta_j} \Big|_{\theta_{\text{lf}}} \log f_\theta(Y) \right) \\ \text{for } j = 1, \dots, p, \quad (17)$$

provided the following four regularity conditions hold true:

- (A1) All partial derivatives up to and including order four of $\log f_\theta(y)$ at θ_{lf} exists and are continuous for F -almost all y . In addition, with $Y \sim F$ we need all fourth order moments of $\nabla_{\theta_{\text{lf}}} \log f_\theta(Y)$ and all second order moments of all second- and third-order partial derivatives of $\log f_\theta(Y)$ to exist and be finite.
- (A2) The matrix J is positive definite.
- (A3) All fourth-order powers of all components of $\sqrt{n}[\hat{\theta} - \theta_{\text{lf}} - (nJ)^{-1} \nabla \ell_n(\theta_{\text{lf}})]$ are uniformly integrable.
- (A4) There exists a function $m : \mathbb{R}^d \rightarrow \mathbb{R}$ such that in a neighborhood of θ_{lf} all fourth-order partial derivatives of $\log f_\theta(y)$ are bounded by $m(y)$ for F -almost all y , and $E m(Y)^4$ exists and is finite.

In addition, c can be estimated consistently by replacing the matrices J and K and V_j and W_j for $j = 1, \dots, p$ by the empirical means, covariances, and variances of the corresponding functions evaluated at $\theta = \hat{\theta}$.

Proof. As mentioned previously, we will only sketch the proof of Theorem 3. For full arguments taking care of all details, consult the Appendix B.

Fix j and let η_j denote the j th component function of $\nabla \ell_n$, that is, $\partial \ell_n / \partial \theta_j$. Taylor expanding the function around θ_{lf} , shows

$$\eta_j(\theta) = \eta_j(\theta_{\text{lf}}) + (\theta - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}}) + (1/2)(\theta - \theta_{\text{lf}})^T H \eta_j(\theta_{\text{lf}})(\theta - \theta_{\text{lf}}) + \epsilon_n(\theta).$$

If η_j is sufficiently regular, we have $\epsilon_n(\theta) = O(n\|\theta - \theta_{\text{lf}}\|^3)$. At $\hat{\theta}$ it therefore holds that

$$0 = \eta_j(\theta_{\text{lf}}) + (\hat{\theta} - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}}) + (1/2)(\hat{\theta} - \theta_{\text{lf}})^T H \eta_j(\theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}}) + O_{pr}(1/\sqrt{n}), \quad (18)$$

since $\hat{\theta} - \theta_{\text{lf}} = O_{pr}(1/\sqrt{n})$. Here we have used that the $\hat{\theta}$ maximizing ℓ_n satisfies $\eta_j(\hat{\theta}) = 0$.

Since θ_{lf} minimizes the Kullback–Leibler divergence, the derivative of $\theta \mapsto E \log f_\theta(Y)$ at θ_{lf} is zero. Hence,

$$E \eta_j(\theta_{\text{lf}}) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \Big|_{\theta=\theta_{\text{lf}}} E \log f_\theta(Y_i) = 0.$$

Because of this, taking the expectation of both sides of (18) shows

$$E[(\hat{\theta} - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}})] + (1/2)E[(\hat{\theta} - \theta_{\text{lf}})^T H \eta_j(\theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}})] + O(1/\sqrt{n}) = 0. \quad (19)$$

We will work with each term in (19) separately, and use the result to find an expression for $E(\hat{\theta} - \theta_{\text{lf}})$ with remainder term smaller than $o(1/n)$. In the following, let $b_n = E(\hat{\theta} - \theta_{\text{lf}})$.

We start with the first term. By definition of η_j and properties of the trace operator,

$$E[(\hat{\theta} - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}})] = E \left((\hat{\theta} - \theta_{\text{lf}})^T \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}}) \right) = \text{Tr} E \left((\hat{\theta} - \theta_{\text{lf}}) \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}})^T \right).$$

By definition of the covariance, this equals

$$\begin{aligned} & \text{Tr} \left[b_n \mathbb{E} \left(\frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}}) \right)^T \right] + \text{TrCov} \left(\hat{\theta} - \theta_{\text{lf}}, \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}}) \right) \\ &= b_n^T \mathbb{E} \left(\frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}}) \right) + \text{TrCov} \left(\hat{\theta} - \theta_{\text{lf}}, \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}}) \right). \end{aligned}$$

We recognize the vector $\mathbb{E} \partial^2 \ell_n(\theta_{\text{lf}}) / \partial \theta_j \partial \theta$ as $-n$ times the j th column of the Fisher matrix J . Let this be denoted by J^j . Furthermore, White (1982) showed that

$$\hat{\theta} - \theta_{\text{lf}} = J^{-1} n^{-1} \ell'_n(\theta_{\text{lf}}) + o_{pr}(1/\sqrt{n}). \quad (20)$$

Hence,

$$\mathbb{E}[(\hat{\theta} - \theta_{\text{lf}})^T \nabla \eta_j(\theta_{\text{lf}})] = -nb_n^T J^j + \frac{1}{n} \text{Tr} \left[J^{-1} \text{Cov} \left(\ell'_n(\theta_{\text{lf}}), \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}}) \right) \right] + o(1).$$

Let u be the score function and u_j its j th component function. Then, since the Y_i s are independent and identically distributed,

$$\text{Cov} \left(\ell'_n(\theta_{\text{lf}}), \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta}(\theta_{\text{lf}}) \right) = \sum_{i=1}^n V_j = nV_j,$$

with V_j defined in (17). Putting all of our results together, shows that the first term of (19) is equal to

$$-nb_n^T J^j + \text{Tr}(J^{-1} V_j) + o(1). \quad (21)$$

The second term of (19) requires less effort. First notice that

$$(\hat{\theta} - \theta_{\text{lf}})^T H \eta_j(\theta_{\text{lf}}) (\hat{\theta} - \theta_{\text{lf}}) = \sqrt{n} (\hat{\theta} - \theta_{\text{lf}})^T n^{-1} H \eta_j(\theta_{\text{lf}}) \sqrt{n} (\hat{\theta} - \theta_{\text{lf}}).$$

By White (1982), $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}}) \xrightarrow{d} U$ where $U \sim N(0, J^{-1} K J^{-1})$. Furthermore, by the law of large numbers $H \eta_j(\theta_{\text{lf}}) / n = W_j + o_{pr}(1)$, where W_j is defined in (17). Hence, $(\hat{\theta} - \theta_{\text{lf}})^T H \eta_j(\theta_{\text{lf}}) (\hat{\theta} - \theta_{\text{lf}}) \xrightarrow{d} U^T W_j U$. This implies

$$\mathbb{E}[(\hat{\theta} - \theta_{\text{lf}})^T H \eta_j(\theta_{\text{lf}}) (\hat{\theta} - \theta_{\text{lf}})] = \text{Tr}(W_j J^{-1} K J^{-1}) + o(1). \quad (22)$$

Here we have again used properties of the trace operator multiple times.

We are now ready to complete the argument. Combining (19), (21), and (22) for $j = 1, \dots, p$, shows

$$0 = -nb_n^T J^j + \text{Tr}(J^{-1} V_j) + \text{Tr}(W_j J^{-1} K J^{-1}) + o(1),$$

for $j = 1, \dots, p$. These p equations hold simultaneously if and only if

$$0 = -nJb_n + \begin{pmatrix} \text{Tr}(J^{-1} V_1) \\ \vdots \\ \text{Tr}(J^{-1} V_p) \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \text{Tr}(W_1 J^{-1} K J^{-1}) \\ \vdots \\ \text{Tr}(W_p J^{-1} K J^{-1}) \end{pmatrix} + o(1).$$

Rearranging the terms and multiplying the equation by $(nJ)^{-1}$, shows (16).

Consistency of the estimators defined in Theorem 3 is shown in the Appendix B. ■

If one of the parameters of the model itself is chosen as focus parameter, Theorem 3 can be used to estimate c . In most situations, however, $\hat{\mu}$ is a function of $\hat{\theta}$, and the above expressions do not suffice. This is a special case of what we will discuss in Section 2.5. First, we will, however, give extensions of Theorem 3 to the more general case of M- and Z-estimators.

2.3.1 | M- and Z-estimators

The calculations made in the proof of Theorem 3 are not limited to maximum likelihood estimators. Similar arguments can be made to find the bias of a more general class of estimators: M- and Z-estimators. Examples of such estimators include copulas fitted using a two-stage procedure, see Joe (2005) and Ko et al. (2019), and robust M-estimators, see for example, Huber (2009).

M-estimators are maximizers of expressions on the form

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^n \zeta(Y_i, \theta), \quad (23)$$

where $\zeta : \mathbb{R}^{d+p} \rightarrow \mathbb{R}$. The most obvious example of such estimators is maximum likelihood estimators. In this case $\zeta(y, \theta) = \log f_{\theta}(y)$. Because of their similarity, M-estimators share many properties with maximum likelihood estimators. It can for instance be shown that the maximizer of (23), $\hat{\theta}$, converges almost surely to θ_0 , the maximizer of the limit function $\theta \mapsto E\zeta(Y, \theta)$, and that $\sqrt{n}(\hat{\theta} - \theta_0)$ has a normal limit, or in particular, that

$$\sqrt{n}(\hat{\theta} - \theta_0) = -V^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial \theta} \Big|_{\theta=\theta_0} \zeta(Y_i, \theta) + o_{pr}(1),$$

where V is the Hessian matrix of the limit function $\theta \mapsto E\zeta(Y, \theta)$ at θ_0 . This expression is an analogue to (20). For more details and proofs, consult Chap. 4 of Van der Vaart (1998).

Looking at the proof of Theorem 3 we notice that the arguments can be repeated for the more general class of M-estimators. This is achieved by replacing $\log f_{\theta}(y)$ by the function ζ and using the properties listed in the previous paragraph in place of the corresponding results for maximum likelihood estimators. The argument is very similar and will not be repeated here, but the result is stated in the following theorem.

Theorem 4. Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from a distribution F and $\hat{\theta} \in \mathbb{R}^p$ be an M-estimator maximizing $n^{-1} \sum_{i=1}^n \zeta(Y_i, \theta)$. Furthermore, let $\theta_0 \in \mathbb{R}$ be the maximizer of the limit function $\theta \mapsto E\zeta(Y, \theta)$ where $Y \sim F$. Then the bias of $\hat{\theta}$ is given by the following formula,

$$c + o(1) = nE(\hat{\theta} - \theta_0) = J^{-1} \left[\begin{pmatrix} \text{Tr}(J^{-1}V_1) \\ \vdots \\ \text{Tr}(J^{-1}V_p) \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \text{Tr}(W_1J^{-1}KJ^{-1}) \\ \vdots \\ \text{Tr}(W_pJ^{-1}KJ^{-1}) \end{pmatrix} \right] + o(1),$$

where J and K are defined as

$$J = E \left(-\frac{\partial^2 \zeta}{\partial \theta \partial \theta^T}(Y, \theta_0) \right) \quad \text{and} \quad K = \text{Var} \left(\frac{\partial \zeta}{\partial \theta}(Y, \theta_0) \right).$$

and

$$V_j = \text{Cov} \left(\frac{\partial \zeta}{\partial \theta}(Y, \theta_0), \frac{\partial^2 \zeta}{\partial \theta \partial \theta_j}(Y, \theta_0) \right) \quad \text{and} \quad W_j = E \left(\frac{\partial^3 \zeta}{\partial \theta \partial \theta^T \partial \theta_j}(Y, \theta_0) \right) \quad \text{for } j = 1, \dots, p,$$

provided the regularity conditions (A1–A4) of Theorem 3 hold true after replacing $\log f$ with ζ , θ_{lf} with θ_0 and the maximum likelihood estimator with the maximizer of $n^{-1} \sum_{i=1}^n \zeta(Y_i, \theta)$.

In addition, c can be estimated consistently by replacing the matrices J and K and V_j and W_j for $j = 1, \dots, p$ by the empirical means, covariances, and variances of the corresponding functions evaluated at $\theta = \hat{\theta}$.

Assume that the function ζ is differentiable in θ . Maximization of (23) can then be rephrased as solving

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \zeta}{\partial \theta}(Y_i, \theta) = 0.$$

Solutions to equations like the one above, are called Z-estimators. In the proof of Theorem 3, it is the fact that the maximum likelihood estimator is a Z-estimator which is used in the arguments. Because of this, the result holds true also for this class of estimators, as stated in the following theorem. The proof is very similar to that of Theorem 3 and is left out.

Theorem 5. Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from a distribution F and $\hat{\theta} \in \mathbb{R}^p$ be a Z-estimator solving $0 = n^{-1} \sum_{i=1}^n \xi(Y_i, \theta)$. Furthermore, let $\theta_0 \in \mathbb{R}$ be the solution to the limit equation $0 = E\xi(Y, \theta)$ where $Y \sim F$. Then the bias of $\hat{\theta}$ is given by the following formula,

$$c + o(1) = nE(\hat{\theta} - \theta_0) = J^{-1} \left[\begin{pmatrix} \text{Tr}(J^{-1} V_1) \\ \vdots \\ \text{Tr}(J^{-1} V_p) \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \text{Tr}(W_1 J^{-1} K J^{-1}) \\ \vdots \\ \text{Tr}(W_p J^{-1} K J^{-1}) \end{pmatrix} \right] + o(1),$$

where J and K are defined as

$$J = E \left(-\frac{\partial \xi}{\partial \theta}(Y, \theta_0) \right) \quad \text{and} \quad K = \text{Var} \xi(Y, \theta_0).$$

and

$$V_j = \text{Cov} \left(\xi(Y, \theta_0), \frac{\partial \xi_j}{\partial \theta}(Y, \theta_0) \right) \quad \text{and} \quad W_j = E \left(\frac{\partial^2 \xi_j}{\partial \theta \partial \theta^T}(Y, \theta_0) \right) \quad \text{for } j = 1, \dots, p,$$

provided the regularity conditions (A1–A4) of Theorem 3 hold true after replacing $\partial \log f / \partial \theta$ with ξ , θ_{lf} with θ_0 and the maximum likelihood estimator with the solution to $0 = n^{-1} \sum_{i=1}^n \zeta(Y_i, \theta)$. In the above ξ_j is the j th component function of ξ .

In addition, c can be estimated consistently by replacing the matrices J and K and V_j and W_j for $j = 1, \dots, p$ by the empirical means, covariances and variances of the corresponding functions evaluated at $\theta = \hat{\theta}$.

Lastly, we remark that Theorem 4 and Theorem 5 hold true even when $\hat{\theta}$ is only “almost” an M- or a Z-estimator. In fact, it is sufficient that $\hat{\theta}$ is an estimator which can be expressed as the root of a function on the form

$$\theta \mapsto \sum_{i=1}^n \xi(Y_i, \theta) + \delta_n(\theta),$$

where $\delta_n(\theta) = o_{pr}(1)$ uniformly in a neighbourhood of the root of the limit function $\theta \mapsto E\xi(Y, \theta)$. This follows from the fact that $\delta_n(\theta)$ can be incorporated into the remainder term of (18).

2.4 | Resampling techniques

The last way of estimating c we will present is resampling techniques. We will discuss two methods: the bootstrap and the jackknife. The main ideas will be presented briefly here, but additional details, proofs, and discussions can be found in for example, Efron (1982), Efron and Tibshirani (1993), or Shao and Tu (1995).

The bootstrap was first introduced in Efron (1979), and has since become a popular statistical tool. To see how the procedure can be used to estimate c , notice that by (1)

$$E[n(\hat{\mu} - \mu)] = nE\epsilon_n = c + o(1).$$

The left-hand side of this equation can be easily estimated with bootstrap. Let $\hat{\mu}_j^*$ be the estimator of μ based on the j th out of B bootstrap samples. Then

$$\hat{c}_{\text{boot}} = \frac{n}{B} \sum_{j=1}^B (\hat{\mu}_j^* - \hat{\mu}),$$

is an estimator of c , see for example, Chap. 10 of Efron and Tibshirani (1993).

The bootstrap procedure is quite general and works for most focus parameters. Hence, this technique is an option when all other methods fail. Using the procedure in practice is, however, often inconvenient, as a very large number of bootstrap samples must be drawn to achieve sufficient precision. To see this, notice that the bootstrap estimate $B^{-1} \sum_{j=1}^B (\hat{\mu}_j^* - \hat{\mu})$ is a Monte Carlo estimate of its expectation. Since the variance of $\hat{\mu}_j^*$ is of order $O(1/n)$ by (1), the central limit theorem ensures that the Monte Carlo error of $B^{-1} \sum_{j=1}^B (\hat{\mu}_j^* - \hat{\mu})$ is of order $O_{pr}(1/\sqrt{nB})$. Hence, the error of $\hat{c}_{\text{boot}}$ is of order $O_{pr}(\sqrt{n}/\sqrt{B})$. For $\hat{c}_{\text{boot}}$ to be consistent we would therefore need $B \gg n$. To achieve a convergence rate of $O_{pr}(1/\sqrt{n})$ for c we would need a million bootstrap samples when $n = 1000$, and even with a mere 50 data points, we would need B to be greater than 2000. Because of this, the computational burden might be too heavy for bootstrap to be practical in many situations.

The jackknife is another resampling technique. It was introduced in Quenouille (1949) and Tukey (1958), and was developed specifically for the purpose of estimating the bias with an error of order $o(1/n)$. It is therefore well suited for estimating c .

The jackknife estimate of the bias of $\hat{\mu}$ is defined as

$$\hat{c}_{\text{jack}} = (n-1) \left(\hat{\mu} - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_{-i} \right).$$

Here $\hat{\mu}_{-i}$ is the same estimator of μ as $\hat{\mu}$, but based on $Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n$, for $i = 1, \dots, n$ rather than the full dataset. More formally, we assume there is a functional T , such that $\hat{\mu} = T(F_n)$ and $\hat{\mu}_{-i} = T(F_n^{-i})$, where F_n^{-i} is the empirical distribution function based on all but the i th data point. Under certain smoothness conditions, $\hat{c}_{\text{jack}}$ is consistent for the true value c . This is shown in Thm. 3.1 of Shao (1991).

The jackknife procedure is a flexible and general tool for bias estimation. That being said, the method requires n applications of the functional T to estimate c . In cases where application of the functional is computationally expensive or n is very large, the jackknife estimate can be slow to calculate. To ease the computational burden, it is possible to use less computer intensive versions of the standard jackknife procedure. Options include the one-step jackknife and the grouped jackknife. See Chap. five of Shao and Tu (1995) for an overview. These methods limit the number of times T needs to be applied, reducing the overall computational cost. In certain situations, however, resampling techniques can be too time consuming to be practical, even with modifications like these. In such cases, the formulas derived in the previous sections can be used instead.

2.5 | Functions of the above

In the previous sections, we derived formulas and described strategies for estimating c in certain situations. We will now show how these results can be used to estimate the bias of even more complex estimators.

Theorem 6. *Let $\hat{\theta}$ be some estimator of θ for which c_θ , the limit of $nE(\hat{\theta} - \theta)$ is known. Assume furthermore that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a $N(0, \Sigma)$ -distributed variable. and that $h : \mathbb{R}^p \rightarrow \mathbb{R}$ is a function for which all partial derivatives up to an including order three exists and are continuous in a neighborhood of θ . Then if $\mu = h(\theta)$ and $\hat{\mu} = h(\hat{\theta})$,*

$$c = \nabla h(\theta)^T c_\theta + (1/2)\text{Tr}[Hh(\theta)\Sigma]. \quad (24)$$

Furthermore, if $\hat{c}_\theta$ is a consistent estimator of c_θ ,

$$\nabla h(\hat{\theta})^T \hat{c}_\theta + (1/2)\text{Tr}[Hh(\hat{\theta})\hat{\Sigma}], \quad (25)$$

converges in probability to c .

Notice the similarities with Theorem 1. When $\hat{\theta}$ is unbiased, the first term of (24) reduces to (4).

Proof. As before, we present a proof sketch only. A full argument is similar to that of Theorem 1. Consult the proof of this theorem in the Appendix B for more details.

A Taylor expansion reveals the following

$$\hat{\mu} = \mu + \nabla h(\theta)^T(\hat{\theta} - \theta) + (1/2)(\hat{\theta} - \theta)^T Hh(\theta)(\hat{\theta} - \theta) + o_{pr}(1/n).$$

Let $b_n = E(\hat{\theta} - \theta)$. Taking expectations on both sides of the above equation, shows

$$nE(\hat{\mu} - \mu) = \nabla h(\theta)^T n b_n + (n/2)E[(\hat{\theta} - \theta)^T Hh(\theta)(\hat{\theta} - \theta)] + o(1),$$

which, by standard rules for the trace operator, is equivalent to $nE(\hat{\mu} - \mu) = \nabla h(\theta)^T n b_n + (1/2)\text{Tr}[Hh(\theta)\Sigma] + o(1)$. By assumption $n b_n$ converges to c_θ , and hence,

$$c = \nabla h(\theta)^T c_\theta + (1/2)\text{Tr}[Hh(\theta)\Sigma].$$

Since $\hat{c}_\theta$ is consistent for c_θ , (25) converges to c by the continuous mapping theorem. ■

Now that we have discussed c in detail, we are ready to move onto our main application: estimation of the MSE and the FIC.

3 | ESTIMATING MSE

We will start by introducing notation and assumptions. In addition, we will derive equations that show explicitly the quantities needed to estimate the MSE.

3.1 | An approximation to the MSE

Assume we have i.i.d. data $Y_1, \dots, Y_n \in \mathbb{R}^d$ from a distribution, F . Let $\hat{\mu}_1$ be an estimator of some focus parameter with true value $\mu_0 \in \mathbb{R}$ based on these data, and assume $\hat{\mu}_1$ admits an influence function, v_1 , with a small enough remainder term, in the sense that

$$\hat{\mu}_1 = \mu_1 + \frac{1}{n} \sum_{i=1}^n v_1(Y_i) + S_n, \quad (26)$$

with $E v_1(Y) = 0$ and $S_n = O_{pr}(1/n)$. Lastly, let c_1 denote the limit of $n E S_n$.

Direct computations show

$$\text{MSE}(\hat{\mu}_1) = E(\hat{\mu}_1 - \mu_0)^2 = b^2 + \frac{\tau}{n} + \frac{2bc_1}{n} + 2E \left(\frac{S_n}{n} \sum_{i=1}^n v_1(Y_i) \right) + E S_n^2 + o(1/n), \quad (27)$$

where $b = \mu_1 - \mu_0$ and τ is the variance of $v_1(Y)$. By the above assumptions $S_n = o_{pr}(1/\sqrt{n})$. Hence, $S_n^2 = o_{pr}(1/n)$. In addition, $n^{-1} \sum_{i=1}^n v_1(Y_i) = O_{pr}(1/\sqrt{n})$ by the central limit theorem. Because of this, $n^{-1} S_n \sum_{i=1}^n v_1(Y_i) = o_{pr}(1/n)$. If we further assume S_n^2 and $S_n n^{-1} \sum_{i=1}^n v_1(Y_i)$ to be uniformly integrable, the above is sufficient for the last two terms in (27) to be of order $o(1/n)$,

see for example, Billingsley (1999, p. 31). Hence, the MSE of $\hat{\mu}_1$ takes the following form

$$\text{MSE}(\hat{\mu}_1) = b^2 + \tau/n + 2bc_1/n + o(1/n).$$

We summarize our findings in a theorem for easier reference.

Theorem 7. Assume $Y_1, \dots, Y_n \in \mathbb{R}^d$ are i.i.d. from a distribution F , and $\hat{\mu}_1$ is an estimator of μ_0 satisfying

$$\hat{\mu}_1 = \mu_1 + \frac{1}{n} \sum_{i=1}^n v_1(Y_i) + S_n,$$

for some $S_n = O_{pr}(1/n)$. Furthermore, let $b = \mu_1 - \mu_0$, and take c_1 to be the limit of nES_n . Then

$$\text{MSE}(\hat{\mu}_1) = b^2 + \tau/n + 2bc_1/n + o(1/n), \quad (28)$$

where τ is the variances of $v_1(Y)$, provided this quantity exists and is finite and nS_n is uniformly integrable.

To estimate the MSE of $\hat{\mu}_1$, we need to approximate all quantities in (28). This will be the topic of the next section.

3.2 | Estimating to the correct order

Looking at (28), we notice that the limiting bias, b , is constant while the remaining terms decrease at the speed of $O(1/n)$. Because of this, b will dominate when n gets large. Hence, minimizing the MSE asymptotically reduces to choosing a $\hat{\mu}_1$ with $b = 0$. In practice, however, we do not have infinite data points, and blindly choosing the estimator with the lowest limit bias is therefore rarely useful. Most applications require extremely flexible estimators to ensure that the bias disappears in the limit. Such procedures often have high variance. As a result, their MSEs tend to be high in finite samples. Because of this, many statistical applications instead attempt to find a trade-off between bias and variance. For our MSE estimates to be useful, they therefore need to take both the bias and variance into account. Since the latter is of order $O(1/n)$, we will need to create an approximation with an error of a smaller order. Otherwise, the variance is on the same scale as the error of the estimator, washing out the effect of the variability.

We will create an estimator of the MSE which is only $o(1/n)$ away from (28) in expected value. To do this, we need to estimate three quantities: the asymptotic bias b , the variance τ of $v_1(Y)$ and the limit c_1 of nES_n . If we can find asymptotically unbiased estimators for these quantities, the two last terms in (28) can be estimated with an error of $o(1/n)$, due to the n appearing in the denominator of these terms. We do, however, need to be extra careful with the estimator of the squared bias. This quantity is not divided by n , and hence, we need to make sure that the bias of the estimator is of order $o(1/n)$.

To estimate the asymptotic bias, we need to approximate the distance from the limit of our estimator, μ_1 , to the truth, μ_0 . This is impossible to do without having some idea of the value of μ_0 . Because of this, we need to introduce a second estimator which is consistent for this true value. We will call this estimator $\hat{\mu}_0$ and make similar assumptions about its form to the ones made for

$\hat{\mu}_1$, that is,

$$\hat{\mu}_0 = \mu_0 + \frac{1}{n} \sum_{i=1}^n v_0(Y_i) + T_n, \quad (29)$$

with T_n being an error term of order $O_{pr}(1/n)$ and $v_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ a function such that $E v_0(Y) = 0$.

The introduction of $\hat{\mu}_0$ might seem slightly counter-intuitive at first glance, and one might wonder why we would bother with $\hat{\mu}_1$ when we have a consistent estimator available. The answer to this concern is that even though $\hat{\mu}_1$ is inconsistent for μ_0 , its MSE might still be lower than that of the consistent estimator $\hat{\mu}_0$, making it preferable for estimation. In addition, $\hat{\mu}_0$ does not need to satisfy all the conditions we require of $\hat{\mu}_1$. The most obvious example of this is when we wish to fit a parametric model to the data. In such cases we might “need” $\hat{\mu}_1$ to be a maximum likelihood estimator, while this condition is not required of $\hat{\mu}_0$. This is a crucial point when working with the FIC, which will be discussed in Section 4.

Since $\hat{\mu}_0$ is consistent for the true value μ_0 , the asymptotic bias can be estimated consistently by the following quantity

$$\hat{b} = \hat{\mu}_1 - \hat{\mu}_0 = b + \frac{1}{n} \sum_{i=1}^n \phi(Y_i) + R_n,$$

where $\phi = v_1 - v_0$ and $R_n = S_n - T_n$. Squaring the above expression shows

$$\hat{b}^2 = b^2 + \left(\frac{1}{n} \sum_{i=1}^n \phi(Y_i) \right)^2 + R_n^2 + 2 \left(\frac{b + R_n}{n} \sum_{i=1}^n \phi(Y_i) + b R_n \right).$$

Taking expectations on both sides of the equation and arguing as in Section 3.1, reveals

$$E \hat{b}^2 = b^2 + \kappa/n + 2bE R_n + o(1/n),$$

under the assumption of uniform integrability of R_n^2 and $n^{-1} R_n \sum_{i=1}^n \phi(Y_i)$. In the above, $\kappa = \text{Var} \phi(Y_i)$. Since $R_n = S_n - T_n$, we further have

$$E \hat{b}^2 = b^2 + 2b c_1/n + \kappa/n - 2bE T_n + o(1/n).$$

Just as for S_n , $E T_n$ will typically be of order $O(1/n)$. So let c_0 be the limit of $nE T_n$. Then, the above equation implies

$$E \hat{b}^2 = b^2 + 2b c_1/n + \kappa/n - 2b c_0/n + o(1/n).$$

We summarize the results in a theorem for easier reference.

Theorem 8. Assume $Y_1, \dots, Y_n \in \mathbb{R}^d$ are i.i.d. from a distribution F , and $\hat{\mu}_1$ and $\hat{\mu}_0$ are estimators satisfying

$$\hat{\mu}_1 = \mu_1 + \frac{1}{n} \sum_{i=1}^n v_1(Y_i) + S_n \quad \text{and} \quad \hat{\mu}_0 = \mu_0 + \frac{1}{n} \sum_{i=1}^n v_0(Y_i) + T_n.$$

for some $S_n, T_n = O_{pr}(1/n)$. Furthermore, let $\hat{b} = \hat{\mu}_1 - \hat{\mu}_0$ and $b = \mu_1 - \mu_0$, and take c_0 and c_1 to be the limits of nET_n and nES_n . Then

$$E\hat{b}^2 = b^2 + 2bc_1/n + \kappa/n - 2bc_0/n + o(1/n), \quad (30)$$

where κ and τ are the variances of $\phi(Y) = v_1(Y) - v_0(Y)$ and $v_1(Y)$, respectively, provided κ and τ exists and are finite and nS_n and nT_n are uniformly integrable.

The first two terms in (30) appear in (28), the asymptotic approximation we wish to estimate. The last two terms, however, do not and need to be corrected for. If we can find consistent estimators, $\hat{c}_0$ and $\hat{\kappa}$, for c_0 and κ , respectively, the following estimator is only $o(1/n)$ away from $b^2 + 2bc_1/n$ in expected value

$$\hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}_0/n. \quad (31)$$

The approximation $\hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}_0/n + \hat{\tau}/n$ of the MSE, therefore has an error of order $o(1/n)$ in expectation. Here $\hat{\kappa}$ and $\hat{\tau}$ are the empirical variances of $\phi(Y_1), \dots, \phi(Y_n)$ and $v_1(Y_1), \dots, v_1(Y_n)$, respectively.

Sometimes, (31) will be negative, leading to an estimate of the squared bias which less than zero. By definition, however, $b^2 \geq 0$. Negative estimates are therefore both counter-intuitive and always further away from the true value than 0. Hence, when negative estimates show up, we will truncate them to zero, as in Jullum and Hjort (2017). This leads to the following estimator of the MSE of $\hat{\mu}_1$

$$\widehat{\text{MSE}}(\hat{\mu}_1) = \max\{0, \hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}_0/n\} + \hat{\tau}/n. \quad (32)$$

Estimators for b , κ , and τ have already been discussed. By definition c_0 is the limit of $nE(\hat{\mu}_0 - \mu_0)$, and estimation of this quantity was discussed in length in Section 2.

Remark 2. The estimator in (32) is not the only way to estimate the MSE of $\hat{\mu}_1$. Rather than adding $2\hat{b}\hat{c}_0/n$ to $\hat{b}^2$ to remove the bias, we could subtract $\hat{c}_0/n$ from $\hat{\mu}_0$ before estimating b . This would correct the bias of $\hat{\mu}_0$ and remove the need to add $2\hat{b}\hat{c}_0/n$ to the squared bias estimate. Such a procedure is indeed an alternative to using (32), but does not lead to a significantly different estimator. By correcting the bias of $\hat{\mu}_0$ before using it to estimate the MSE of $\hat{\mu}_1$, we are left with the estimator $\hat{b}^2 + 2\hat{b}\hat{c}_0/n + \hat{c}_0^2/n^2 - \hat{\kappa}/n$ of the squared bias. The expected difference between this expression and (30) is a term of order $O(1/n^2)$. As we are attempting to create estimators with biases of order $o(1/n)$, the difference between the two approaches is negligible.

That being said, estimating and subtracting $\hat{c}_0$ is only one out of many ways to remove bias from $\hat{\mu}_0$. Modifying the score function as in Firth (1993), removes the $O(1/n)$ part of the bias of the maximum likelihood estimator. For M-estimators, similar alterations can be made to the moment equations to achieve approximate unbiasedness, see Kim (2016). For situations where such techniques are available, correcting the bias of $\hat{\mu}_0$ before using it to estimate b is indeed an alternative to the methods suggested in this article.

4 | THE FOCUSED INFORMATION CRITERION

As mentioned previously, estimation of the MSE has been developed recently in the literature concerning the FIC. This criterion attempts to choose the model with the most precise estimate of a prespecified focus parameter, μ , rather than the one fitting the data the best overall. This is achieved by ranking candidate models by the MSE of their estimate of μ . The estimator in (32) can therefore be seen, not only as a way to approximate MSEs, but indeed as an information criterion.

Assume we fit a parametric model f_θ using maximum likelihood. The resulting estimator $\hat{\theta}$ of θ then satisfies the following equation

$$\hat{\theta} = \theta + J(\theta_{\text{lf}})^{-1} \frac{1}{n} \sum_{i=1}^n u(Y_i, \theta_{\text{lf}}) + O_{pr}(1/n),$$

where u is the score function and θ_{lf} and $J(\theta_{\text{lf}})$ are as defined in Section 3.1. See for example, Thm. 5.39 in Van der Vaart (1998) for a proof and a set of conditions. Applying the delta method ensures that the focus parameter satisfies (26) with $v_1(y) = \nabla \mu(\theta_{\text{lf}})^T J(\theta_{\text{lf}})^{-1} u(y, \theta_{\text{lf}})$, where ∇ denotes the gradient operator. Because of this, the MSE of $\psi(\hat{\theta})$ can be estimated using (32), provided a consistent estimator of the true value exists and satisfies the conditions stated previously in this chapter. Since the FIC of the parametric model is the MSE of $\psi(\hat{\theta})$, we can compute the FIC of f_θ using (32). A related idea is used in Jullum and Hjort (2017), and their formulas are indeed quite similar to ours. We will now discuss the FIC of Jullum and Hjort (2017) in closer detail and use (32) to define a new FIC.

4.1 | The new and old FIC

In Jullum and Hjort (2017), the authors derive a FIC, allowing for comparison of multiple nonnested parametric models and nonparametric alternatives, by evaluating how well a certain pre-specified parameter of interest μ is estimated. This quantity is called the focus parameter, and a model's FIC score is an approximation to the MSE of its estimator $\hat{\mu}_1$ of μ .

To derive the criterion, the authors proceed more or less as we have done in Section 3. They argue that $\hat{b}^2 = (\hat{\mu}_1 - \hat{\mu}_0)^2$ tends to overshoot the actual bias of $\hat{\mu}_1$ and that $\kappa/n = \text{Var}\phi(Y_i)/n$ should be subtracted from the expression to correct for this. Using our notation, their final formula for the FIC is as follows,

$$\text{FIC}_{\text{old}} = \max\{0, \hat{b}^2 - \hat{\kappa}/n\} + \hat{\tau}/n.$$

This expression is very similar to the one given in (32). The only difference is that the above expression lacks the $2\hat{b}\hat{c}_0/n$ term, which is present in (32). Because of this, FIC_{old} is easier to compute than the expression in (32), but, unless b or c is zero, the expected value of the quantity is $O(1/n)$ away from the true MSE of $\hat{\mu}_1$. Hence, the FIC of Jullum and Hjort (2017) will, in many cases, underemphasize the importance of the variance since the error of the estimated squared bias is on the same scale as this quantity. We therefore propose the following new form of the FIC, which in some sense can be seen as a corrected version of FIC_{old} ,

$$\text{FIC}_{\text{new}} = \max\{0, \hat{b}^2 - \hat{\kappa}/n + 2\hat{b}\hat{c}/n\} + \hat{\tau}/n.$$

As argued previously, this estimator is only $o(1/n)$ from the true MSE of $\hat{\mu}_1$. With this new version of the FIC, we need to estimate c_0 , as was discussed in Section 2. All other quantities can be approximated by the canonical estimators.

4.2 | How much does it matter?

Even though the the FIC from Jullum and Hjort (2017) is slightly off its mark, this version of the FIC has had success in many applications. See, for instance, Cunen, Walløe, and Hjort (2020), Claeskens et al. (2019), Wang and Hobæk Haff (2019), and Hermansen et al. (2015). It is therefore natural to ask “how wrong” this formula really is. In this section, we will present some examples illustrating the differences.

Firstly, notice that when $\hat{\mu}_0$ is unbiased, $c = 0$. When this is the case, the term distinguishing FIC_{old} from FIC_{new} is equal to zero and the two versions of the criterion agree. This happens when $\hat{\mu}_0$ is a mean or some bias corrected estimator. Another perhaps less trivial example is linear combinations of the regression coefficients used in linear regression. These estimators are unbiased for the true values, provided the model is correctly specified. Hence c_0 is zero in this case. Because of this, the use of the FIC in, for instance, Cunen, Walløe, and Hjort (2020) and Claeskens et al. (2019) is unproblematic.

Another situation where FIC_{old} and FIC_{new} give very similar results, is when either b or c are very small. If either $\hat{\mu}_1$ or $\hat{\mu}_0$ are almost unbiased for the true value μ_0 , this can happen, and the term distinguishing the two versions of the criterion is too small to have any real effect. In addition, we truncate negative estimates of the squared bias to zero. In some cases, the estimated squared bias will be less than zero regardless of whether we subtract $2\hat{b}\hat{c}_0/n$ or not. Hence, the two criteria can end up being very similar, if not equal, when the bias is small, relatively speaking. An example of such a situation, is when estimating the difference in median life lengths of Roman era Egyptian men and women using data from Pearson (1902). Jullum and Hjort (2017) analyze this data set in their article and use FIC_{old} to choose the model estimating the focus parameter with the highest precision. We repeated their computations in addition to calculating FIC_{new} for the same data. In this example, the two information criteria gave almost identical results. The reason for this, is that the bias estimate is truncated for all but one model. This happens for both FIC_{new} and FIC_{old} , and hence, the two criteria agree for all but the Gompertz model. For this one case, FIC_{new} was 0.022 larger than FIC_{old} . The Gompertz model is, however, by far the worst of the ones considered in Jullum and Hjort (2017) with very high estimated MSE of the focus parameter. Because of this, adding 0.022 to the result, does not change the conclusions we would draw from using the FIC in this case.

The above example might lead one to believe that the difference between the new and old FIC is negligible in all but a few situations. It might therefore be tempting to use FIC_{old} rather than FIC_{new} because of its relatively easier formulation. This, however, is not to be advised, as there in general is no bound on the size of term $2bc/n$. To see this, consider fitting an exponential model to i.i.d. data points $Y_1, \dots, Y_n \sim \text{Gamma}(k, \sqrt{k})$ for a fixed number k in order to estimate the SD in the distribution, which is 1 for all k . In this situation, the limiting bias in the model will grow with k , and as a result FIC_{old} will lose precision when k is large. A small simulation supports this claim. For a selection of k -values, we simulated one million data sets of size $n = 100$. In each case, we fitted an exponential model and computed FIC_{old} as well as FIC_{new} using the SD as focus parameter. We used the empirical SD as the consistent estimator $\hat{\mu}_0$. In Figure 1, we illustrate the

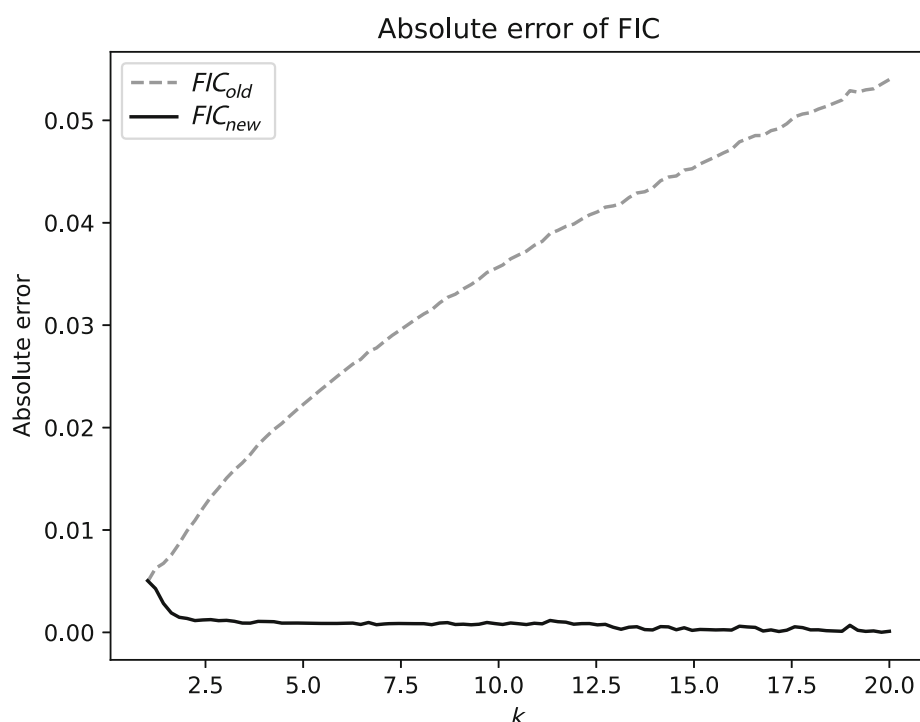


FIGURE 1 The figure displays the absolute value of the difference between the mean of 10^6 simulated FIC-values and the true mean squared error as functions of the shape parameter in the underlying Gamma($k, \sqrt{k}$)-distribution of the data. The error for the old version of the information criterion is shown by a grey dashed line, while the error of new version is plotted as a solid black line. The focus parameter is the SD in the distribution and the empirical SD is used as the consistent estimator $\hat{\mu}_0$. The sample size is $n = 100$.

results. We have plotted the estimated absolute value of the error of both FIC_{old} and FIC_{new} as functions of k , the shape parameter in the true underlying Gamma distribution. The figure shows clearly that not taking the term $2bc/n$ into account, causes the error of FIC_{new} to grow large as k increases.

This is of course only a toy example, but shows that there, at least in theory, is no bound on $2bc/n$ and that in certain situations ignoring its effect can be very problematic. We therefore recommend using FIC_{new} unless working exclusively with unbiased estimators. We will now illustrate this even further by trying out the new and old FIC on a dataset where the two criteria give very different results.

4.3 | Application—battle deaths in inter-state wars

In Pinker (2011) it is claimed that the world is gradually becoming more peaceful. Such statements have since been a heated topic of discussion. See for example, Cirillo and Taleb (2016a, 2016b), Clauset (2018), or Cunen, Hjort, and Nygård (2020). When investigating claims like that of Pinker statistically, proper modeling is important.

In this section, we will use the FIC to evaluate suitability of different models for checking the claim of Pinker based on the Correlates of War (CoW) dataset (Sarkees & Wayman, 2010). This dataset consists of the number of battle deaths in the 95 most recently concluded wars. The number of battle deaths is by no means a complete measure of violence. Hence, using the CoW data set to assess the claim of Pinker is of course somewhat lacking. Nevertheless, we use

these data to estimate the difference in median battle deaths between newer and older inter-state wars.

To compare recent wars to older ones, we need to decide which conflicts should be considered new. Cunen, Hjort, and Nygård (2020) investigate this question. Their analysis finds that the Korean War marks a change in the number of battle deaths in inter-state wars. Building on this, we define the “old” wars to be all conflicts before and including the Korean war. The remaining wars are treated as “new” conflicts.

We consider the data points as independent. Furthermore, we assume that the battle deaths in the recent wars are identically distributed and that the same holds true for the older wars. In our analysis, we consider eight different models and investigate how well they estimate the difference in median deaths between newer and older inter-state wars. For each case, we fit the corresponding model to the data sets consisting of newer and older wars separately. Afterwards, we use the fitted models to estimate the difference between median number of battle deaths in older and more recent wars. To evaluate the different models, we compute FIC_{new} for each case. As the consistent estimator, $\hat{\mu}_0$, we use the difference between the empirical median for the older and newer wars. We use the default method in R with `type` equal to 7 to estimate this quantity. Since the two datasets are independent, κ/n and τ/n are estimated by the sum of the corresponding quantities for each separate data set. To approximate c_0 and b , we take the difference between the corresponding estimates for older and more recent wars. Afterwards, the squared bias is estimated and truncated to zero in the cases where the approximation is negative. It is worth noting that the forms of κ , τ , b , and c_0 are relatively simple in this case where the focus parameter is the difference between two quantiles. For more complicated functions, for example, $\hat{\mu}_B/\hat{\mu}_A$ where $\hat{\mu}_A$ and $\hat{\mu}_B$ are the median number of battle deaths in newer and older wars, respectively, the delta method and the result of Section 2.5 need to be applied.

We worked with a slightly modified version of the data. In the CoW dataset only wars with more than 1001 battle deaths are registered. To avoid problems with observations on the bounds of the support, we replaced all observations of 1001 with 1001.01. The same trick is used in Cunen, Hjort, and Nygård (2020). See this article for more explanations and discussions. After transforming the data, we fitted eight parametric models. Firstly, Dagum, Log-logistic, Pareto IV, Lomax, log-normal and log-Cauchy distribution were fit to the data shifted by 1001 to the left. Secondly, we fitted log-Gamma and log-Weibull distributions to the data scaled by 1000. The resulting values of FIC_{new} for each model can be found in the plot in the left panel of Figure 2 together with the maximum likelihood estimate of the focus parameter for each model. From the figure, we notice that using the log-normal distributions results in the lowest MSE for the focus parameter. With this model, the estimated difference between median battle deaths before and after the Korean War is 2066.

To illustrate the difference between FIC_{new} and FIC_{old} , we also computed the criterion of Julum and Hjort (2017). The results are displayed in Figure 2. From the plot, we notice that the two methods indeed give very different estimates of the MSE for many of the models, though not all. In fact, using FIC_{new} rather than FIC_{old} actually leads to different models being chosen. With the model selected by FIC_{old} the difference between median number of battle deaths before and after the Korean War is estimated to be 3958, which is almost twice as large as the estimate we got from the model chosen by FIC_{new} . This shows that the additional term in FIC_{new} does have a real implication and is important for analysing the CoW dataset correctly.

The above analysis can be repeated for other quantiles than the median. In the right panel of Figure 2 we have computed and displayed the FIC values for and estimates of the difference in the third quartile in the two distributions. This can be used as a measure of how violence in the

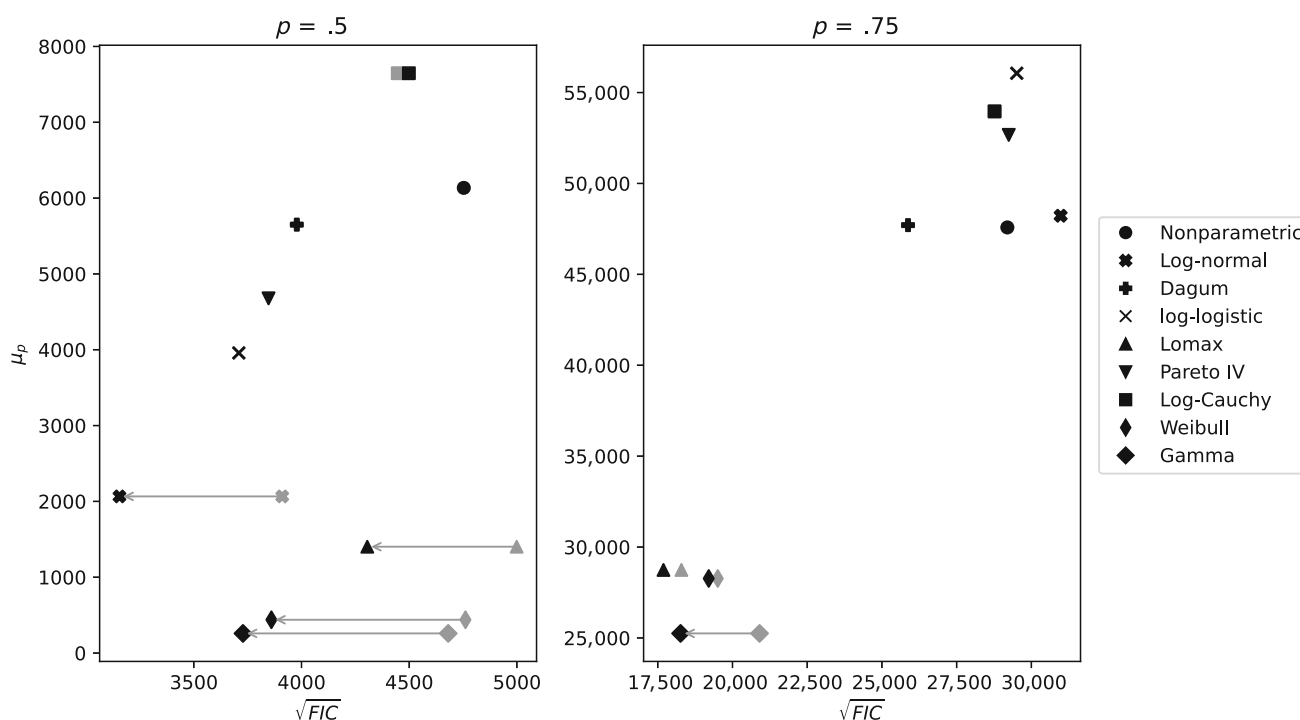


FIGURE 2 The figure shows FIC_{old} (colored grey) and FIC_{new} (colored black) for different models fitted to the correlates of war (CoW) dataset together with the models corresponding estimate of the focus parameter. The focus parameter is the difference in the p -quantile for older and newer wars. The value of p is shown in the title of each plot.

more deadly wars has changed in recent years. From the plot we notice that the two criteria give similar results in this case, with both the new and the old FIC selecting the Lomax model with an estimate of 28,740 for the focus parameter.

The plots in Figure 2 give some insight into how the new and old version of FIC differ. For the case of the median, c_0/n was estimated to be approximately 1227. This is a positive number, and hence the nonparametric estimate of the focus parameter is biased upward. Because of this, we would expect models underestimating the focus parameter compared to the nonparametric estimate, to be “less off” than what the naive estimate $\hat{b}$ suggests. Models estimating the focus parameter to be even larger than $\hat{\mu}_0$, on the other hand, are likely to be more biased than what $\hat{b}$ would lead one to believe. This intuitive idea is what is caught by FIC_{new} , but not FIC_{old} . If $\hat{b}$ and $\hat{c}_0$ have the same signs, $2\hat{b}\hat{c}/n > 0$ and the model is penalized by the criterion developed in this article, whereas when $\hat{\mu}_0$ is biased in the opposite direction of $\text{sign}(\hat{b})$, $2\hat{b}\hat{c}/n < 0$, resulting in a lower value of FIC_{new} than FIC_{old} unless the estimate of the squared bias is truncated to zero.

The above calculations show that there can be real and impactful differences between the old and new FIC. This is, however, not a guarantee that FIC_{new} performs better than FIC_{old} in this example. To investigate whether this is the case, we need to compute the true value of the focus parameter and compare the true MSEs with the corresponding FIC_{new} and FIC_{old} scores. In theory, this is a simple task, but in practice we do not know the true underlying random structure of battle deaths in inter-state wars. Because of this, we cannot check the performance of FIC_{new} and FIC_{old} directly. To get some indication on the quality of the MSE estimates we will therefore instead work with a distribution similar to our data set where the truth is known. To achieve this, we fitted a Lomax model to the data as explained previously. Afterwards, we simulated 10,000 datasets from this model. For each simulated dataset, we computed $\hat{\mu}_1$, FIC_{old} and FIC_{new} for each of the models considered in this section. We used the difference in median number of battle

deaths before and after the Korean war as the focus parameter. Afterwards, we used the 10,000 samples to estimate the expected value of FIC_{new} and FIC_{old} in addition to the true MSE for $\hat{\mu}_1$ for each model. For five out of the eight models considered, the average value of FIC_{new} was closer to the true MSE than FIC_{old} . For the remaining three cases FIC_{old} performed better than FIC_{new} . The two criteria were more or less equally variable, with $\text{Var}FIC_{\text{new}}/\text{Var}FIC_{\text{old}}$ ranging from 0.655 to 1.186 for the different models.

Comparing only the FIC-scores gives somewhat of a limited view. FIC_{old} and FIC_{new} differ only in their estimate of the squared bias. The estimated variance is identical in the two criteria. Because of this, an over- or underestimated variance might make it beneficial to under- or over-estimate the squared bias, respectively, as this “cancels” parts of the error made by the variance estimate. Although this might lead to more precise estimates of the MSE in certain cases, such cancellation happens mostly due to “luck” and is in general not something one should hope or aim for. We therefore believe it is more relevant to compare the quality of the estimates of the squared bias made by each criterion rather than the full MSE scores. Looking only at the estimated squared bias shows that the cancellation described above has happened in two of the cases where FIC_{new} performed better than FIC_{old} . We find that the estimate of the squared bias in the new criterion is indeed closer to the true value than the estimate provided by FIC_{old} in all but one model.

Choosing to fit a Lomax model and simulate from this is of course not the only way one can generate data similar to the number of battle deaths in the 95 most recent and concluded inter-state wars. We could fit any of the models considered in this section or even draw bootstrap samples from the dataset itself. We have tried out a number of these methods, and similar results were found in all cases. For the sake of clarity and brevity, we will therefore not go into depth about all settings, but taking the average over all simulation settings gives the following result. More often than not, FIC_{new} performed better than FIC_{old} when it comes to estimation of the MSE. Furthermore, the squared bias estimate provided by FIC_{new} were on average more precise than the one from FIC_{old} for a little more than five out of eight models. In addition, the average value of $\text{Var}FIC_{\text{new}}/\text{Var}FIC_{\text{old}}$ ranged from 0.457 to 1.025 between the different models.

5 | CONCLUDING REMARKS

In this article, we have only considered the framework of i.i.d. data, but extensions to more complicated situations are certainly possible. For regression and classification settings with responses Y_i and predictors X_i for $i = 1, \dots, n$, the formulas derived in this article are directly applicable when the covariates are considered random. This is because the tuples $Z_i = (Y_i, X_i)$ for $i = 1, \dots, n$ can be considered as i.i.d. data points in this case. In addition, results like the Lindeberg–Feller theorem (see e.g., Billingsley, 1995, p. 359) allows most of the formulas derived in this article to be generalized to situations where the covariates are considered nonstochastic. It is, however, worth noting that in regression and classification settings, nonparametric estimators satisfying (29) can be hard to come by. This is especially true when the number of covariates is high. In such cases, a solution can be to replace the true parameter, μ_0 , by the least false parameter in a widest model. This is in some sense the closest we can get to the true value, and the FIC should therefore measure the MSE of estimators relative to this quantity. Claeskens et al. (2019) and Cunen, Walløe, and Hjort (2020) take this approach for the simpler FIC of Jullum and Hjort (2017). Modifications along the same lines can be made for the FIC derived in this article as well, but to use the more precise information criterion developed in this article, c_0 needs to be estimated. This can be achieved

using Theorem 3. Further extensions of the iid framework include dependent and censored data. Arguments similar to those given in Hermansen et al. (2015) and Jullum and Hjort (2019), respectively, can be used to generalize the results of this paper to these more complicated data structures.

FIC is a popular research topic and there exists many extensions. Most of which can also be made for the criterion developed in this article as well. We will mention two such extensions, but many other possibilities exist. For an overview of techniques, consult Claeskens and Hjort (2008).

In certain situations, multiple focus parameters are of interest and should be taken into account when performing model selection. One way of achieving this for the FIC is to use average FIC (AFIC). Assume the focus parameters can be written as $\mu(u)$ for u in some index set U . A risk function taking all parameters into account is the following expression,

$$E \left(\int_U [\hat{\mu}(u) - \mu(u)]^2 dW(u) \right) = \int_U \text{bias}[\hat{\mu}(u)]^2 dW(u) + \int_U \text{Var} \hat{\mu}(u) dW(u),$$

where W is some distribution over U indicating how much each focus parameter matters relative to the others. The risk of the estimators $\hat{\mu}(u)$ for $u \in U$ can therefore be estimated by the following expression

$$\text{AFIC} = \max \left\{ 0, \int_U \left(\hat{b}(u)^2 - \frac{\hat{\kappa}(u)}{n} + \frac{2\hat{b}(u)\hat{c}_0(u)}{n} \right) dW(u) \right\} + \frac{1}{n} \int_U \hat{\tau}(u) dW(u).$$

Here $\hat{b}(u)$, $\hat{c}_0(u)$, $\hat{\kappa}(u)$, and $\hat{\tau}(u)$ are the estimators discussed previously in this article computed for the estimator $\hat{\mu}(u)$. Notice that AFIC does not equal the integral of the individual FIC scores as we truncate possibly negative estimates of the integrated squared bias rather than the individual bias estimates for each $u \in U$ to zero. For more information about and variants of AFIC, see Claeskens and Hjort (2008, pp. 179–183).

FIC scores can be used for more than strictly selecting the optimal model for estimating the focus parameter. An alternative approach, is to use model averaging. Assume we have p models with estimates $\hat{\mu}_1, \dots, \hat{\mu}_p$ of the focus parameter. Rather than using the FIC to choose the estimator with the lowest MSE, we could use the weighted average $\hat{\mu}_{\text{avg}} = \sum_{j=1}^p w_j \hat{\mu}_j$, where w_j are weights summing to one. The weights should be related to the FIC scores such that $\hat{\mu}_j$ s with low MSE are given more weight than $\hat{\mu}_j$ s for which $\text{FIC}(\hat{\mu}_j)$ is large. Inspired by Eq. 6.1 in Jullum and Hjort (2017), we suggest the following form for the weights

$$w_j = \frac{\exp[-\lambda \text{FIC}(\hat{\mu}_j)]}{\sum_{k=1}^p \exp[-\lambda \text{FIC}(\hat{\mu}_k)]},$$

for some tuning parameter λ .

ACKNOWLEDGMENTS

This research is funded in part by The Norwegian Research Council 237718 through the Big Insight Centre for research-driven innovation. The authors are grateful for support from Centre for Advanced Research, Oslo, via the Stability and Change project, led by Hjort. We appreciate the anonymous reviewers for their careful reading, constructive comments, and suggestions, which contributed to improving our the manuscript.

ORCID

Ingrid Dæhlen  <https://orcid.org/0000-0002-2029-4451>

REFERENCES

- Billingsley, P. (1995). *Probability and measure* (3rd ed.). Wiley.
- Billingsley, P. (1999). *Convergence of probability measures*. John Wiley & Sons.
- Cirillo, P., & Taleb, N. N. (2016a). *The decline of violent conflicts: What do the data really say?* [Conference presentation]. The Nobel Foundation Symposium 161: The Causes of Peace, 1–26.
- Cirillo, P., & Taleb, N. N. (2016b). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications*, 452, 29–45.
- Claeskens, G., Croux, C., & Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, 62, 972–979.
- Claeskens, G., Cunen, C., & Hjort, N. L. (2019). Model selection via focused information criteria for complex data in ecology and evolution. *Frontiers in Ecology and Evolution*, 7, 415.
- Claeskens, G., & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98, 900–916.
- Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Books.
- Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science. Advances*, 4, eaao3580.
- Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society Series B (Methodological)*, 30, 248–275.
- Cunen, C., Hjort, N. L., & Nygård, H. M. (2020). Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research*, 57, 221–234.
- Cunen, C., Walløe, L., & Hjort, N. L. (2020). Focused model selection for linear mixed models with an application to whale ecology. *The Annals of Applied Statistics*, 14, 872–904.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1–26.
- Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics.
- Efron, B., & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24, 2431–2461.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27–38.
- Hermansen, G. H., Hjort, N. L., & Jullum, M. (2015). *Parametric or nonparametric: The FIC approach for stationary time series* [Conference presentation]. Proceedings of the 60th World Statistics Congress of the International Statistical Institute, 4827–4832.
- Hjort, N. L., & Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 23, 882–904.
- Hjort, N. L., & Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24, 1619–1647.
- Huber, P. J. (2009). *Robust statistics* (2nd ed.). Wiley.
- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94, 401–419.
- Jullum, M., & Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, 27, 951–981.
- Jullum, M., & Hjort, N. L. (2019). What price semiparametric Cox regression? *Lifetime Data Analysis*, 25, 406–438.
- Kim, K. I. (2016). Higher order bias correcting moment equation for M-estimation and its higher order efficiency. *Econometrics*, 4, 48.
- Ko, V., Hjort, N. L., & Hobæk Haff, I. (2019). Focused information criteria for copulas. *Scandinavian Journal of Statistics*, 46, 1117–1140.
- Lindstrøm, T. L. (2017). *Spaces: An introduction to real analysis* (Vol. 29). American Mathematical Society.
- Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics*, 24, 1602–1618.
- Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika*, 1, 261–264.
- Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Viking.

- Quenouille, M. H. (1949). *Approximate tests of correlation in time-series 3*. In *Mathematical proceedings of the Cambridge philosophical society* (Vol. 45, pp. 483–484). Cambridge University Press.
- Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353–360.
- Sarkees, M. R., & Wayman, F. (2010). *Resort to war: 1816 - 2007*. CQ Press.
- Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica*, 1, 185–202.
- Shao, J. (2003). *Mathematical statistics* (2nd ed.). Springer Science & Business Media.
- Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. Springer Science & Business Media.
- Shao, J., & Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, 17, 1176–1197.
- Singh, R. (1979). Mean squared errors of estimates of a density and its derivatives. *Biometrika*, 66, 177–180.
- Tukey, J. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29, 614.
- Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge university press.
- Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Chapman & Hall/CRC.
- Wang, Y., & Hobæk Haff, I. (2019). Focussed selection of the claim severity distribution. *Scandinavian Actuarial Journal*, 2019, 129–142.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.
- Zhang, X., & Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39, 174–200.

How to cite this article: Dæhlen, I., Hjort, N. L., & Hobæk Haff, I. (2024). Accurate bias estimation with applications to focused model selection. *Scandinavian Journal of Statistics*, 51(2), 724–759. <https://doi.org/10.1111/sjos.12696>

APPENDIX A. KERNEL DENSITY ESTIMATION

In Section 2.2, we derived an estimator for c when $\hat{\mu}$ is a quantile. To use the formula in (10), however, we need consistent estimators of both the density f and its derivative f' at μ_p .

The classic kernel density estimators are

$$\hat{f}(x) = \frac{1}{nh_0} \sum_{i=1}^n K\left(\frac{x_i - x}{h_0}\right) \quad \text{and} \quad \hat{f}'(x) = -\frac{1}{nh_1^2} \sum_{i=1}^n K'\left(\frac{x_i - x}{h_1}\right). \quad (\text{A1})$$

In the above, h_1 and h_2 are called the bandwidths and K is the kernel. In principle, the kernel can be any nonnegative and differentiable function, but to simplify calculations we will assume it is symmetric. The obvious choice for such a kernel is the standard normal density, ϕ .

The standard strategy is to minimize the asymptotic integrated MSE. Minimizing the L^2 distance between $\hat{f}$ and $\hat{f}'$ and the corresponding true values is reasonable when a good overall fit is desired. In our situation, however, we are only interested in using the kernel density estimates to approximate $f(\mu_p)$ and $f'(\mu_p)$. Hence, we should choose bandwidths minimizing the MSE of these estimators and not the full integrated MSE. Arguing as in Singh (1979), one can show that this is achieved by

$$h_1 = \left(\frac{R(K)f(\mu_p)}{S(K)^2 f^{(2)}(\mu_p)^2} \right)^{1/5} n^{-1/5} \quad \text{and} \quad h_2 = \left(\frac{3R(K')f(\mu_p)}{S(K)^2 f^{(3)}(\mu_p)^2} \right)^{1/7} n^{-1/7},$$

where $f^{(k)}$ denote the k th derivative of f and

$$R(H) = \int_{\mathbb{R}} H(u)^2 du \quad \text{and} \quad S(H) = \int_{\mathbb{R}} u^2 H(u) du.$$

See Chap. 2 of Wand and Jones (1995) for more details.

To estimate the optimal bandwidths, we need to know the value of $f(\mu_p)$, $f^{(2)}(\mu_p)$ and $f^{(3)}(\mu_p)$. These quantities are, however, unknown by definition of the problem. To estimate them, we therefore suggest fitting parametric models to the data. This can lead to inconsistent estimators of $f(\mu_p)$, $f^{(2)}(\mu_p)$ and $f^{(3)}(\mu_p)$, but this will only result in bandwidths being slightly off from the optimal ones. The estimators in (A1) will still converge in probability to the true values, $f(\mu_p)$ and $f'(\mu_p)$ when inconsistent estimators of $f(\mu_p)$, $f^{(2)}(\mu_p)$, and $f^{(3)}(\mu_p)$ are used to find optimal bandwidths.

We recommend the following rule of thumb: If the distribution has a heavy tail, fit a Pareto distribution to the data. Otherwise, fit a normal distribution. These distributions are chosen mainly because their maximum likelihood estimators have closed form expressions, allowing $f(\mu_p)$, $f^{(2)}(\mu_p)$, and $f^{(3)}(\mu_p)$ to be estimated without numerical optimization.

Combining all of the above, leads to the following rule of thumb for estimating $f(\mu_p)$ and $f'(\mu_p)$:

$$\hat{f}(x) = \frac{1}{nh_0} \sum_{i=1}^n \phi\left(\frac{x_i - x}{h_0}\right) \quad \text{and} \quad \hat{f}'(x) = \frac{1}{nh_1^2} \sum_{i=1}^n \phi'\left(\frac{x_i - x}{h_1}\right).$$

where ϕ is the standard normal density and

$$\begin{aligned} h_0 &= \hat{\mu}_p \left[2\sqrt{\pi}(Y_{(1)}/\mu_p)^{\hat{\alpha}} \hat{\alpha}(\hat{\alpha}+1)^2(\hat{\alpha}+2)^2 n \right]^{-1/5} \quad \text{and} \quad h_1 \\ &= \hat{\mu}_p \left[(4/3)\sqrt{\pi}(Y_{(1)}/\mu_p)^{\hat{\alpha}} \hat{\alpha}(\hat{\alpha}+1)^2(\hat{\alpha}+2)^2(\hat{\alpha}+3)^2 n \right]^{-1/7}, \end{aligned}$$

when the distribution of Y is heavy tailed and

$$\begin{aligned} h_0 &= \sigma^2 \exp \left[(\hat{\mu}_p - \bar{Y})^2 / 10\hat{\sigma}^2 \right] \left\{ \sqrt{2}\hat{\sigma} \left[(\hat{\mu}_p - \bar{Y})^2 - \hat{\sigma}^2 \right]^2 n \right\}^{-1/5} \quad \text{and} \\ h_1 &= \sigma^2 \exp \left[(\hat{\mu}_p - \bar{Y})^2 / 14\hat{\sigma}^2 \right] \left\{ (2\sqrt{2}/3)\hat{\sigma}(\hat{\mu}_p - \bar{Y})^2 \left[(\hat{\mu}_p - \bar{Y})^2 - 3\hat{\sigma}^2 \right]^2 n \right\}^{-1/7}, \end{aligned}$$

otherwise. In the above, $\hat{\sigma}$, $\bar{Y}$, and $Y_{(1)}$ are the empirical SD, mean, and minimum value in the sample, respectively. In addition, $\hat{\mu}_p$ is the empirical quantile and

$$\hat{\alpha} = \left[\frac{1}{n} \sum_{i=1}^n \log(Y_i/Y_{(1)}) \right]^{-1}.$$

Lastly, we would like to remark that there exists more sophisticated methods for estimating $f(\mu_p)$ and $f'(\mu_p)$ than the ones presented here. For instance, rather than using a purely non-parametric or parametric estimate, a semi-parametric approach is possible. Parametric and nonparametric models can be combined as in, for example, Hjort and Glad (1995) and Efron and Tibshirani (1996). Another approach is to use local likelihoods, see Hjort and Jones (1996) and Loader (1996).

APPENDIX B. COMPLETE PROOFS

We will use the following notation. If A is a matrix $A_{j,k}$ denotes the (j, k) th component of A . Furthermore A^j is notation for the j th column vector of A . Similarly, a_j denotes the j th component of a if a is a vector. Lastly we will use $\nabla_{x_0} f(x)$ and $H_{x_0} f(x)$ as the gradient and hessian matrix of f at x_0 , respectively.

B.1 Uniform integrability

To give full proofs and sets of conditions for the theorems in the article, we will need to show that multiple quantities are uniformly integrable. We give the definition here, for the sake of completion.

Definition 1 (Uniformly integrable, Billingsley (1999)). A sequence of random variables $(X_n)_n \subseteq \mathbb{R}$ is uniformly integrable if the following holds,

$$\lim_{K \rightarrow \infty} \sup_n E(|X_n| : |X_n| \geq K) = 0.$$

We also state a lemma containing multiple results concerning uniform integrability. These results are not new and proofs can be found in most textbooks on the subject. Properties we will use are nevertheless included here for easier reference in the arguments that follow.

Lemma 2. *In the following, let $(X_n)_n, (Y_n)_n \subseteq \mathbb{R}$ be two sequences of random variables. Furthermore, let $A_n \in \mathbb{R}^{p \times q}$ and $B_n \in \mathbb{R}^{q \times p'}$ be random matrices and $C \in \mathbb{R}^{q \times p''}$ a fixed matrix.*

- (i) X_n is uniformly integrable if and only if $|X_n|$ is uniformly integrable.
- (ii) Let $a, b \in \mathbb{R}$ and assume that X_n and Y_n are uniformly integrable, then $aX_n + bY_n$ is uniformly integrable
- (iii) Let X_n^2 and Y_n^2 be uniformly integrable, then $X_n Y_n$ is uniformly integrable.
- (iv) If X_n^p is uniformly integrable, X_n^q is uniformly integrable for all $1 \leq q < p$.
- (v) If X_n converges in distribution to X and $E|X_n|$ converges to $E|X|$, X_n is uniformly integrable.
- (vi) If each component of A_n is uniformly integrable, each component of CA_n is uniformly integrable as well.
- (vii) If the square of each component of A_n and B_n both are uniformly square integrable, each component of $A_n B_n$ is uniformly integrable.

Proof. (i) holds by definition. (ii) follows from the triangle inequality and (iii) from Hölder's inequality. Since EX_n^p converges to EX^p by uniform integrability of X_n^p , $\sup_n E(|X_n|^{q+\epsilon}) < \infty$ where $\epsilon = (p - q)/q > 0$. By (3.18) in Billingsley (1999, p. 31), this implies that X_n^q is uniformly integrable, proving (iv). Case (v) is a consequence of Thm. 3.6 in Billingsley (1999, p. 32), the continuous mapping theorem and case (i). The two remaining points follow from (ii) and (iii) respectively. ■

B.2 Functions of unbiased estimators

We start with the simplest case: functions of unbiased estimators.

Theorem 9 (Functions of unbiased estimators). *Let $\hat{a} \in \mathbb{R}^p$ be an estimator of a_0 such that $\sqrt{n}(\hat{a} - a_0)$ converges in distribution to some U with finite second moment. Assume that all components of $\sqrt{n}(\hat{a} - a_0)$ to the power of three are uniformly integrable and that there exists a neighborhood $\mathcal{N}$ of a_0 on which $h : \mathbb{R}^p \rightarrow \mathbb{R}$ has continuous partial derivatives up to order 3. Then*

$$\lim_{n \rightarrow \infty} n E[h(\hat{a}) - h(a_0)] = \frac{1}{2} \text{Tr}[Hh(a_0)EUU^T].$$

Proof. Since all components of $\sqrt{n}(\hat{a} - a_0)$ are uniformly square integrable, case (vii) of lemma 2 ensures that all components of $En(\hat{a} - a_0)(\hat{a} - a_0)^T$ are uniformly integrable as well. Hence, $En(\hat{a} - a_0)(\hat{a} - a_0)^T$ converges to EUU^T and $nE(\hat{a} - a_0)^T Hh(a_0)(\hat{a} - a_0) = Hh(a_0)EUU^T + o(1)$. To show the theorem, we therefore only need to prove that, $nE\epsilon_n(\hat{a}) = o(1)$, where $\epsilon_n(\hat{a})$ is the remainder term in (6).

Choose a compact subset K of $\mathcal{N}$ with a_0 in its interior. Since $\hat{a}$ converges in probability to a_0 , it lies in the interior of K with probability tending to one. Without loss of generality, we will therefore assume that $\hat{a}$ lies in the interior of K . Since all third-order partial derivatives of h are continuous on this set, the extreme value theorem guarantees that they are bounded by some $M \in \mathbb{R}$. Hence, by standard results for Taylor expansions of multivariate functions,

$$|\epsilon_n(\hat{a})| \leq \frac{M}{3!} \|\hat{a} - a_0\|_1^3.$$

See for example, Coroll. 6.5.8 in Lindstrøm (2017). In the above, $\|\cdot\|_1$ denotes the L^1 norm on $\mathbb{R}^p$. Since all components of $\sqrt{n}(\hat{a} - a_0)$ to the power of three are uniformly integrable, and $n^{3/2}\|\hat{a} - a_0\|_1^3$ is $O_{\text{Pr}}(1)$ by the continuous mapping theorem, we have $E|\epsilon_n(\hat{a})| = O(n^{-3/2})$. ■

For the special case of empirical means, the theorem simplifies somewhat.

Corollary 1 (Functions of means). *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be random variables with empirical mean $\hat{\mu}$ and expected value μ_0 . Assume further that the expected value of all components of Y_i to the power of four exists, and that h is a function with continuous partial derivatives up to order three in a neighborhood of μ_0 . Then*

$$\lim_{n \rightarrow \infty} n E[h(\hat{\mu}) - h(\mu_0)] = \frac{1}{2} \text{Tr}[Hh(\mu_0)\text{Var}Y_i].$$

Proof. By the central limit theorem $\sqrt{n}(\hat{\mu} - \mu_0)$ converges to a normal distribution with variance $\text{Var}Y_i$. Because of this, the result follows from Theorem 9 provided all components of $\sqrt{n}(\hat{\mu} - \mu_0)$ to the power of three are uniformly integrable. By case (iv) of lemma 2, this holds true when all components of $\sqrt{n}(\hat{\mu} - \mu_0)$ to the power of four are uniformly integrable. We will show this latter statement.

Fix j . By the continuous mapping theorem, $n^2[(\hat{\mu})_j - (\mu_0)_j]^4$ converges in distribution to U^4 where $U \sim N[0, (\text{Var}Y_i)_{jj}]$. Furthermore, utilizing that $Y_1, \dots, Y_n$ are independent and have mean zero one can show

$$\begin{aligned} & \mathbb{E} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n [(Y_i)_j - (\mu_0)_j] \right)^4 = \\ & \frac{1}{n^2} \sum_{i=1}^n \{ \mathbb{E}[(Y_i)_j - (\mu_0)_j]^4 - 3(\text{Var} Y_i)_{jj}^2 \} + 3(\text{Var} Y_i)_{jj}^4 \rightarrow 3(\text{Var} Y_i)_{jj}^2, \end{aligned}$$

which is equal to the fourth moment of the j th component of U . Hence, $n^2[(\hat{\mu})_j - (\mu_0)_j]^4$ is uniformly integrable and Theorem 9 can be applied. ■

B.3 Quantiles

The second case is quantiles.

Theorem 10 (Quantiles). *Let $Y_1, \dots, Y_n$ be i.i.d. with distribution F and let μ_p be the true p -quantile in the distribution. Assume $|\sqrt{n}(\hat{\mu}_p - \mu_p)|^3$ is uniformly integrable and that there exists a neighbourhood $\mathcal{N}$ of μ_p on which $F : \mathbb{R}^p \rightarrow \mathbb{R}$ has continuous partial derivatives up to order 3, then*

$$n\mathbb{E}(\hat{\mu}_p - \mu_p) = \frac{n}{f(\mu_p)} \left(\frac{j + \gamma}{n + 1} - p \right) - \frac{f'(\mu_p)p(1-p)}{2f(\mu_p)^3} + O(n^{-3/2}),$$

where f is the density function in the distribution F and $\hat{\mu}_p$ is the empirical quantile defined by $\hat{\mu}_p = (1 - \gamma)Y_{(j)} + \gamma Y_{(j+1)}$ with j and γ defined by either of the rows in Table 1.

Proof. The only thing we need to prove is that $\mathbb{E}\delta_n = o(1/n)$, where δ_n is defined in (14) of the article. The rest of the argument is similar to those given in the article. Since δ_n is a sum of remainder terms, it suffices to show that the expected value of each of them are $o(1/n)$. We will focus on δ'_n , defined in Equation (13) of the article. The other case is similar.

By standard results for quantiles, $\hat{\mu}_p$ is consistent for μ_p and $\sqrt{n}(\hat{\mu}_p - \mu_p)$ converges in distribution to a $N[0, p(1-p)/f(\mu_p)]$ distribution. This allows us to argue as in the proof of Theorem 9, to show

$$n^{3/2}\mathbb{E}\delta'_n \leq \frac{M}{3!}\mathbb{E}|\sqrt{n}(\hat{\mu}_p - \mu_p)|^3.$$

Since $|\sqrt{n}(\hat{\mu}_p - \mu_p)|^3$ is uniformly integrable by assumption and $|\sqrt{n}(\hat{\mu}_p - \mu_p)|^3 = O(1)$ by the continuous mapping theorem, the above equation ensures $\mathbb{E}\delta'_n = O(n^{-3/2})$. ■

B.4 Maximum likelihood estimators

We will now prove Theorem 3 of the article. We start by showing a lemma.

Lemma 3. *Let $Y_1, \dots, Y_n$ be i.i.d. data points from a distribution F and $\mathcal{F} = \{\log f_\theta \mid \theta \in \Theta\}$ a family of densities indexed by an open set Θ . Let θ_{lf} denote the minimizer of the Kullback–Leibler divergence from F to $\mathcal{F}$ and $\hat{\theta}$ the maximum likelihood estimator. Furthermore, let*

$$J_n = -\frac{1}{n}H_{\theta_{\text{lf}}}\ell_n(\theta) \quad \text{and} \quad J = -H_{\theta_{\text{lf}}}\mathbb{E}\log f_\theta(Y),$$

and

$$W_n = \frac{1}{n} \frac{\partial^3 \ell_n}{\partial \theta \partial \theta^T \partial \theta^j}(\theta_{\text{lf}}) \quad \text{and} \quad W^j = \frac{\partial^3}{\partial \theta \partial \theta^T \partial \theta^j} \Big|_{\theta_{\text{lf}}} \mathbb{E} \log f_\theta(Y),$$

where ℓ_n denotes the likelihood function based on $Y_1, \dots, Y_n$ and $Y \sim F$. Then the following quantities are uniformly integrable:

- (1) $n^2(\hat{\theta}_j - (\theta_{\text{lf}})_j)^4$ for each $j = 1, \dots, p$
- (2) $n[(J_n)_{j,k} - J_{j,k}]^2$ for each $j, k = 1, \dots, p$
- (3) $n[(W_n)_{j,k} - W_{j,k}]^2$ for each $j, k = 1, \dots, p$,

provided the following assumptions hold true:

- (A1) All partial derivatives up to order 4 of $\log f_\theta(y)$ at θ_{lf} exists and are continuous for almost all y . In addition, all fourth moments of $\nabla_{\theta_{\text{lf}}} \log f_\theta(Y)$ and all second moments of all second- and third-order partial derivatives of $\log f_\theta(Y)$ at θ_{lf} exists and are finite.
- (A2) The matrix J is positive definite.
- (A3) All fourth-order powers of $\sqrt{n}[\hat{\theta} - \theta_{\text{lf}} - (nJ)^{-1} \nabla \ell_n(\theta_{\text{lf}})]$ are uniformly integrable.

Proof. We start with (1). Fix j and introduce the two variables

$$N_n = n^{-1/2} (J^{-1})^j \nabla \ell_n(\theta_{\text{lf}}) \quad \text{and} \quad \delta_n = \sqrt{n} \left\{ \hat{\theta}_j - (\theta_{\text{lf}})_j - [(nJ)^{-1}]^j \nabla \ell_n(\theta_{\text{lf}}) \right\}.$$

Then,

$$\left[\sqrt{n}(\hat{\theta} - \theta_{\text{lf}}) \right]^4 = (N_n + \delta_n)^4 = N_n^4 + \delta_n^4 + 4N_n\delta_n^3 + 4N_n^3\delta_n + 6N_n^2\delta_n^2.$$

We will show that each term is uniformly integrable. By case (ii) of Lemma 2, this ensures that $n^2(\hat{\theta}_j - (\theta_{\text{lf}})_j)^4$ is uniformly integrable.

By the central limit theorem, N_n converges in distribution to a random variable $N \sim N(0, \sigma^2)$ where

$$\sigma = [(J^{-1})^j]^T \text{Var}[\nabla_{\theta_{\text{lf}}} \log f_\theta(Y_i)] (J^{-1})^j.$$

Furthermore, elementary computations show

$$\mathbb{E} N_n^4 = \frac{1}{n} (\mathbb{E} \{ [(J^{-1})^j]^T \nabla_{\theta_{\text{lf}}} \log f_\theta(Y_i) \}^4 - 3\sigma^4) + 3\sigma^4 \rightarrow 3\sigma^4 = \mathbb{E} N^4,$$

since all fourth-order moments of $\nabla_{\theta_{\text{lf}}} \log f_\theta(Y_i)$ exist and are finite. Hence, by cases (v) and (iv) of Lemma 2, N_n^p is uniformly integrable for $p \leq 4$.

By assumption and case (iv) of Lemma 2, δ_n^p is uniformly integrable for $p \leq 4$. The remaining terms now follow directly. Since δ_n^2 and N_n^2 are uniformly integrable, case (iii) of Lemma 2 ensures that $N_n \delta_n$ is uniformly integrable. This, together with uniform integrability of N_n^4 and δ_n^4 and yet another application of case (iii) in Lemma 2, shows that $N_n \delta_n^3$ and $N_n^3 \delta_n$ are uniformly integrable.

We will now prove (2). The argument for (3) is similar and hence omitted. Fix j and k . By the law of large numbers $n[(J_n)_{j,k} - J_{j,k}]^2$ converges almost surely to $\text{Var}H_{\theta_{\text{lf}}} \log f_{\theta}(Y_i)_{j,k}$. Furthermore, the expected value of $n[(J_n)_{j,k} - J_{j,k}]^2$ is equal to $\text{Var}H_{\theta_{\text{lf}}} \log f_{\theta}(Y_i)_{j,k}$. Hence, by case (v) of Lemma 2, $n[(J_n)_{j,k} - J_{j,k}]^2$ is uniformly integrable. ■

We are now ready to prove Theorem 3 of the article.

Theorem 11. *Let all quantities be defined as in Lemma 2 and assume that (A1–A3) hold true. Then the bias of $\hat{\theta} - \theta_{\text{lf}}$ is on the form given in Theorem 3 of the article, provided the following additional assumption:*

- *There exists a neighborhood, $\mathcal{N}$ of θ_{lf} and an integrable function m such that all fourth-order partial derivatives are bounded by $m(y)$ and $\text{Em}(Y)^4$ exists and is finite.*

Proof. By the mean value theorem, the following holds for all $i, j, k = 1, \dots, p$,

$$\left\| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \log f_{\theta}(y) - \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \Big|_{\theta_{\text{lf}}} \log f_{\theta}(y) \right\| \leq \left\| \frac{\partial^4}{\partial \theta \partial \theta_i \partial \theta_j \partial \theta_k} \Big|_{\theta^*} \log f_{\theta}(y) (\theta - \theta_{\text{lf}}) \right\|,$$

for some θ^* on the line segment between θ and θ_{lf} . Hence, if we assume without loss of generality that $\mathcal{N}$ is convex, the existence of m as described in (A4), ensures that for all $\theta \in \mathcal{N}$,

$$\left\| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \log f_{\theta}(y) - \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \Big|_{\theta_{\text{lf}}} \log f_{\theta}(y) \right\| \leq C m(y) \|\theta - \theta_{\text{lf}}\|,$$

for some $C > 0$. If we further, without loss of generality, assume that $\mathcal{N}$ is compact we can use the above equation and the triangle inequality to show

$$\left\| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \log f_{\theta}(y) \right\| = \left\| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \Big|_{\theta_0} \log f_{\theta}(y) \right\| + C' m(y),$$

for some positive $C' \in \mathbb{R}$. The right-hand side does not depend on θ and has finite expectation, hence all third-order partial derivatives of $\log f_{\theta}(y)$ are bounded by a fixed integrable function. This together with condition (A1) and (A2) of Lemma 3, ensures that $\hat{\theta}$ exists with probability tending to 1, $\hat{\theta}$ converges in probability to θ_{lf} and that

$$\hat{\theta} = \theta_{\text{lf}} + (nJ)^{-1} \ell_n(\theta_{\text{lf}}) + o_{\text{Pr}}(1). \quad (\text{B1})$$

This follows from Thms. 5.41 and 5.42 in Van der Vaart (1998). Since $\hat{\theta}$ converges to θ_{lf} in probability, the probability of $\hat{\theta}$ lying in $\mathcal{N}$ goes to one. In the following, we will therefore assume that $\hat{\theta} \in \mathcal{N}$ without loss of generality.

To show Theorem 3 in the article we need to prove that the expected value of the remainder term $\epsilon_n(\hat{\theta})$ is of order $O(1/\sqrt{n})$, and that each component of $W_n^j n(\hat{\theta} - \theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}})^T$ and $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\sqrt{n}(J_n^j - J^j)^T$ is uniformly integrable. The rest of the argument is similar to that given in the article.

We start with the remainder term $\epsilon_n(\hat{\theta})$. Since m bounds all fourth order partial derivatives of $\log f_{\theta}(y)$ in $\mathcal{N}$ for almost all y , standard results concerning remainder terms in Taylor expansions can be used to show

$$|\epsilon_n(\hat{\theta})| \leq \frac{1}{3!} \sum_{i=1}^n m(Y_i) \|\hat{\theta} - \theta_{\text{lf}}\|_1^3.$$

Taking expectations on both sides of the inequality shows,

$$\mathbb{E} \sqrt{n} |\epsilon_n(\hat{\theta})| \leq \frac{1}{6} \mathbb{E} \left[m(Y) \|\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\|_1^3 \right].$$

Furthermore, applying Hölder's inequality twice, reveals

$$\mathbb{E} \sqrt{n} |\epsilon_n(\hat{\theta})| \leq \frac{1}{6} (\mathbb{E} m_2(Y)^4)^{1/4} \left(\mathbb{E} \|\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\|_1^4 \right)^{3/4}.$$

By (B1) $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}}) = O(1)$. Applying the continuous mapping theorem, ensures $n^2 \|\hat{\theta} - \theta_{\text{lf}}\|_1^4 = O(1)$. Furthermore, uniform integrability of $n^2 \|\hat{\theta} - \theta_{\text{lf}}\|_1^4$ can be shown using case (1) of lemma 3 and case (ii) and (iii) of Lemma 2. Hence, $\mathbb{E} \|\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})\|_1^4 = O(1)$, showing that $\mathbb{E} \epsilon_n(\hat{\theta}) = O(1/\sqrt{n})$.

Uniform integrability of each component of $W_n^j n(\hat{\theta} - \theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}})^T$ holds true as long as the square of each component of W_n^j is uniformly integrable and each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ to the power of four is uniformly integrable. This follows from case (vii) of Lemma 2. Each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ to the power of four is uniformly integrable by (1) in Lemma 3. For W_n^j , notice that the square of each component of W_n^j is uniformly integrable by (3) in Lemma 3. Hence, $W_n^j n(\hat{\theta} - \theta_{\text{lf}})(\hat{\theta} - \theta_{\text{lf}})^T$ is uniformly integrable.

For the last case, notice that each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}}) \sqrt{n}(J_n^j - J^j)^T$ is uniformly integrable by case (vii) in Lemma 2 if the square of each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ and $\sqrt{n}(J_n^j - J^j)$ is uniformly integrable. By case (1) of Lemma 3 and case (iv) of Lemma 2, the square of each component of $\sqrt{n}(\hat{\theta} - \theta_{\text{lf}})$ is uniformly integrable, while the square of each component of $\sqrt{n}(J_n^j - J^j)$ is uniformly integrable by case (2) of Lemma 3. ■

To show consistency of the estimator of c defined in Theorem 3, we first state and prove a lemma.

Lemma 4. Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F , $\hat{\theta} \in \mathbb{R}^p$ a consistent estimator of $\theta_0 \in \mathbb{R}^p$ and $\psi : \mathbb{R}^{d+p} \rightarrow \mathbb{R}^q$ a function. Then $n^{-1} \sum_{i=1}^n \psi(Y_i, \hat{\theta})$ converges in probability to $\mathbb{E} \psi(Y, \theta_0)$ where $Y \sim F$, if there exists an F -integrable function $m : \mathbb{R}^{d+p} \rightarrow \mathbb{R}$ and a neighborhood $\mathcal{N}$ for θ_0 such that $\|\psi(y, \theta)\| \leq m(y)$ for F -almost all y and all $\theta \in \mathcal{N}$ and $\psi(y, \theta)$ is continuous in θ with probability 1 for all $\theta \in \mathcal{N}$.

Proof. Since $\hat{\theta}$ is consistent for θ_0 , there exists a compact subset K of $\mathcal{N}$ such that $\hat{\theta} \in K$ with probability tending to 1. We will assume $\hat{\theta} \in K \subseteq \mathcal{N}$, without loss of generality.

Since K is compact and ψ is continuous in θ with probability 1 on K , the existence of m as in the theorem, guarantees that the uniform law of large numbers holds true on

K . Hence, $\sup_{\theta \in K} \|n^{-1} \sum \psi(Y_i, \theta) - E\psi(Y, \theta)\|$ converges in probability to 0. This implies that $\|n^{-1} \sum \psi(Y_i, \hat{\theta}) - E\psi(Y, \theta_0)\| \xrightarrow{pr} 0$, showing the theorem. ■

Theorem 12. *Let all quantities be defined as in Lemma 3, and assume that condition (A1–A4) of lemma 3 and Theorem 11 hold true. Then, replacing J , K , V_j , and W_j for $j = 1, \dots, p$ in Theorem 3 with the empirical variances, covariance or means of the corresponding functions evaluated at $\hat{\theta}$ yields a consistent estimator of c .*

Proof. By the continuous mapping theorem, it suffices to show that the proposed estimators of J , K , V_j , and W_j for $j = 1, \dots, p$ are consistent.

Arguing as in the beginning of the proof of Theorem 11, one can show that there exists functions $m_1(y)$, $m_2(y)$ and $m_3(y)$ with finite expectation bounding all first-, second-, and third-order partial derivatives of $\log f_\theta(y)$ in a neighborhood of θ_{lf} , respectively. This, combined with Lemma 4, ensures that the estimators for J and W_j for $j = 1, \dots, p$ are consistent. The estimator of the matrix K is

$$\frac{1}{n} \sum_{i=1}^n \nabla \log f_{\hat{\theta}}(Y_i) \nabla \log f_{\hat{\theta}}(Y_i)^T.$$

The existence of m_1 guarantees that the norm of the above expression is bounded by $Cm_1(y)^2$ for some $C > 0$. Direct computations show $m(y) = C_1g(y) + C_2m(Y)$ where $g(y)$ is some linear combination of first-, second-, and third-order partial derivatives of $\log f_\theta(y)$ at θ_{lf} . By condition (A1), $Eg(Y)^2$ exists and is finite and hence, Hölder's inequality implies

$$Em_1(Y)^2 \leq Eg(Y)^2 + 2(Eg(Y)^2 Em(Y)^2)^{1/2} + Em(Y)^2.$$

Each term on the right-hand side of this equation is finite, showing that $Em_1(Y)^2 < \infty$. By lemma 4, the estimator of K is consistent for the true value of the matrix. The argument for the estimators of V_j for each $j = 1, \dots, p$ is similar and left out. ■

Paper II

Model robust hybrid likelihood

Model robust hybrid likelihood

Ingrid Dæhlen and Nils Lid Hjort

Department of Mathematics, University of Oslo

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

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

1. Introduction and summary

Parametric models are of course widely used in statistics. Procedures associated with these are often interpretable, and under model conditions they are in some sense asymptotically optimal. When the model fitted is incorrectly specified, however, these properties are no longer guaranteed to hold and instead bias can be introduced. Nonparametric methods, on the other hand, typically have little to no bias. They can, however, be quite variable, which might make the methods too imprecise to be useful in practice. In this article, we study hybrid likelihood theory, a middle ground between these two extreme approaches. We derive new and model agnostic results for the framework and greatly extend its applicability by proposing ways of deciding ‘how parametric’ the method should be.

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

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

where $\text{EL}_n[\mu(\theta)]$ is the empirical likelihood function, described in Section 2, a nonparametric counterpart to $L_n(\theta)$, evaluated at a context-driven control parameter $\mu(\theta)$, and a is a tuning parameter in $[0, 1)$ deciding how much weight should be put on the parametric and nonparametric parts of HL_n . The control parameter μ is some pre-specified parameter we wish to estimate robustly, i.e. without necessarily trusting the parametric model. Working with HL_n instead of either of its components separately allows us to choose a middle ground between purely parametric and nonparametric approaches.

For the nonparametric part of HL_n , the empirical likelihood function is used. This concept was first introduced in Owen (1988), and has since found numerous extensions and applications. For an overview, see e.g. Owen (2001) or Hjort, McKeague, and Van Keilegom (2009). So far, however, there have been few attempts to combine parametric and empirical likelihoods. Certain works study ways of combining the two methods when a parametric model is available for a subset of the data only, see e.g. Qin (1994). This is, however, quite different from the framework of Hjort et al. (2018), where information from empirical and parametric likelihoods constructed using the whole sample are combined. Because of this, our hybrid likelihood framework turns out to be more related with semi-parametric methods like those of e.g. Olkin and Spiegelman (1987) or Mazo and Portier (2021). Furthermore, the framework has certain conceptual similarities with methods developed from a rather different Bayesian semi-parametric setup in Hjort (1986). There a given parametric model has its likelihood modified by certain control sets, via a pinned down Dirichlet process on well-defined residuals.

In this article we also derive new results for empirical likelihood theory. In Section 2, we use arguments similar to those of Molanes Lopez, Van Keilegom, and Veraverbeke (2009) to derive an alternative characterization of the empirical likelihood function and the maximum empirical likelihood estimator (MEL estimator), defined and studied in Qin and Lawless (1994). Rather than arguing as in that article, however, we use the framework of generalized estimating equations (GEE), see e.g. Shao (2003, chapter 5). The method of GEEs allows us to study the limiting behaviour of the MEL estimator when plug-in estimators are used in the empirical likelihood function, i.e. $\hat{\mu}(\hat{P})$, the maximizer of $\mu \mapsto EL_n(\mu, \hat{P})$ where $\hat{P}$ is some estimator. The behaviour of $\log EL_n[\hat{\mu}(\hat{P}), \hat{P}]$ has been studied in Hjort et al. (2009), and Molanes Lopez et al. (2009) worked with $\hat{\mu}(\hat{P})$ when the plug-in estimator is a function of μ , equal to the maximizer of $EL_n(\mu, P)$ for each fixed μ . So far, however, no general result concerning the limiting behaviour of $\hat{\mu}(\hat{P})$ for a general plug-in estimator has been derived. We believe our results fill this gap in the literature.

In Section 3 we study the asymptotic properties of the hybrid likelihood function and its maximizer. Hjort et al. (2018) did this, to some degree, by deriving asymptotic results for the hybrid likelihood function and its maximizer when the parametric model used in construction of the hybrid likelihood function is correctly specified, but their results do not hold true under model misspecification. Furthermore, we also derive results covering the case with plug-in estimators, which was not studied in Hjort et al. (2018). In Section 3 we use the theorems of Section 2 to show that the MHL estimator aims for the minimizer of a well-defined divergence $d_{a,\mu}$ and extend the results of Hjort et al. (2018) to the case of possible model misspecification. We also mention that some preliminary results were reached in the master thesis Dæhlen (2021). Our theorems also hold for the MEL estimator, making the theorems in Section 3.1 tools also for empirical likelihood theory in the context of possible model misspecification, i.e. when $Em(Y, \mu) = 0$ has no solution. This topic is apparently not previously studied in the literature. Both Schennach (2007) and Chib et al. (2018) study somewhat similar situations, but they both work with exponentially tilted empirical likelihoods, resulting in slightly different estimators than the MEL estimator.

For hybrid likelihood to be useful in practice, we need to choose how much weight should be put on the parametric and nonparametric part of HL_n . In Section 4, we construct a focused information criterion (FIC) for models fitted with hybrid likelihood. The FIC was first introduced in Claeskens and Hjort (2003) for models in a local misspecified modelling

framework and in Hjort et al. (2018) the authors derive a FIC similar in spirit to those of Claeskens and Hjort (2003). Our version of the FIC, however, is based upon the relatively newer version of the criterion introduced in Jullum and Hjort (2017), with important modifications in Dæhlen, Hjort, and Hobæk Haff (2023). The criterion directly estimates the mean squared error of some pre-specified parameter of interest, which may or may not be equal to the control parameter. Using our proposed strategy to set a should therefore, in theory, be asymptotically optimal and lead to more precise estimation than either of the methods can achieve on their own. This holds true without requiring model misspecification to disappear in the limit or to be absent altogether. Our derivations are based on the model robust results derived in Section 3.1, and would not be possible with previous methods like those of Qin and Lawless (1994), White (1982), or Hjort et al. (2018). We therefore believe we offer a new way to fit a hybrid parametric-nonparametric model in a way that is theoretically justified and guaranteed to lead to improved asymptotic inference.

After having discussed also regression settings, in Section 5 and presented some simulations in 6, we go on to illustrate the hybrid likelihood machinery in Section 7, pertaining to the important battle death series for all major interstate wars from 1823 to the 2022; see the Correlates of War data set, starting with Sarkees and Wayman (2010) and later curated by other scholars. Our analysis is a contribution to the ongoing discussion of whether wars have become somewhat less deadly over time.

2. Empirical likelihood

Before we can study the hybrid likelihood function and its maximizer, we define and take a closer look at the empirical likelihood function. This section will serve as motivation for the theory we derive for the hybrid likelihood function in Section 3, in addition to establish notation which will be used throughout this article.

Empirical likelihood is a nonparametric tool for making inference about solutions to equations on the form $0 = E[m(Y, \mu)]$, where $Y \in \mathbb{R}^d$ is a random variable, $m: \mathbb{R}^{d+s} \rightarrow \mathbb{R}^q$ is some function and $\mu \in \mathbb{R}^s$ is the parameter we are solving the equation with respect to. Such expressions are called estimating equations, and m is referred to as the estimating function. Many quantities can be expressed as solutions to equations on the form of $E[m(Y, \mu)] = 0$ for some function m . Examples include quantiles, moments, probabilities, and regression coefficients.

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

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

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

Using the Lagrange method, one can show

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

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

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

under the condition that $1 + \lambda^T m(Y_i, \mu) \leq 0$ for all $i = 1, \dots, n$. When no such $\lambda_n(\mu)$ exists, $\text{EL}_n(\mu) = 0$. For derivations, see e.g. Owen (2001) or Hjort et al. (2018).

The right-hand side of (2.4) is a mean of functions. Because of this, $\lambda_n(\mu)$ is a so-called Z-estimator for each fixed value of μ . A Z-estimator is an estimator which can be expressed as the zero of an expression on the form $\sum_{i=1}^n \psi(Y_i, \theta)$ for some known function ψ . Such estimators have been well studied in the literature (see e.g. Van der Vaart (2000, Ch. 5)), and we will now use the relevant theory to derive an asymptotically equivalent expression for $\log \text{EL}_n(\mu)$. For simplicity, we will leave μ out of the notation and replace $m(Y_i, \mu)$ by M_i and $m(Y, \mu)$ for a general $Y \sim F$ by M .

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

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

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

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

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

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

Let μ_0 solve the estimating equation $E[M(\mu)] = 0$. Then $E\{M(\mu_0)/[1 + 0 \cdot M(\mu_0)]\} = 0$. Hence, $\lambda(\mu_0) = 0$. Plugging this into (2.6) shows $-2 \log \text{EL}_n(\mu_0) = V_n(\mu_0)^T S(\mu_0)^{-1} V_n(\mu_0) + o_{\text{pr}}(1) \xrightarrow{d} \chi_q^2$ as $V_n(\mu_0)$ converges in distribution to a $N[0, S(\mu_0)]$ distribution by the central limit theorem. Hence, (2.6) allows us to rediscover the main theorem of empirical likelihood theory. This is not the only application of the alternative characterization in (2.6), however. After dividing by n we notice that the right-hand side of the equation is asymptotically equivalent to $-n^{-1} \sum_{i=1}^n \log[1 + \lambda(\mu)^T m(Y_i, \mu)]$. In the following section, we will use this fact to derive model robust theory in the context of hybrid likelihood.

3. Hybrid likelihood

Hybrid likelihood inference was introduced in Hjort et al. (2018) and is based on the study of the associated hybrid likelihood function. As explained in the introduction, this is a combination of the likelihood function of the data, L_n , and the empirical likelihood function of a pre-specified parameter of interest called the control parameter and denoted by μ . We will assume that μ can be identified by the estimating equation $E[m(Y, \mu)] = 0$. Furthermore, for a given parametric model f_θ with $\theta \in \Theta \subseteq \mathbb{R}^p$, we will let $\mu(\theta)$ refer to the value the control parameter takes in the model f_θ . With this notation, we define the hybrid likelihood function as $\text{HL}_n(\theta) = L_n(\theta)^{1-a} \text{EL}_n[\mu(\theta)]^a$, where $\text{EL}[\mu(\theta)]$ is the empirical likelihood function for μ evaluated at $\mu(\theta)$ and constructed using the estimating function m , and $a \in [0, 1]$ is some tuning parameter deciding how much weight should be put on the parametric and nonparametric part of HL_n . The quantity a is called the balance parameter.

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

In generalization of the above, we may construct a histogram controlled hybrid likelihood framework as follows. For bins $C_1, \dots, C_k$, associated with a histogram, use estimating equations $m_j(u, \mu_j) = I(y \in C_j) - \mu_j$ for $j = 1, \dots, k-1$. This corresponds to using the probabilities $\mu_j = \Pr(Y \in C_j)$ for $j = 1, \dots, k$ as control parameters, and hence fit into the hybrid likelihood framework by using $m(y, \mu) = (m_1(y, \mu), \dots, m_{k-1}(y, \mu))^T$ and $\mu =$

$(\mu_1, \dots, \mu_{k-1})^T$ where $m_j(y, \mu) = I(y \in C_j) - \mu_j$ and $\mu_j(\theta) = \Pr_\theta(Y \in C_j)$ for $j = 1, \dots, k-1$. Here $\Pr_\theta(Y \in C_j)$ is the probability of Y landing within C_j in the parametric model f_θ for each j . By using these control parameters and corresponding estimating functions, the hybrid likelihood framework can be used to fit f_θ in a way ensuring that the probability of landing within each cell $C_1, \dots, C_{k-1}$ is robustly estimated. This corresponds to ensuring that the histogram of the data with bins $C_1, \dots, C_k$ fits well with the fitted model.

We illustrate this idea on a data set taken from Pearson (1902) containing life lengths of 141 men and women in Roman era Egypt. The data range from 1.5 to 96.0, and for this illustration we will fit a Weibull model to the life lengths. We will do this using the hybrid likelihood framework with $C_1 = [0, 20), \dots, C_5 = [80, 100)$ and use m as described above. In this case, $\mu_j(k, \lambda) = \exp[-(20(j-1)/\lambda)^k] - \exp[-(20j/\lambda)^k]$. The fit obtained using maximum likelihood is displayed in Figure 1, together with the fits resulting from maximizing the hybrid likelihood function for different values of the balance parameter. From the figure we notice that as a increases in value, the fitted density moves from the maximum likelihood estimate towards a curve which agrees somewhat more with the histogram of the data. This is a consequence of the empirical likelihood part of the hybrid likelihood function becoming more dominant as a increases. As the Weibull-model fits the data quite well in this case, the fits are quite similar for different values of a .

In other situations, however, the difference may be more substantial than for the data with Egyptian life lengths. For illustration of this, consider the dataset of 66 measurements of the speed of light, carried out by Simon Newcomb in 1882 (see e.g. Stigler, 1977). Each observation is measured in nanoseconds and recorded as the deviation from 24.8. The true value is 33.2. A histogram is shown in Figure 2 together with normal densities fitted using histogram controlled maximum hybrid likelihood for different values of a . From the figure we see that we get very different fits with different choices of balance parameter. This is a consequence of the outliers in the data set dominating the maximum likelihood fit and increased robustness as $a \rightarrow 1$.

Now that we have introduced the framework of hybrid likelihood, we will study the asymptotic behaviour of the maximizer of the hybrid likelihood function. We will refer to this as the maximum hybrid likelihood estimator (the MHL estimator). Hjort et al. (2018) consider the case when the parametric model f_θ is correctly specified. When this assumption holds true, the MHL estimator converges in probability to the true value of θ_0 and at the speed of $O(1/\sqrt{n})$. As a result, the MHL estimator of μ also converges to the true value of μ in the underlying distribution of the data. Classic results concerning the empirical likelihood function in shrinking neighbourhoods of the true value (e.g. Hjort et al. (2009) and Molanes Lopez et al. (2009)) can therefore be used to derive limit results. When, on the other hand, the true underlying density is not assumed to be a member of the parametric family, there is no guarantee that the MHL estimator of the control parameter converges to the true solution of the estimating equation. Because of this, the results of e.g. Molanes Lopez et al. (2009) and Hjort et al. (2009) are not sufficient. Instead, we will use the arguments of the previous section.

3.1 The hybrid likelihood function and its maximizer

Similarly as we did for the empirical likelihood function, we start by deriving a pointwise asymptotically equivalent expression for $n^{-1} \log \text{HL}_n(\theta)$, so $a \in [0, 1)$ and let $h_n: \mathbb{R}^p \rightarrow \mathbb{R}$ be the hybrid log-likelihood function, defined as $h_n(\theta) = (1 - a)\ell_n(\theta) + a \log \text{EL}_n[\mu(\theta)]$,

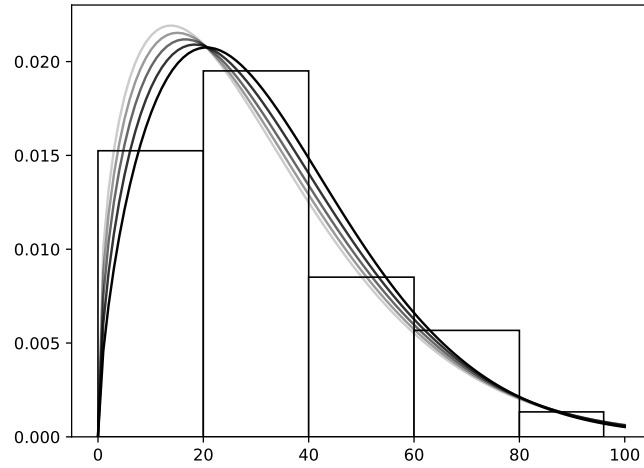


Figure 1: The figure shows a histogram of the life lengths for 141 men and women in Roman era Egypt together with the density of four fitted Weibull models. The model fitting was carried out by maximizing the hybrid log-likelihood function for $a = 0, 0.25, 0.5, 0.75, 0.99$. The value of a is indicated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

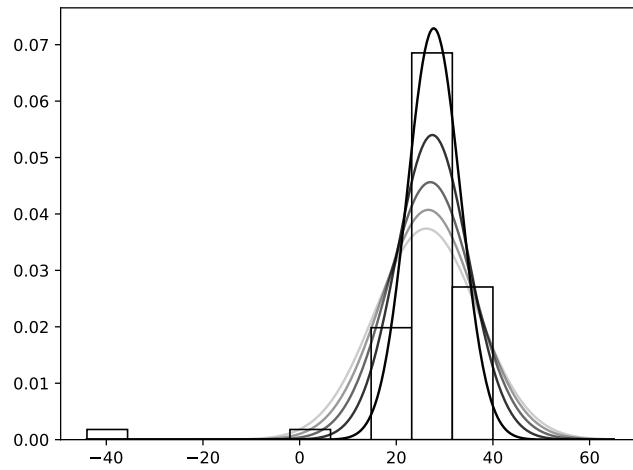


Figure 2: The figure shows a histogram of transformed measurements of the speed of light together with the density of four fitted Normal models. The model fitting was carried out by maximizing the hybrid log-likelihood function for $a = 0, 0.25, 0.5, 0.75, 0.99$. The value of a is indicated by the colour of the line with light values corresponding to low values of a .

where ℓ_n is the log-likelihood of the data. Applying (2.6) shows $h_n(\theta) = \sum_{i=1}^n \psi[Y_i, \mu(\theta)] + V_n[\mu(\theta)]^T S[\mu(\theta)]^{-1} V_n[\mu(\theta)] + o_{\text{pr}}(1)$, where $\psi(y, \theta) = a \log f_\theta(y) - (1-a) \log\{1 + \lambda[\mu(\theta)]^T m[y, \mu(\theta)]\}$. In particular, the above shows that $n^{-1}h_n(\theta)$ is asymptotically equivalent to a mean. Because of this, we would expect its maximizer to be ‘almost’ an M-estimator, a maximizer of a mean of functions (see e.g. Van der Vaart (2000, Ch. 5) for definitions and properties of M-estimators). The following theorem shows that this intuition is indeed correct. The proof is given in the supplementary material, and to ease the notation, we will let $M_i(\theta)$ act as short-hand for $m[Y_i, \mu(\theta)]$ and $M(\theta)$ for $m[Y, \mu(\theta)]$ for a general $Y \sim F$. Furthermore, we will let $\lambda(\theta)$ be the solution to $0 = E\{M(\theta)/[1 + \lambda^T M(\theta)]\}$, which is assumed to be unique.

In the following theorem, we will need certain regularity conditions to be full-filled. Some of these are relatively straightforward, requiring the involved functions to be smooth. More specifically, for a certain set $\mathcal{N}$, we will need μ to be twice continuously differentiable and for $(\theta, \lambda) \mapsto E\{\partial_\theta \psi(Y, \theta), M(\theta)^T/[1 + \lambda^T M(\theta)]\}^T$ to be continuously differentiable. Furthermore, we will need both m and $\log f_\theta$ to be regular on this set, in the sense that both the norm of m and all partial derivatives of m with respect to μ needs to be bounded by functions not depending on μ with finite fourth moments. Similarly, we will need all second order partial derivatives of m as well as the score function in the model and all its partial derivatives to be bounded by square integrable functions. Under these conditions, the following theorem holds true.

Theorem 1. *Let u be the score function in the model f_θ and fix $a \in [0, 1]$. Suppose that $\hat{\theta}$ is the MHL estimator and unique critical point of h_n in a open set Θ . Define $\psi(y, \theta) = (1-a) \log f_\theta(y) - a \log[1 + \lambda(\theta)^T m(y, \theta)]$ and let θ_0 be the maximizer and unique critical point of $\theta \mapsto E\psi(Y, \theta)$ in Θ . Then $\hat{\theta}$ is consistent for θ_0 and*

$$\hat{\theta} = \theta_0 + n^{-1}(J_a^*)^{-1} \sum_{i=1}^n \nabla_{\theta_0} \psi(Y_i, \theta_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (3.7)$$

where $J_a^* = -H_{\theta_0} E\psi(Y, \theta)$ and ∇_{θ_0} and H_{θ_0} denotes the gradient and hessian matrix evaluated at θ_0 respectively, provided there exists a neighbourhood $\mathcal{N}$ of $(\theta_0^T, \lambda(\theta_0)^T)^T$ on which the regularity conditions discussed in the paragraph preceding the theorem holds true, and in addition, that the function $(\theta, \lambda) \mapsto E\{\partial_\theta \psi(Y, \theta), M(\theta)^T/[1 + \lambda^T M(\theta)]\}^T$ has non-singular derivative at $(\theta_0^T, \lambda(\theta_0)^T)^T$, and that there exists some positive L such that $\Pr(1 + \lambda^T M(\theta) > L) = 1$ on $\mathcal{N}$. In particular, we have $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, (J_a^*)^{-1} K_a^* (J_a^*)^{-1}]$, where $K_a^* = \text{Var}[\psi(Y, \theta_0)]$.

Remark 1. Most of the assumptions of Theorem 1 are more or less standard conditions ensuring sufficient regularity. The condition that, in $\mathcal{N}$, $\Pr(1 + \lambda^T M(\theta) > L) = 1$ for some $L > 0$, however, warrants some discussion, as it excludes situations where $m(Y, \mu)$ has unbounded support, leading to several important situations not being covered. The perhaps most obvious example is $m(y, \mu) = y - \mu$ when F has unbounded support. The assumption is, however, necessary, see e.g. Schennach (2007, Thm. 1). In practice, this problem can be avoided by truncating data to lie within some arbitrary large and fixed interval.

Remark 2. For $a = 0$, θ_0 is by definition equal to the minimizer of the Kullback-Leibler divergence from f_θ to F . Furthermore, we notice that J_0^* and K_0^* are equal to the fisher information matrix and the variance of the score function in the model f_θ , respectively.

Hence, Theorem 1 reduces to the results of White (1982) when $a = 0$. Additionally, even though the hybrid likelihood function is defined only for $a \in [0, 1]$, the theorem above can be applied when $a = 1$ and $h_n(\theta) = \log \text{EL}_n[\mu(\theta)]$. In this case, it is most natural to let $\mu(\theta) = \theta$, for which Theorem 1 simplifies to a version of Qin and Lawless (1994, Thm. 1).

If the true underlying distribution is a member of the parametric family fit to the data, there is a vector $\theta_0 \in \mathbb{R}$ such that the distribution F has density function f_{θ_0} . In this case, $\mu(\theta_0)$ solves the estimating equation $\text{E}\{m[Y, \mu(\theta_0)]\} = 0$ and $\text{E}\{m(Y, \mu)/[1 + 0]\} = 0$. Hence $\lambda[\mu(\theta_0)] = 0$, which together with direct computations shows that the derivative of $-\text{E}\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$ at $\mu(\theta_0)$ is zero. Because of this, $-\text{E}\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$ attains its maximum at $\mu(\theta_0)$, and as a result, both $\text{E}[\log f_{\theta}(Y)]$ and $-\text{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$ are maximized by θ_0 , making this vector the maximizer of $\theta \mapsto (1-a)\text{E}[\log f_{\theta}(Y)] - a\text{E}\{\log[1 + \lambda(\theta)^T M(\theta)]\}$. Furthermore, direct computations show

$$\nabla_{\theta_0} \psi(y, \theta) = (1-a)u(y, \theta_0) - a\text{E}[M'(\theta_0)]^T \text{E}[M(\theta_0)M(\theta_0)^T]^{-1} \text{E}[M'(\theta_0)],$$

Plugging this into Theorem 1 shows $K_a^* = (1-a)^2 J + a^2 \xi_0^T W^{-1} \xi_0 - a(1-a)(C W^{-1} \xi_0 + \xi_0^T W^{-1} C^T)$, where J is the Fisher matrix, W is the covariance matrix of $m[Y, \mu(\theta_0)]$, C is the covariance between $u(Y, \theta_0)$ and $m[Y, \mu(\theta_0)]$, and $\xi_0 = \text{E} M'(\theta_0)$. In addition, $J_a^* = (1-a)J + a\xi_0^T W^{-1} \xi_0$ by the chain rule. Hence, if the model is correctly specified, Theorem 1 simplifies to the corresponding results given in Hjort et al. (2018), making Theorem 1 a genuine model robust extension of the results of Hjort et al. (2018).

In certain situations, it can be beneficial to use a plug-in estimator inside the empirical likelihood part of h_n . Say that we, for instance, want to control for robust estimation of the variance σ^2 in a distribution. This can be achieved by using the hybrid likelihood framework with estimating function $m(y, \mu, \sigma^2) = (y - \mu, (y - \mu)^2 - \sigma^2)$ this method however introduces the mean as a second control parameter. A perhaps more intuitive way could be to replace μ with the empirical mean of the observations $\hat{\mu}$ and use $m(y, \sigma^2) = (y - \hat{\mu})^2 - \sigma^2$. The following result allows for such constructions. The proof is given in the supplementary material, and to improve readability, we adopt the following notation: Let $m: \mathbb{R}^{d+r+s} \rightarrow \mathbb{R}^q$ be an estimating function depending on a parameter P which we wish to estimate using a plug-in estimator, i.e. $(y, \mu) \mapsto m(y, \mu, \hat{P})$ is used as estimating function in the empirical likelihood function for some estimator $\hat{P}$. Let $\mu(\theta, P)$ be defined as the solution to $\text{E}\{m[Y, P, \mu(\theta, P)]\} = 0$ when $Y \sim f_{\theta}$. Lastly, let $\lambda(\theta, P)$ be the solution to $0 = \text{E}\{M(\theta, P)/[1 + \lambda^T M(\theta, P)]\}$ for each θ and for each P .

Again we will need a set of regularity conditions. These are relatively similar to those given in the paragraph preceeding Theorem 1. For a certain set $\mathcal{N}$, we will need μ to be twice continuously differentiable and for $(\mu, \lambda, P) \mapsto \text{E}\{(\partial_{\theta} \psi_P(Y, \theta, P), M(\theta, P)^T/[1 + \lambda^T M(\theta, P)], h(Y, P)^T)\}^T$ to be continuously differentiable. Furthermore, we will need the norm of m and all partial derivatives of m with respect to μ and P to be bounded by functions not depending on μ or P with finite fourth moments. Similarly, we will need all second order partial derivatives of m as well as the score function in the model and all of its partial derivatives to be bounded by square integrable functions. Under these conditions, the following theorem holds true.

Theorem 2. Fix $a \in [0, 1]$ and assume that $\hat{P} \in \mathbb{R}^r$ is an estimator satisfying $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_p(1/\sqrt{n})$ for some function $h: \mathbb{R}^{d+r} \rightarrow \mathbb{R}^r$ and that $\hat{P}$ is used as a plug-in estimator for

P in the empirical likelihood part of h_n . Let u be the score function in the parametric family, and $\hat{\theta}(\hat{P})$ be the maximizer and unique critical point of h_n . Define $\psi_P(y, \theta, P) = (1 - a) \log f_\theta(y) - a \log[1 + \lambda(\theta, P)^T m(y, \theta, P)]$. For each P assume that $\theta(P)$ maximizes and is the unique critical point of $\theta \mapsto E \psi_P(Y, \theta, P)$. Then the MHL estimator $\hat{\theta}(\hat{P})$ is consistent for $\theta_0 = \theta(P_0)$ and

$$\hat{\theta}(\hat{P}) = \hat{\theta}(P_0) + \theta'(P_0)(\hat{P} - P_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (3.8)$$

provided there exists a neighbourhood $\mathcal{N}$ of $\alpha_0 = (\theta_0^T, \lambda(\theta_0, P_0)^T, P_0^T)^T$ on which the regularity conditions stated in the paragraph preceeding the theorem hold true, and in addition that the function $(\mu, \lambda, P) \mapsto E\{(\partial_\theta \psi_P(Y, \theta, P), M(\theta, P)^T/[1 + \lambda^T M(\theta, P)], h(Y, P)^T)\}^T$ has non-singular derivative at α_0 . Furthermore, we need there to exist a positive L such that $\Pr(1 + \lambda^T m(Y, \mu, P) > L) = 1$ on $\mathcal{N}$, and that on $\mathcal{N}$ the norms of h and $\partial h/\partial P$ are bounded by F -square integrable functions not depending on P or θ .

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

Let V be the Jacobian matrix of $\theta \mapsto -E h(Y, P)$ at P_0 . Then we have $\hat{P} - P_0 = n^{-1} V^{-1} \sum_{i=1}^n h(Y_i, P_0) + o_p(1/\sqrt{n})$, by the conditions above and Van der Vaart (2000, Thm. 5.21). Furthermore, $\hat{\theta}(P_0)$ satisfies (3.7) with $m(y, \mu)$ replaced by $(y, \mu) \mapsto m(y, \mu, P_0)$ by Theorem 1. Hence, the central limit theorem guarantees that $\sqrt{n}[\hat{\theta}(P_0) - \theta(P_0)]$ and $\sqrt{n}(\hat{P} - P_0)$ converge jointly in distribution to a central normal distribution with covariance matrix

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

where $V = E[\partial_{P_0} h(Y, P)]$ and C_θ is the covariance between $\partial_{\theta_0} \psi_P(Y, \theta, P_0)$ and $h(Y, P_0)$ where $Y \sim F$. This together with (3.8) ensures that $\sqrt{n}[\hat{\theta}(\hat{P}) - \theta(P_0)]$ converges to a normal distribution with mean zero and variance $(I_p, \mu'(P_0)^T) \Sigma_\theta (I_p, \mu'(P_0)^T)^T$.

3.2 The hybrid divergence

By Theorem 1, the maximum hybrid likelihood estimator converges in probability to the maximizer of $E \psi(Y, \theta) = (1 - a) E[\log f_\theta(Y)] - a E\{\log[1 + \lambda(\theta)^T M(\theta)]\}$. This function is useful for showing consistency of the MHL estimator, but the expression is somewhat hard to interpret, and it is not clear from the formula why we should want to maximize it. In this section, we will take a closer look at $E \psi(Y, \theta)$ and show that maximizing this expression is equivalent to minimizing a divergence taking both model fit and precise estimation of the control parameter into account.

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

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

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

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

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

The first term in (3.10) is small when the overall model fit is good, i.e. when h is close to g overall. The second term, on the other hand, does not emphasise overall model fit, but is small when $\mu(h)$ is close to $\mu(g)$. Hence, d_a measures the distance from g to h by a convex combination of the lack of overall model fit and error involved when using $\mu(h)$ in place of $\mu(g)$. By Theorem 1, the MHL estimator converges to θ_0 such that f_{θ_0} minimizes d_a in the class f_θ for $\theta \in \Theta$. Hence, the limit of the MHL estimator is the value of θ making f_θ as close as possible to the true density g , and distance is measured as a trade-off between lack of model fit and bias of the control parameter, and how much weight should be put on each term in the divergence is determined by the value of the tuning parameter.

4. The balance parameter

To use the hybrid likelihood theory in practice, we need to choose a control parameter, μ , and set the value of the balance parameter, a . The choice of the former should reflect what parameters are of particular interest in the specific application, and should not be chosen in an automatic way. In some situations, the same might be true for a . Usually, however, how much weight should be put on the parametric and nonparametric part of HL_n is dependent on the quality of the model fit and estimator of the control parameter. In this section, we will propose a way to choose a taking these considerations into account.

4.1 FIC for models fitted with hybrid likelihood

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

The focused information criterion (see Claeskens and Hjort (2008) for an introduction) ranks models by the quality of their estimate of the focus parameter. The information criterion is created to choose between different parametric models, but the estimators defined can in principle be used for other purposes. We propose to use the information criterion to choose the optimal value for the balance parameter a used in construction of the hybrid likelihood function. This idea was briefly visited in Hjort et al. (2018), where the authors derive formulas for a focused information criterion for models in shrinking neighbourhoods of

the true underlying density. More precisely, Hjort et al. (2018) formulate an FIC for models on the form $f(y, \theta)$ when the true underlying distribution of the data is $f_{\text{true}}(y, \theta, \gamma_0 + \delta/\sqrt{n})$ and $f_{\text{true}}(y, \theta, \gamma_0) = f(y, \theta)$. Their version of the information criterion therefore requires the model to be asymptotically correctly specified, and to be able to compute their FIC one needs to define a wider model $f_{\text{true}}(y, \theta, \gamma)$, which is then assumed to be correct. This is not too problematic in e.g. the context of variable selection, but can be quite restrictive for more general use. We will therefore instead create a fully model agnostic FIC, based upon the formulas and results of Jullum and Hjort (2017) with important modifications in Dæhlen et al. (2023). These articles find asymptotically unbiased approximations to the mean squared error of estimators of the focus parameter in a large number of cases. Their results can be used to find estimators of the mean square error of maximum hybrid likelihood estimators of focus parameters. In the following theorem, we state a version of the results of Jullum and Hjort (2017) and Dæhlen et al. (2023) valid in the hybrid likelihood context.

Theorem 3. *Fix a balance parameter a and let $\hat{\theta}$ be the MHL estimator. Let $\xi \in \mathbb{R}$ be some focus parameter and $\hat{\xi}$ a consistent estimator of ξ satisfying*

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

for some function $\phi: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\epsilon_n = O_{\text{pr}}(1/n)$. Assume further that all conditions of Theorem 1 hold true and that there is a continuously differentiable function $g: \mathbb{R}^p \rightarrow \mathbb{R}$ such that $\xi = g(\theta)$ in the model f_{θ} . Then $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$ is only $o(1/n)$ away from the true MSE of $g(\hat{\theta})$ in expected value where $\hat{b} = g(\hat{\theta}) - \hat{\xi}$, $c = \lim_{n \rightarrow \infty} nE(\epsilon_n)$, $\kappa = \text{Var}[\nabla g(\theta_0)^T (J_a^*)^{-1} \psi_{\theta}(Y, \theta_0) - \phi(Y, \xi)]$ and $\tau = (J_a^*)^{-1} K_a^* (J_a^*)^{-1}$, provided the assumptions of Jullum and Hjort (2017) and Dæhlen et al. (2023) hold.

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

In practice, the quantities c , κ and τ are usually unknown. To approximate the MSE of the MHL estimator using Theorem 3, we therefore need to first define estimators of these parameters and approximations to the expression $\hat{b}^2 - n^{-1}(2bc + \kappa - \tau)$. This is provided in the following definition.

Definition 1. Let all quantities be as defined in Theorem 3. We define the following approximation to the MSE of $g(\hat{\theta})$,

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

where $\hat{c}$ is either of the estimators given in Dæhlen et al. (2023), $\hat{\tau} = \nabla g(\hat{\theta})^T (\hat{J}_a^*)^{-1} \hat{K}_a^* (\hat{J}_a^*)^{-1} \nabla g(\hat{\theta})$ and $\hat{\kappa}$ is given by $\hat{\tau} + n^{-1} \sum_{i=1}^n \phi(Y_i, \hat{\xi})^2 - 2 \nabla g(\hat{\theta})^T (\hat{J}_a^*)^{-1} \hat{V}$, with $\hat{J}_a^* = n^{-1} H h_n(\hat{\theta})$, $\hat{K}_a^* = n^{-1} \sum_{i=1}^n \nabla_{\hat{\theta}} \hat{\psi}(Y_i, \theta) \nabla_{\hat{\theta}} \hat{\psi}(Y_i, \hat{\theta})^T$ and $\hat{V} = n^{-1} \sum_{i=1}^n \nabla_{\hat{\theta}} \hat{\psi}(Y_i, \theta) \phi(Y_i, \hat{\xi})^T$. Here $\hat{\psi}(Y_i, \theta)$ is equal to $\hat{\psi}(Y_i, \theta)$ with $\lambda(\theta)$ replaced by $\hat{\lambda}$, the solution to $0 = V_n(\lambda, \theta) = \sum_{i=1}^n M_i(\theta) / [1 + \lambda^T M_i(\theta)]$ and $\lambda'(\theta)$ by $-(S_n^1)^{-1} S_n^2$ where $S_n^1 = \partial_{\lambda} V_n(\lambda, \hat{\theta})$ and $S_n^2 = \partial_{\theta} V_n(\theta, \hat{\lambda})$.

Remark 4. In most applications the control and focus parameter will be the same, but there is no technical reason for this, and, if useful, μ and ψ can be chosen to be different parameters. See Section 7.

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

The above HFIC is only valid when plug-in estimators are not present. Thanks to Theorem 2, however, the criterion can also be defined in this context. We give the formulas below.

Definition 2. Let all quantities be defined as in Theorem 2 and assume the conditions of this theorem hold. Furthermore, let $\xi \in \mathbb{R}$ be some focus parameter taking the value $g(\theta)$ in the parametric model and $\hat{\xi}$ a consistent estimator of ξ satisfying (4.11) for some ϕ . Define $\psi_{\text{FIC}}(y, \theta, \lambda, P) = (m(y, \mu, P)^T / [1 + \lambda^T m(y, \mu, P)], \nabla_\theta \psi_P(y, \theta, P)^T, h(y, P)^T)^T$, and let $J\psi_{\text{FIC}}(y, \theta, \lambda, P)$ denote its Jacobian matrix with respect to all parameters but y . Let $\hat{A} = n^{-1} \sum_{i=1}^n J\psi_{\text{FIC}}(Y_i, \hat{\theta}, \hat{\lambda}, \hat{P})$ and $\hat{B} = n^{-1} \sum_{i=1}^n \psi_{\text{FIC}}(Y_i, \hat{\theta}, \hat{\lambda}, \hat{P})\psi_{\text{FIC}}(Y_i, \hat{\theta}, \hat{\lambda}, \hat{P})^T$ and $\hat{C} = \hat{A}^{-1}\hat{B}(\hat{A}^{-1})^T$. With this we define the HFIC as in Definition 1 with $\hat{b} = g[\hat{\theta}(\hat{P})] - \hat{\xi}$, $\hat{c}$ equal to either of the estimators given in Dæhlen et al. (2023) and $\hat{\tau} = (\nabla g[\hat{\theta}(\hat{P})], 0)^T \hat{C} (\nabla g[\hat{\theta}(\hat{P})]^T, 0)^T$ and

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

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

By Lemma 5 in the supplementary material and the results of Jullum and Hjort (2017) and Dæhlen et al. (2023) the focused information criterion is only $O(1/n)$ away from the true MSE of $g[\hat{\theta}(\hat{P})]$ in expected value.

5. Regression

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

Let $Y \in \mathbb{R}$ be the response variable and X the regressor, and assume that $Z_i = (Y_i, Y_i)$ for $i = 1, \dots, n$ are i.i.d. from some distribution F . Suppose furthermore that we wish to fit a linear regression model to these data. This corresponds to maximizing the function

$\ell_n(\beta_0, \beta) = -\sum_{i=1}^n (Y_i - \beta_0 - \beta^T X_i)^2$. We will use ℓ_n as the likelihood function in the hybrid likelihood framework.

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

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

In practice, estimation of μ and Σ is rarely of interest. Because of this, it can be beneficial to replace them by $\hat{\mu}$, the empirical mean, and $\hat{\Sigma}$, the empirical covariance, respectively rather than optimizing with respect to these variables. This corresponds to using the hybrid likelihood framework with plug-in estimators for μ and Σ . Indeed, let $h(y, x, \mu, \Sigma) = (x^T - \mu^T, (x - \mu)(x - \mu)^T - \Sigma)^T$, then $(\hat{\mu}, \hat{\Sigma})$ satisfies $n^{-1} \sum_{i=1}^n h(Z_i, \hat{\mu}, \hat{\Sigma}) = o_p(1/\sqrt{n})$. Hence, inference of the hybrid likelihood estimator of β using plug-in estimators for μ and Σ can be made using Theorem 2. As a plug-in result is lacking from Hjort et al. (2018), this greatly improves the applicability of the preliminary ideas for regression which are discussed in Hjort et al. (2018).

Remark 5. It is worth noting that when $\mathcal{S}$ consists of all indices, both ℓ_n and $\log \text{EL}_n$ are maximized at $\hat{\beta}_{\text{OLS}}$, the ordinary least square estimator of β . Hence, $\hat{\beta}_{\text{OLS}}$ maximizes h_n for all values of a and is therefore equal to the MHL estimator regardless of the choice of balance parameter. This holds true, whether the parametric model used is specified correctly or not. When $\mathcal{S} \neq \{1, \dots, d\}$, however, $\log \text{EL}_n$ and ℓ_n do not have the same maxima and hence the MHL estimator is not constant as a function of a .

6. Simulations

To make our results more trustworthy we performed a small simulation study, and did the following $B = 100$ times. We drew $n = 100$ data points from a Poisson distribution with rate parameter 1.5 and fitted geometric distributions to the data with maximum hybrid likelihood for a selection of a -values. The geometric distribution has density $\log f_p(y) = y \log(1 - p) + \log p$ for $y = 0, 1, \dots$. We used the probability of exceeding 1 as the control parameter and $m(y, \mu) = I(Y \leq 2) - \mu$ as the estimating function in the empirical likelihood function. We will refer to the estimators as $\hat{p}_b(a)$ for $b = 1, \dots, B$ and $a = 0, 0.01, \dots, 0.99$. The mean of $\hat{p}(a)_1, \dots, \hat{p}_B(a)$ for each of the selected a values are shown in Figure 3 together with the theoretical limit, the minimizer of $d_a(g, f_p)$ where g is the pmf in the $\text{Pois}(2)$

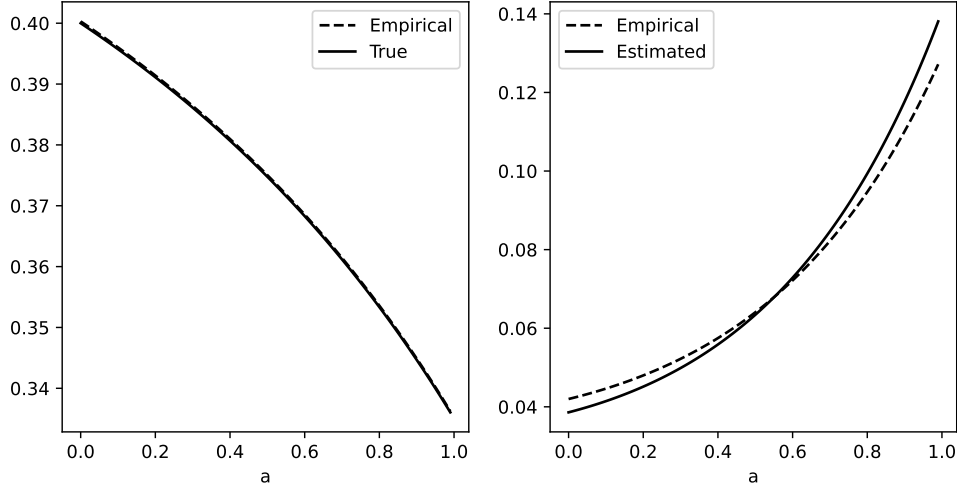


Figure 3: The plot to the left shows the mean of $\hat{p}_1(a), \dots, \hat{p}_B(a)$ for $a = 0, 0.01, \dots, 0.99$ together with the theoretical limit. To the right, the empirical variance of $\sqrt{n}\hat{p}_1(a), \dots, \sqrt{n}\hat{p}_B(a)$ for each a is displayed together with the mean of the estimator of the variance obtained using Definition 1.

distribution. From the figure, we notice that the mean is close to the theoretical limit. In each iteration we also computed the estimated variance of $\sqrt{n}\hat{p}(a)$ for each a using the estimators defined in 1. The mean of the B estimator is shown in Figure 3 together with the empirical variance of $\sqrt{n}\hat{p}_1(a), \dots, \sqrt{n}\hat{p}_B(a)$ for each a . Again, the theory and estimators seem to agree with the simulation results.

We also computed the HFIC score for each a in every iteration. We used the control parameter as the focus parameter and the empirical mean of $I(Y_i \leq 1)$ for $i = 1, \dots, n$ as consistent estimator. This satisfies (4.11) with $\phi(y, \xi) = I(y \leq 1) - \xi$. A plot of the mean of the B HFIC curves is displayed in Figure 4 together with the observed mean squared error of the MHL estimator of the focus parameter. As before, the approximation seem to work very well. Since, the HFIC curves were computed in every iteration, we also registered the balance parameter minimizing the criterion in each case. A histogram of the a chosen is shown in Figure 4, and reveals that $a \neq 0, 0.99$ is often chosen, indicating that a true hybrid model is often preferred. Furthermore, the values of a being most frequently chosen are those close to the minimizer of the MSE, indicating that HFIC is working as intended when it comes to choosing the balance parameter.

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

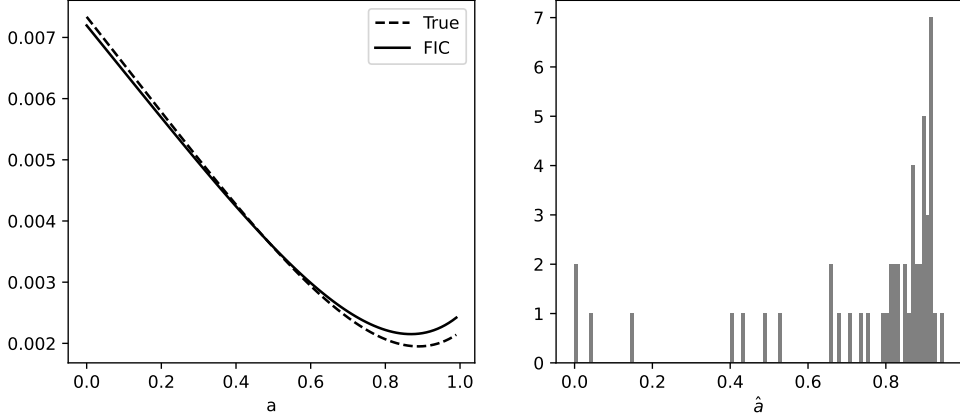


Figure 4: The plot to the left shows the mean of HFIC for $a = 0, 0.01, \dots, 0.99$ together with the true MSE. To the right, a histogram of $\hat{a}_1, \dots, \hat{a}_B$ is displayed.

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

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

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

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

We chose separate balance parameters for each model. In each case, this was done by computing HFIC for each value of a in a grid of points between 0 and 1. The median was used as the focus parameter for both models, and the empirical median in each data set as the consistent estimator $\hat{\xi}$ in Definition 1. Since the median satisfies (4.11) with $\phi(x, \xi) = (I(Y \leq \xi) - 0.5)/f(\xi)$, where f is the true density of the data, this is a valid choice. We estimated all quantities in HFIC as given in Definition 1, with $f(\mu)$ replaced

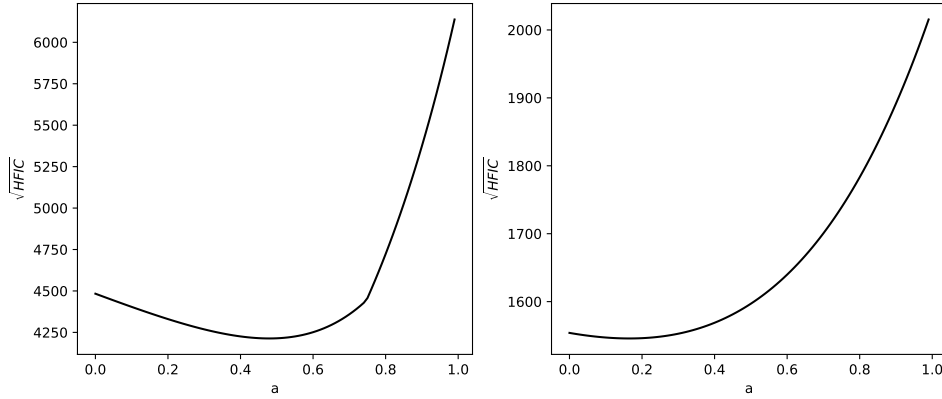


Figure 5: The figure displays root-HFIC scores as functions of a (the balance parameter) for older and newer wars, in the left and right plots respectively. The minimum was attained at $a = 0.48$ for the older wars and $a = 0.17$ for the newer conflicts.

by non-parametric kernel density estimates. For more details see Dæhlen et al. (2023), in particular their Appendix A. The results are shown in Figure 5. We see that neither minima are attained at 0 or 1, i.e. a genuine hybrid method is chosen for both data sets. The optimal balance values are $a = 0.48$ for the older wars and $a = 0.17$ for the newer conflicts.

With all parameters set, we fitted the models to each data set using maximum hybrid likelihood. The resulting MHL estimators of the median for the two data sets were 8192 and 4750 for the older and newer wars respectively. In both cases this is closer to the empirical median in the data sets (11375 and 5240 respectively) than the maximum likelihood estimates (6799 and 4733). In addition, we estimated the standard deviation of the MHL estimator of the medians in the two populations. This was done by applying Theorem 1 in conjunction with the delta method and the estimators defined in Definition 1. Estimated standard deviations were found to be 3488 and 1546 for the older and newer wars respectively. This is only marginally larger than the estimated standard deviations for the ML estimator for the older wars and more or less equal to the that of the the ML estimator for the newer wars (2746 and 1554 respectively). Hence, the added robustness of hybrid likelihood comes at a low price in this case.

We can now make inference about the difference in number of battle deaths between newer and older wars using hybrid likelihood. The MHL estimator of this parameter was found to be 3441, which, relatively speaking, is a lot larger than the corresponding ML estimator of 2066. We can compute variances for this quantity as well by applying the delta method and Theorem 1. We found the standard deviation for the MHL estimator to be 3815, while the ML estimator had a slightly smaller standard deviation of 3155. With these numbers we can construct confidence intervals. A 95% confidence interval obtained with hybrid likelihood was computed to be $[-4036, 10919]$. With maximum likelihood, we get a slightly smaller interval shifted to the left ($[-4135, 8232]$). In addition, we can construct hypothesis tests for testing if the median number of battle deaths is the same for older and newer wars, against the alternative that violence has gone down. The p-value for testing this hypothesis using hybrid likelihood was 0.18, while the test based on maximum likelihood had a p-value of 0.26. These findings are in line with e.g. Clauset (2018).

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

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

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

8. Concluding remarks

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

The perhaps most pressing issue with the theory developed in this article was discussed in Remark 1 following Theorem 1. If m is the estimating function used in the construction of the empirical likelihood function, we require $m(y, \mu)$ to have bounded support for each fixed μ in a neighbourhood of μ_0 . Theoretically, this assumption is quite problematic, but as is shown in Schennach (2007, Thm. 1), such a condition is necessary for the EL-framework to function outside of model conditions. In practice, however, it is less of a problem, as approximations using truncated distributions often give very good results. Take for instance, the situation where $m(y, \mu) = y - \mu$ and the data is normally distributed. Replacing each data point Y by $\text{sign}(Y) \max\{K, |Y|\}$ for some large number K makes all theorems go through while making virtually no difference to the analysis.

Another avenue for further research is to consider the case of non-smooth estimating functions m . The present theorems require m to be differentiable. This excludes the case

of quantiles as $m(y, \mu) = I(y \leq \mu) - p$ for $p \in [0, 1]$ is not even continuous, let alone differentiable in μ . The results do, however, hold true for smoothed approximations to m . Further investigation into the behaviour as the smoothed approximations get better and better might therefore lead to results covering the case of quantiles.

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

Supplementary Materials

Proofs of Theorem 1 and 2 can be found in the article’s supplementary material.

Acknowledgements

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

References

- Chen, S. X. and I. Van Keilegom (2009). A review on empirical likelihood methods for regression. *Test* 18, 415–447.
- Chib, S., M. Shin, and A. Simoni (2018). Bayesian estimation and comparison of moment condition models. *Journal of the American Statistical Association* 113(524), 1656–1668.
- Cirillo, P. and N. N. Taleb (2016a). The decline of violent conflicts: What do the data really say? *Nobel Foundation Symposium 161: The Causes of Peace*, 1–26.
- Cirillo, P. and N. N. Taleb (2016b). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications* 452, 29–45.
- Claeskens, G. and N. L. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association* 98(464), 900–916.
- Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge, England: Cambridge University Press.
- Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances* 4(2), eaao3580.
- Cunen, C., N. L. Hjort, and H. M. Nygård (2020). Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research* 57(2), 221–234.

- Dæhlen, I., N. L. Hjort, and I. Hobæk Haff (2023). Accurate bias estimation with applications to focused model selection. *Scandinavian Journal of Statistics*. Online publication. DOI: 10.1111/sjos.12696.
- Dæhlen, I. (2021). Empirical and hybrid likelihood. Master thesis.
- Gray, J. (2015). Steven Pinker is wrong about violence and war. *The Guardian*, Friday 13 March 2015. Available at: <https://www.theguardian.com/books/2015/mar/13/john-gray-steven-pinker-wrong-violence-war-declining>.
- Hjort, N. L. (1986). Discussion: On the consistency of Bayes estimates. *The Annals of Statistics* 14(1), 49–55.
- Hjort, N. L., I. W. McKeague, and I. Van Keilegom (2009). Extending the scope of empirical likelihood. *The Annals of Statistics* 37, 1079–1111.
- Hjort, N. L., I. W. McKeague, and I. Van Keilegom (2018). Hybrid combinations of parametric and empirical likelihoods. *Statistica Sinica* 28(4), 2389–2407.
- Jullum, M. and N. L. Hjort (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica* 27, 951–981.
- Mazo, G. and F. Portier (2021). Parametric versus nonparametric: The fitness coefficient. *Scandinavian Journal of Statistics* 48(4), 1344–1383.
- Molanes Lopez, E. M., I. Van Keilegom, and N. Veraverbeke (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics* 36(3), 413–432.
- Olkin, I. and C. H. Spiegelman (1987). A semiparametric approach to density estimation. *Journal of the American Statistical Association* 82(399), 858–865.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2), 237–249.
- Owen, A. B. (1991). Empirical likelihood for linear models. *The Annals of Statistics* 19, 1725–1747.
- Owen, A. B. (2001). *Empirical likelihood*. Boca Raton, FL: CRC press.
- Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika* 1(2), 261–264.
- Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Toronto: Viking.
- Qin, J. (1994). Semi-empirical likelihood ratio confidence intervals for the difference of two sample means. *Annals of the Institute of Statistical Mathematics* 46, 117–126.
- Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *the Annals of Statistics* 22(1), 300–325.
- Sarkees, M. R. and F. Wayman (2010). *Resort to war: 1816 - 2007*. Washington DC: CQ Press.
- Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics* 35, 634–672.
- Shao, J. (2003). *Mathematical statistics*. New York: Springer Science & Business Media.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.
- Wang, Q.-H. and B.-Y. Jing (1999). Empirical likelihood for partial linear models with fixed designs. *Statistics & Probability Letters* 41(4), 425–433.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society* 50, 1–25.
- Yan, L. and X. Chen (2015). Empirical likelihood for generalized linear models with fixed and adaptive designs. *Statistics* 49(5), 978–988.

Ingrid Dæhlen

Department of mathematics, University of Oslo

E-mail: ingrdae@math.uio.no

Nils Lid Hjort
Department of mathematics, University of Oslo
E-mail: nils@math.uio.no

Paper III

**The asymptotic effect of tuning
parameters**

The asymptotic effect of tuning parameters

Ingrid Dæhlen, Nils Lid Hjort and Ingrid Hobæk Haff

Department of Mathematics, University of Oslo

Address for correspondence: Ingrid Dæhlen, Department of Mathematics, University of Oslo, Moltke Moes vei 35, 0851 Oslo, Norway, email: ingrdae@math.uio.no

Abstract

Tuning parameters are parameters involved in an estimating procedure for the purpose of reduce the risk of some other estimator. Examples include the degree of penalization in penalized regression and likelihood problems, as well as the balance parameter in hybrid methods. Typically tuning parameters are set to the minimizers of some estimator of the risk, a step which introduces additional randomness and makes standard methodology inapplicable. We derive precise asymptotic theory for this situation. Our framework allows for smooth, but otherwise arbitrary, loss functions and for the risk to be estimated by cross-validation procedures. Results include consistency of the optimal estimator towards a well-defined quantity and asymptotic normality after proper scaling and centring. We give explicit forms and estimators for the limiting variance matrix and results sharply characterizing the distance from the training error to the cross-validated estimator of the risk.

Keywords: cross validation, information criteria, large-sample theory, regularisation, two-stage estimators

1 Introduction

Suppose we have data $Z_1, \dots, Z_n$ and wish to estimate a parameter θ . Assume further that for each fixed value of some other parameter λ this can be done by maximising a

function $\Gamma_n(\theta, \lambda)$ defined in terms of the data. Examples include hybrid estimators where $\Gamma_n(\theta, \lambda) = \lambda\gamma_n^1(\theta) + (1 - \lambda)\gamma_n^2(\theta)$ and penalized likelihood methods, where for each λ , θ is estimated by the maximizer of $\ell_n(\theta) - \lambda R(\theta)$ for some function R . This latter situation covers the ridge regression. An example of the former case is the hybrid generative-discriminative estimator defined as the maximizer of $\lambda\ell_n^{\text{Gen}}(\theta) + (1 - \lambda)\ell_n^{\text{Disc}}(\theta)$ where ℓ_n^{Gen} and ℓ_n^{Disc} are the log-likelihoods in generative and corresponding discriminative models (see e.g. Bouchard and Triggs (2004)).

For each fixed λ , we can maximise $\Gamma_n(\theta, \lambda)$ with respect to θ to get the estimator $\hat{\theta}(\lambda)$. Furthermore, if Γ_n is sufficiently smooth in θ and a relatively standard set of regularity conditions are satisfied, $\hat{\theta}(\lambda)$ is consistent for a well-defined quantity $\theta_0(\lambda)$ and $\sqrt{n}\{\hat{\theta}(\lambda) - \theta_0(\lambda)\}$ converges in law to some known limit distribution. In practice, however λ is rarely fixed, and is instead set to minimise some estimated loss, usually approximated by means of an information criterion or cross-validation. There is therefore an additional randomness involved when working with tuning parameters. Because of this, the standard limit results for $\hat{\theta}(\lambda)$ are not immediately applicable to $\hat{\theta}(\hat{\lambda})$. In practice, this point is often ignored and inference about θ is made by treating λ as fixed at $\hat{\lambda}$ and by using pointwise results for $\hat{\theta}(\lambda)$. Using such a method ignores part of the fitting procedure and added randomness introduced by tuning λ , however. Because of this, the variance of $\hat{\theta}(\hat{\lambda})$ tends to be underestimated.

Figure 1 illustrates the ideas outlined above for a data set concerning female Pima Indians and prevalence of diabetes (publicly available in e.g. the R-package `mlbench`). In Section 6.1 we fit a penalized logistic regression model to these data where the tuning parameter dictating the size of the regularization term is set to the minimizer of the cross-validated Brier score. For each regression coefficient, Figure 1 shows histograms based on 2000 non-parametric bootstrap iterations, together with the densities in approximate distributions. The dotted curve corresponds to the distribution we get when using the pointwise results as described in the previous paragraph. The dashed lines correspond to the alternative approximate distributions derived in this article,

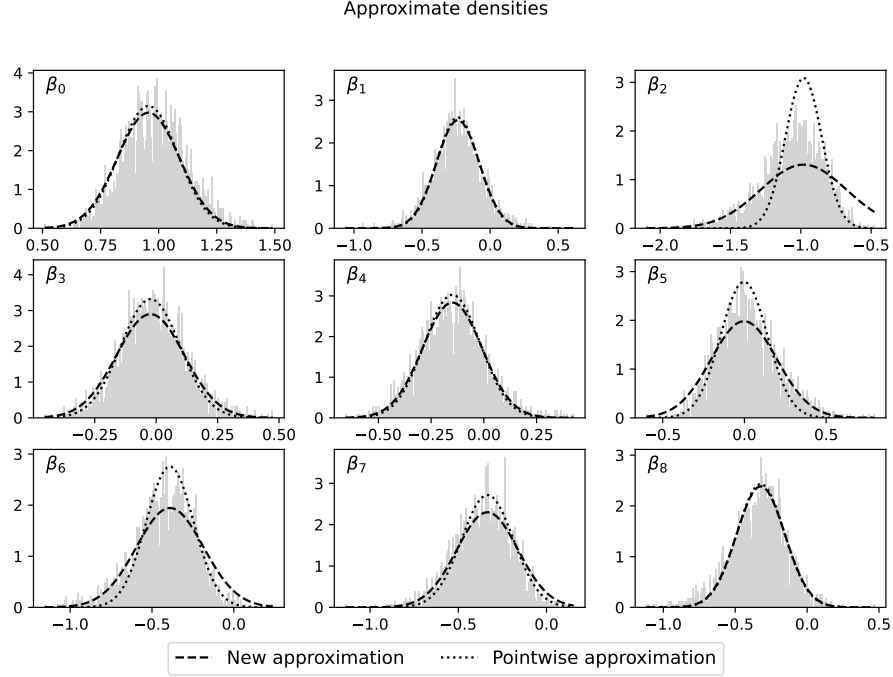


Figure 1: Histograms of draws from the bootstrap distribution of $\hat{\beta}(\hat{\lambda})_j$ for $j = 0, \dots, 8$ together with the density in the approximate distribution.

which take the tuning procedure into account. Note that the curves corresponding to the pointwise estimator seem a bit too narrow, while the curves corresponding to the new estimators are less sharp. This is a consequence of including and not including the effect of the tuning step in the limiting distribution.

In the present article, we fill a gap in literature by deriving precise limit results for the estimator $\hat{\theta}(\hat{\lambda})$, allowing inference about θ to be made while taking the added randomness from tuning λ into account. We will start by deriving limit results when $\hat{\lambda}$ is estimated by minimising the relatively simpler training error (Section 2) or an information criterion (Section 3). We also derive estimators for the limiting variance matrix of $\hat{\theta}(\hat{\lambda})$ (Section 2.2). In Section 4, we show that the results of Section 2 hold true also when the cross validated error is used to estimate λ . This will require us to take a closer look at the cross validation procedure, and in Section 4.1 we show a result proving exactly how far off the training error is from the cross-validated counterpart. In Section 5, we consider a particular situation where the theory simplifies greatly

and in Section 6, we go through a couple of applications. In particular, we will return to the above analysis of diabetes in Pima Indians in Section 6.1. Lastly, Section 7 contains some concluding remarks.

1.1 Related work

As far as we know, no results concerning the limiting distribution of $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set to the minimizer of some general risk estimator exists in literature. Arcones (2005) studies a similar situation, by deriving limit results for $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set to the minimizer of some estimator of the limiting variance of $\hat{\theta}(\lambda)$. This result is closely related to our theorems, but Arcones' framework only works for one-dimensional parameters θ and λ . Furthermore, $\hat{\lambda}$ cannot minimise a general loss function or a cross-validated estimator of the risk.

Our results are different from those of Arcones (2005) in yet another way, as his results are based on theory of stochastic processes, while we use the framework of Z-estimators. We reformulate the tuning procedure as a two-stage estimation problem and show that $\hat{\theta}(\hat{\lambda})$ can be written as components of a Z-estimator (see e.g. Van der Vaart (2000, Ch. 5)). Many of our proofs are therefore based upon and inspired by results for two-stage estimators, see e.g. Joe (2005), Ko and Hjort (2019) or Dæhlen and Hjort (2024), and bears similarities with the framework of generalized estimating equations used frequently in econometrics, see e.g. Hansen (1982), Hall and Inoue (2003) or for a more thorough introduction, Shao (2003, Ch. 5).

To apply the two-stage arguments also in the context of cross-validation, we derive sharp results characterizing the difference between the cross-validated estimator of the risk and the training error in Theorem 6. This theorem can be seen as an extension of the result given in Stone (1977), where the author writes that the cross-validated likelihood is asymptotically equivalent to the AIC. Stone (1977) only writes down an informal argument, however, and such intuitive explanations are often repeated when the paper is cited, see e.g. Claeskens et al. (2008) and Konishi and Kitagawa

(2008). To the best of our knowledge no formal proof of the statement has ever been given, and we therefore believe that our Theorem 6 fills this gap in the literature. Furthermore, our theorem is more general than that of Stone (1977) as it allows for general Z -estimators instead of only maximum likelihood estimators, and the functions used in the cross-validation procedure and estimation of θ need not be the same.

We also use Theorem 6 to discuss whether cross-validation is precise enough to capture the difference between different models, a topic which has been discussed by e.g. Shao (1993) or Shao (1997) in the context of linear models. Furthermore, our theory allows us to characterize the optimism of the training error, something which is studied in e.g. Efron (1983), Efron (1986), Tibshirani and Knight (1999) or Efron (2004), albeit with a different approach.

Our methodology is in some sense related to the field of post-selection inference, see e.g. Berk et al. (2013), Lee et al. (2016), Bachoc et al. (2020) or indeed Kuchibhotla et al. (2022) for a recent review. In this field, the distribution of regression parameters is derived in a way that takes the model selection procedure into account. This is typically done by either conditioning on the model selection procedure or by creating confidence intervals by taking the supremum over all candidate models. We, on the other hand, include the selection of continuous tuning parameters into the fitting procedure itself. Because of this, the field of post-selection inference is related to the present article in spirit only. Furthermore, we provide explicit forms and estimators for the limiting variance of $\hat{\theta}(\hat{\lambda})$, something which is not obtained through the most popular post-selection inference techniques (see e.g. the aforementioned review article).

Another topic which bears some similarities with our work is the research concerning asymptotic optimality of different parameter tuning schemes. For a given prediction depending on some tuning parameter λ , these works investigate its distance to the “best” prediction we can possibly make, when λ is tuned according to some criterion. Craven and Wahba (1978), Li (1985) and Speckman (1985), Li (1986) and Li (1987) all investigate this problem in linear models, and more recently Mu et al. (2018) and

Mu et al. (2021) have proven asymptotic optimality in a more advanced context. See also Mu et al. (2021) which contains a more comprehensive list of relevant literature. So far, however, the main concern of this field of research has been whether the tuning procedure ensures that the obtained and optimal predictions are close, rather than the asymptotic effect tuning parameters have on the limit distribution of $\hat{\theta}(\hat{\lambda})$. We believe this sets our results apart from the literature of asymptotic optimality.

1.2 Notation and definitions

Before we begin with the theory we will introduce notation which will be used throughout this article. For a given function $f(\theta)$ we will let $\nabla_{\theta=\theta_0}f(\theta)$ and $H_{\theta=\theta_0}f(\theta)$ be the gradient and Hessian matrix of f at θ_0 respectively, while $\partial_{\theta=\theta_0}f(\theta) = f'(\theta_0) = \partial f / \partial \theta(\theta_0)$ will denote the Jacobian matrix of f at θ_0 . When there is no room for misunderstanding, we will simplify notation and write $\nabla_{\theta_0}f$, $H_{\theta_0}f$ and $\partial_{\theta_0}f$ rather than the above. Furthermore, we will often identify matrices in $\mathbb{R}^{p \times q}$ with vectors in $\mathbb{R}^{pq}$ in the following way $A = (a_1, a_2, \dots, a_q) \in \mathbb{R}^{p \times q} \rightsquigarrow (a_1^T, a_2^T, \dots, a_q^T)^T \in \mathbb{R}^{pq}$ and hence, if A is a matrix, $\|A\| = (\sum_j \sum_k a_{jk}^2)^{1/2}$. We will also let $\xrightarrow{\text{Pr}}$ and $\xrightarrow{d}$ denote convergence in probability and distribution/law respectively.

We will let $Z_1, \dots, Z_n \in \mathbb{R}^d$ be independent and identically distributed (i.i.d.) variables from a distribution F and assume that for each fixed value of $\lambda \in \mathbb{R}^q$, $\theta \in \mathbb{R}^p$ is estimated by solving the equation $n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta, \lambda) = 0$, for some function $\varphi: \mathbb{R}^{d+p+q} \rightarrow \mathbb{R}^p$. We will let $\hat{\theta}(\lambda)$ denote the solution. If $\hat{\theta}(\lambda)$ is a maximizer of an expression on the form $n^{-1} \sum_{i=1}^n \phi(Z_i, \theta, \lambda)$ for some function ϕ , this is satisfied with $\varphi(z, \theta, \lambda) = \nabla_{\theta} \phi(z, \theta, \lambda)$ under general conditions. In particular, if $\hat{\theta}(\lambda)$ maximises a penalized likelihood on the form $n^{-1} \sum_{i=1}^n \log f(Y_i, \theta) - \lambda P(\theta)$, $\varphi(z, \theta, \lambda) = u(Y_i, \theta) - \lambda P'(\theta)$ where u is the score function in the model. We will also comment that regression and classification models fit into this framework by considering the covariates as random and letting $Z = (Y, X^T)^T$ where Y is the response and X the predictor.

Lastly, fix a loss function $\psi: \mathbb{R}^{d+p} \rightarrow \mathbb{R}$. The training error, which will be denoted by

$TE(\lambda)$, is equal to $n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}(\lambda)\}$. The leave-one-out cross validation (LOOCV) estimator of the risk is $n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}_{(-i)}(\lambda)\}$ and will be referenced by $CV(\lambda)$. We will assume that $\hat{\lambda}$ is set to the minimizer of some error estimator and refer to this process as “tuning” λ

2 Two-stage tuning

We will start by investigating the asymptotic properties of $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set by tuning with respect to the training error. If this function is convex in λ and the minimizer lies in the interior of the parameter space for λ , this is equivalent to solving $TE'(\lambda) = 0$. Hence, under these conditions, $\hat{\lambda}$ satisfies $TE'(\hat{\lambda}) = \sum_{i=1}^n \hat{\theta}'(\hat{\lambda})^T \nabla \psi[Z_i, \hat{\theta}(\hat{\lambda})]$. If $n^{-1} \sum_{i=1}^n \partial_{\hat{\theta}(\hat{\lambda})} \varphi(Z_i, \theta, \hat{\lambda})$ is invertible, the quantity $\hat{\theta}'(\hat{\lambda})$ exists by the implicit function theorem. Furthermore, it satisfies $n^{-1} \sum_{i=1}^n [\partial_{\hat{\theta}(\hat{\lambda})} \varphi(Z_i, \theta, \hat{\lambda}) \hat{\theta}'(\hat{\lambda}) + \partial_{\hat{\lambda}} \varphi\{Y_i, \hat{\theta}(\hat{\lambda}), \lambda\}] = 0$ by the same result. Because of these relations as well as the definition of $\hat{\theta}(\lambda)$, we can express the vector $\hat{\alpha} = (\hat{\theta}(\hat{\lambda})^T, \hat{\lambda}^T, \hat{\theta}'(\hat{\lambda})^T)^T \in \mathbb{R}^{p+q+pq}$ as the solution to a set of equations. This implies that the full vector $\hat{\alpha}$ is what is called a Z-estimator, which are typically consistent for well-defined quantities and asymptotically normal after proper scaling and centring, see e.g. Van der Vaart (2000, Ch. 5). Since $\hat{\alpha}$ is Z-estimator by the previous arguments, we would expect the same to be true for this estimator. The following theorem shows that this is indeed correct.

Theorem 1. *For each fixed λ , let $\hat{\theta}(\lambda)$ be the unique solution to $n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta, \lambda) = 0$. Furthermore, assume that $\hat{\lambda}$ is the unique solution to $TE'(\lambda) = o_p(1/\sqrt{n})$. Let $\alpha = (\theta^T, \lambda^T, D^T)^T \in \mathbb{R}^{p+q+pq}$ and $\hat{\alpha} = (\hat{\theta}(\hat{\lambda})^T, \hat{\lambda}^T, \hat{\theta}'(\hat{\lambda})^T)^T$. Define the function $\eta: \mathbb{R}^{d+p+q+pq} \rightarrow \mathbb{R}^{p+d+pq}$ as $\eta(z, \alpha) = (\eta_1(z, \alpha)^T, \eta_2(z, \alpha)^T, \eta_3(z, \alpha)^T)^T$ where $\eta_1(z, \alpha) = \varphi(z, \theta, \lambda)$, $\eta_2(z, \alpha) = D^T \nabla_{\theta} \psi(z, \theta)$ and $\eta_3(z, \alpha) = \partial_{\theta} \varphi(z, \theta, \lambda) D + \partial_{\lambda} \varphi(Z, \theta, \lambda)$. Assume the solution to $0 = \Psi(\alpha) = E\{\eta(Z, \theta, \lambda, D)\}$ is unique and equal to α_0 . Then $\hat{\alpha}$ is*

consistent for α_0 and furthermore,

$$\hat{\alpha} = \alpha_0 - n^{-1} \Psi'(\alpha_0)^{-1} \sum_{i=1}^n \eta(Z_i, \theta_0, \lambda_0, D_0) + o_p(1/\sqrt{n}), \quad (1)$$

provided the function Ψ is continuously differentiable in a neighbourhood of α_0 , that $\Psi'(\alpha_0)$ is non-singular and that in a neighbourhood of α_0 the norm of $\eta(z, \alpha)$ and all first order partial derivatives of η with respect to α are bounded by functions $m_0(z)$ and $m_1(z)$ respectively with $E m_0(Z)^2, E m_1(Z)^2 < \infty$. In particular, $\sqrt{n}(\hat{\alpha} - \alpha_0)$ converges in law to a $N(0, V_\alpha)$ distribution where $V_\alpha = \Psi'(\alpha_0)^{-1} \text{Var} \eta(Z, \alpha_0) \{\Psi'(\alpha_0)^{-1}\}^T$.

Proof. By the arguments preceding the theorem, $\sum_{i=1}^n \eta(Z_i, \hat{\alpha}) = 0$. Hence, consistency follows from the assumptions and Lemma S1.1 and Corollary S1.2 in the supplementary material. Equation (1) follows from Van der Vaart (2000, Th. 5.21). ■

In practice, we are rarely interested in the asymptotic behaviour of the full vector $\hat{\alpha}$, but rather in the limiting properties of $\hat{\theta}(\hat{\lambda})$ alone.

Corollary 1. Assume the conditions of Theorem 1 hold true, and for each λ let $\theta(\lambda)$ be solution to $E \varphi(Z, \theta, \lambda) = 0$. Then $\hat{\theta}(\hat{\lambda})$ is consistent for $\theta_0 = \theta(\lambda_0)$ where λ_0 is the solution to $\nabla_\lambda E \psi\{Z, \theta_0(\lambda)\} = 0$. Let $D = \theta'(\lambda_0)$, $Z_1 = H_{\lambda_0} E \psi\{Z, \theta_0(\lambda)\}$, $Z_2 = H_{\theta_0} E \psi(Z, \theta)$, $J = -\partial_{\theta_0} E \varphi(Z, \theta, \lambda_0)$ and $b = \nabla_{\theta_0} E \psi(Z, \theta)$, then

$$\begin{aligned} \hat{\theta}(\hat{\lambda}) = & \theta_0 + A_1 \frac{1}{n} \sum_{i=1}^n \varphi(Z_i, \theta_0, \lambda_0) + A_2 \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_0} \psi(Z_i, \theta) \\ & + A_3 \frac{1}{n} \sum_{i=1}^n \{\partial_{\theta_0} \varphi(Z_i, \theta, \lambda_0) D + \partial_{\lambda_0} \varphi(Z_i, \theta_0, \lambda)\}, \end{aligned} \quad (2)$$

where $A_1 = J^{-1} - D Z_1^{-1} \{D^T Z_2 + W\} J^{-1}$, $A_2 = -D Z_1^{-1} D^T$ and $A_3 = -D Z_1^{-1} M$, where M is a $q \times pq$ block diagonal matrix where the diagonal blocks are each equal to $b^T J^{-1}$ and W is a $q \times p$ matrix which can be written as $W = ((b^T J^{-1} W^1)^T, \dots, (b^T J^{-1} W^q)^T)^T$

with

$$W^j = \begin{pmatrix} D_j^T H_{\theta_0} \mathbb{E} \varphi^1(Z, \theta, \lambda_0) \\ D_j^T H_{\theta_0} \mathbb{E} \varphi^2(Z, \theta, \lambda_0) \\ \vdots \\ D_j^T H_{\theta_0} \mathbb{E} \varphi^p(Z, \theta, \lambda_0) \end{pmatrix} + \partial_{\lambda_0^j} \partial_{\theta_0} \mathbb{E} \varphi(Z, \theta, \lambda),$$

where D_j denotes the j -th column of D . In particular, $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0(\lambda_0)\} \rightarrow N\{0, A^* K^* (A^*)^T\}$ where $A^* = (A_1, A_2, A_3)$ and $K^* = \text{Var} \eta(Z, \theta_0, \lambda_0, D)$ with η defined in Theorem 1.

Proof. This follows from Theorem 1 and formulas for the inverse of block matrices. ■

If $q = 1$, we have $M = b^T J^{-1}$ and $W = b^T J^{-1} W^1$ making all formulas a lot more readable. Furthermore, the expressions above often simplify greatly, see e.g. Section 5.

2.1 Truncated estimators

In Corollary 1 we assume that $\hat{\lambda}$ is found by solving $\text{TE}'(\lambda) = 0$. This is, however, not how λ is usually tuned in practice. Typically, $\hat{\lambda}$ is set to the minimizer of $\text{TE}(\lambda)$ in some compact set Λ . We will now discuss how Corollary 1 can be used to make inference also in this situation. For simplicity, we will take $q = 1$ and $\Lambda = [0, 1]$, but all results should be generalizable to higher dimensions and other parameter sets.

As long as TE is differentiable and $\hat{\lambda}$ lies in the interior of $[0, 1]$, minimising TE in Λ is equivalent to solving $\text{TE}'(\lambda) = 0$. Oftentimes, however, $\text{TE}(\lambda)$ might be increasing or decreasing with $\text{TE}'(\lambda) \neq 0$ for all values of λ . In such cases $\hat{\lambda}$ is set to 0 or 1 depending on which of the boundary points achieves the lowest value of $\text{TE}(\lambda)$, and the condition $\text{TE}'(\hat{\lambda}) = 0$ might not be satisfied. To study this more complicated estimator, we will work with a modified version of $\hat{\theta}(\lambda)$ defined in the following way:

$$\hat{\theta}_T = \hat{\theta}(0)I(\lambda \leq 0) + \hat{\theta}(\lambda)I(0 < \lambda < 1) + \hat{\theta}(1)I(\lambda \geq 1), \quad (3)$$

where I is the indicator function. This estimator agrees with $\hat{\theta}(\lambda)$ for all $\lambda \in [0, 1]$. Outside of the unit interval, however, the estimator is “truncated” to either $\hat{\theta}(0)$ or $\hat{\theta}(1)$ depending on which side of $[0, 1]$ λ lies. Hence, as long as $TE(\lambda)$ is convex and has a global minimizer $\hat{\lambda}_G \in \mathbb{R}$, $\hat{\theta}_T(\hat{\lambda}_G)$ will be equal to $\hat{\theta}(\hat{\lambda})$ where $\hat{\lambda}$ is the minimizer of $TE(\lambda)$ in $[0, 1]$. Furthermore, when the global minimizer of $TE(\lambda)$ lies in $[0, 1]$, $\hat{\lambda}$ is the global minimizer ensuring $\hat{\theta}_T(\hat{\lambda}_G) = \hat{\theta}(\hat{\lambda})$.

Theorem 2. *Assume the conditions of Corollary 1 hold true, and let $\hat{\theta}_T(\lambda)$ be defined as in (3). Furthermore, let $\hat{\lambda}_G$ be the global minimizer of $TE(\lambda)$. The following hold:*

- (a) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 < 0$ and $\sqrt{n}\{\hat{\theta}(0) - \theta_0\} \xrightarrow{d} N$, where N is some normal distribution. Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}) - \theta_0\} \xrightarrow{d} N$.*
- (b) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 > 1$ and $\sqrt{n}\{\hat{\theta}(1) - \theta_0\} \xrightarrow{d} N$. Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0\} \xrightarrow{d} N$.*
- (c) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 = 0$ and that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}_G)^T - \theta_0^T, \hat{\theta}(0)^T - \theta_0^T, \hat{\lambda}\}^T \xrightarrow{d} (N_1^T, N_2^T, N_3)^T \sim N(0, V)$ for some matrix V . Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}) - \theta_0\} \xrightarrow{d} I(N_3 \geq 0)N_1 + I(N_3 < 0)N_2$.*
- (d) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 = 1$ and that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}_G)^T - \theta_0^T, \hat{\theta}(1)^T - \theta_0^T, \hat{\lambda} - 1\}^T \xrightarrow{d} (N_1^T, N_2^T, N_3)^T \sim N(0, V)$ for some matrix V . Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0\} \xrightarrow{d} I(N_3 \leq 0)N_1 + I(N_3 > 0)N_2$.*

Remark. From Z-estimation theory, $\sqrt{n}\{\hat{\theta}(\lambda) - \theta(\lambda)\} = J(\lambda)^{-1}n^{-1/2}\sum_{i=1}^n \varphi\{Z_i, \theta(\lambda), \lambda\} + o_p(1)$ where $\theta(\lambda)$ is the root of $E\varphi(Z, \theta, \lambda)$ and $J(\lambda) = -\partial_{\theta(\lambda)} E\varphi(Z, \theta, \lambda)$ for $\lambda = 0, 1$. Combining this with (2) and the central limit theorem shows the joint convergence required for case (c) and (d).

Proof. We will only prove (a) and (c). The remaining cases are similar. Since $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 < 0$ the probability of $\hat{\lambda}_G < 0$ tends to 1. Furthermore, conditioned on this event, $\hat{\theta}(\hat{\lambda}_G) = \hat{\theta}(0)$. Hence, $\Pr\{\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0\lambda_0\} \leq x\} = \lim_{n \rightarrow \infty} [\Pr\{\sqrt{n}\{\hat{\theta}(0) - \theta_0\} \leq$

$x\} \Pr(\hat{\lambda}_G < 0) + o(\Pr(\hat{\lambda}_G \geq 0))]$. The right hand side of this equation converges to $\Pr(N \leq x)$. This proves (a).

For (c) note that $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0(0)\} = I\{\sqrt{n}(\hat{\lambda}_G - 0) \geq 0\}\sqrt{n}\{\hat{\theta}(\hat{\lambda}_G) - \theta_0(0)\} + I\{\sqrt{n}(\hat{\lambda}_G - 0) < 0\}\sqrt{n}\{\hat{\theta}(0) - \theta_0(0)\}$. Now since the function $g(x, y, z) = I(z \geq 0)x + I(z < 0)y$ has the following set of discontinuity points $D_g = \{(x, y, 0) \mid x, y \in \mathbb{R}\}$, and this set has measure zero in the multivariate normal distribution, the continuous mapping theorem guarantees that $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0(0)\}$ converges in law towards $g(N_1, N_2, N_3) = I(N_3 \geq 0)N_1 + I(N_3 < 0)N_2$. This concludes the proof. $\blacksquare$

2.2 Estimators of the variance

Now that we know how $\hat{\theta}(\hat{\lambda})$ behaves asymptotically, we will define estimators of its limiting variance. We start by giving approximations to the quantities in Corollary 1.

Theorem 3. *Let $\hat{\lambda}$ be the minimizer of $TE(\lambda)$ and $\hat{\theta}(\hat{\lambda})$ the solution to $0 = \Phi_n(\theta, \hat{\lambda})$, where $\Phi_n(\theta, \lambda) = n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta, \lambda)$. Define the following estimators: $\hat{Z}_1 = n^{-1} \sum_{i=1}^n H_{\hat{\lambda}} \psi\{Z_i, \hat{\theta}(\hat{\lambda})\}$, $\hat{Z}_2 = H_{\hat{\theta}(\hat{\lambda})} \Psi_n(\theta)$ and $\hat{b} = \nabla_{\hat{\theta}(\hat{\lambda})} \Psi_n(\theta)$, where $\Psi_n(\theta) = n^{-1} \sum_{i=1}^n \psi(Z_i, \theta)$. Furthermore, let $\hat{J} = -\partial_{\hat{\theta}(\hat{\lambda})} \Phi_n(\theta, \hat{\lambda})$, $\hat{D} = \hat{J}^{-1} \partial_{\hat{\lambda}} \Phi_n\{\hat{\theta}(\hat{\lambda}), \lambda\}$, $\hat{K}^* = n^{-1} \sum_{i=1}^n \hat{\eta} \hat{\eta}^T$, where $\hat{\eta} = \eta\{z, \hat{\theta}(\hat{\lambda}), \hat{\lambda}, \hat{\theta}'(\hat{\lambda})\}$ and*

$$\hat{W}^j = \frac{1}{n} \sum_{i=1}^n \left\{ \begin{pmatrix} \hat{D}_j^T H_{\hat{\theta}} \varphi^1(Z_i, \theta, \hat{\lambda}) \\ \vdots \\ \hat{D}_j^T H_{\hat{\theta}} \varphi^p(Z_i, \theta, \hat{\lambda}) \end{pmatrix} + \partial_{\hat{\lambda}_j} \partial_{\hat{\theta}} \varphi(Z_i, \theta, \lambda) \right\} \text{ for } j = 1, \dots, q,$$

with η is as defined in Theorem 1. These estimators are consistent for the corresponding quantities, if there exists functions depending only on z that bound the norms of all partial derivatives of $\partial_{\theta} \varphi$, $\partial_{\lambda} \varphi$, $H_{\lambda} \psi$, $H_{\theta} \psi$, $\nabla_{\theta} \psi$, $(z, \theta, \lambda) \mapsto \eta(z, \theta, \lambda) \eta(z, \theta, \lambda)^T$, $H_{\theta} \varphi^j$ and $\partial_{\lambda_j} \partial_{\theta} \varphi$ for $j = 1, \dots, p$ with respect to $(\theta, \lambda, \theta')$ in a neighbourhood of $(\theta_0^T, \lambda_0^T, D^T)^T$.

Proof. This follows from standard theory, see e.g. Dæhlen et al. (2024, Le. 4, Ap. B). $\blacksquare$

We are now ready to define estimators of the limiting variance of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0(\lambda_0)\}$.

Theorem 4. Assume we are minimising $TE(\lambda)$ in $[0,1]$ and the conditions of Theorem 3 and Corollary 1 hold true. Then, $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ is asymptotically normal with variance V which can be estimated consistently by $\hat{V}_1 = \hat{A}^* \hat{K}^* (\hat{A}^*)^T$ if $\hat{\lambda}$ lies in the interior of $[0,1]$, or $\hat{V}_2 = \hat{J}^{-1} \hat{K} (\hat{J}^{-1})^T$ when TE is minimised in λ with $TE'(\lambda) \neq 0$ for λ equal to 0 or 1. Here $\hat{K} = n^{-1} \sum_{i=1}^n \varphi\{Z_i, \hat{\theta}(\hat{\lambda}), \hat{\lambda}\}$, $\hat{A}^* = (\hat{A}_1, \hat{A}_2, \hat{A}_3)$ with $\hat{A}_1 = \hat{J}^{-1} - \hat{D} \hat{Z}_1^{-1} \{\hat{D}^T \hat{Z}_2 + \hat{W}\} \hat{J}^{-1}$, $\hat{A}_2 = -\hat{D} \hat{Z}_1 \hat{D}^T$ and $\hat{A}_3 = -\hat{D} \hat{Z}_1^{-1} \hat{M}$, where $\hat{M}$ is a block diagonal matrix with $\hat{b}^T \hat{J}^{-1}$ on each of its entries and $\hat{W} = ((\hat{b}^T \hat{J}^{-1} \hat{W}^1)^T, \dots, (\hat{b}^T \hat{J}^{-1} \hat{W}^q)^T)^T$.

Proof. This follows directly from Corollary 1 and Theorem 2, in conjunction with Theorem 3 and the continuous mapping theorem. ■

It could of course happen that λ_0 is exactly equal to either of the endpoints of Λ . Then, case (c) or (d) of Theorem 2 applies and $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ is not guaranteed to be asymptotically normal, so that approximating the it by $N(0, \hat{V}_1)$ or $N(0, \hat{V}_2)$ will not be correct. In Section 5 we will, however, see that the distribution in cases (c) and (d) often simplifies to a central multivariate normal distribution also in this case, and that the variance can be estimated by either $\hat{V}_1$ or $\hat{V}_2$, defined in the preceding theorem.

3 Information criteria

The theory presented in the previous sections is not limited to cases where $\hat{\lambda}$ is set to the minimizer of the training error. The results also hold true when the tuning parameter minimises quantities which are “almost” equal to TE. The most obvious example of this is perhaps when $\hat{\lambda}$ is set to the minimizer of an information criterion.

Let $f_\lambda(z, \theta)$ be some parametric model and $\ell_{\lambda,n}(\theta)$ the corresponding likelihood. Furthermore, for each λ , let $\hat{\theta}(\lambda)$ be the maximizer of $\ell_{\lambda,n}(\theta)$. Then by definition Akaike’s information criterion (AIC), the Bayesian information criterion (BIC) and Takeuchi’s information criterion (TIC) take the following forms for each fixed λ : $AIC(\lambda) = -n^{-1} \ell_{\lambda,n}\{\hat{\theta}(\lambda)\} + n^{-1}p$, $BIC(\lambda) = -n^{-1} \ell_{\lambda,n}\{\hat{\theta}(\lambda)\} + n^{-1}p \log n$ and $TIC(\lambda) =$

$-n^{-1}\ell_{\lambda,n}\{\hat{\theta}(\lambda)\} + n^{-1}\text{Tr}\{\hat{J}(\hat{\theta}, \lambda)^{-1} \hat{K}(\hat{\theta}, \lambda)\}$, where p is the dimension of θ , the number of parameters in the model $f_\lambda(z, \theta)$ at a fixed λ . Hence, provided $\hat{J}$ and $\hat{K}$ are sufficiently regular, the above information criteria are all only $o_p(1/\sqrt{n})$ away from $n^{-1}\ell_{\lambda,n}\{\hat{\theta}(\lambda)\}$, allowing Theorem 1 to be applied after replacing φ by $\nabla_\theta \log f_\lambda(z, \theta)$ and $\eta_2(z, \alpha)$ by $(z, \alpha) \mapsto \partial_\theta \log f_\lambda(z, \theta)D + \partial_\lambda \log f_\lambda(z, \theta)$. Corollary 1 and Theorem 2 also hold after making similar modifications.

The fact that the results of this article can be applied also when λ is tuned by minimising the AIC, BIC or TIC follows more or less directly from the forms of the criteria. We will now discuss the focused information criterion (FIC), a newer criterion for which the result is less immediate. The FIC was introduced in Claeskens and Hjort (2003), and is an information criterion with a slightly different aim than the classic criteria discussed above. Rather than ranking models according to how well they fit the data overall, the FIC evaluates models by how well they estimate some pre-specified parameter of interest. This parameter of interest is called the focus parameter and should be chosen to reflect the main goal of an analysis. If, for instance, the goal is to estimate the median in a population, the FIC prefers models for which the estimator of the median is good. The quality of an estimator is evaluated by its mean squared error (MSE), and hence, the FIC of a model is the estimated MSE of that model's estimator of the focus parameter. We will work with the relatively newer version of the criterion introduced in Jullum and Hjort (2017) and modified in Dæhlen et al. (2024).

As before, let $f_\lambda(z, \theta)$ be some parametric model and $\hat{\theta}(\lambda)$ the maximizer of $\theta \mapsto \ell_{\lambda,n}(\theta) = \sum_{i=1}^n \log f_\lambda(Z_i, \theta)$ for each λ . Let μ be the focus parameter with true value μ_0 and assume μ takes the value $\mu_\lambda(\theta)$ in the model $f_\lambda(z, \theta)$. The FIC of the model $f_\lambda(z, \theta)$ for each λ is the MSE of $\mu_\lambda\{\hat{\theta}(\lambda)\}$. Dæhlen et al. (2024) gives an estimator of this quantity: $\text{FIC}(\lambda) = \hat{b}_\lambda^2 + n^{-1}(2\hat{b}_\lambda\hat{c}_\lambda - \hat{\kappa}_\lambda + \hat{\tau}_\lambda)$, where $\hat{c}_\lambda$, $\hat{\kappa}_\lambda$ and $\hat{\tau}_\lambda$ are consistent estimators of certain population quantities defined in the paper, and $\hat{b}_\lambda = \mu\{\hat{\theta}(\lambda)\} - \hat{\mu}_0$ for some estimator $\hat{\mu}_0$, which is assumed to be consistent for μ_0 . Assume now that $\hat{\mu}_0$ is the solution to an equation on the form $0 = n^{-1} \sum_{i=1}^n \xi(Z_i, \mu)$. We

then have $\text{FIC}(\lambda) = n^{-1} \sum_{i=1}^n [\mu\{\hat{\theta}(\lambda)\} - \hat{\mu}_0]^2 + O_p(1/n)$, and hence, provided sufficient regularity, Theorem 1 can be applied by replacing $\theta(\lambda)$ by $(\theta(\lambda)^T, \mu)^T$, $\varphi(z, \theta, \lambda)$ by $(\varphi(z, \theta, \lambda)^T, \xi(z, \mu_0))^T$ and $\eta_2(z, \alpha)$ by $(\theta, \mu, \lambda, D) \mapsto 2\{\mu(\theta) - \mu\}(D^T \nabla_{\theta} \mu(\theta)^T, -1)^T$. In Section 6.2, we illustrate this in an example.

Remark. It is worth noting that we require the criterion used for fitting λ to be differentiable as a function of the tuning parameter. Because of this, our theorems do not cover situations where λ is discrete. In particular, our theorems cannot be applied when an information criterion is used to e.g. choose between a finite number of models.

4 Cross-validation

When λ is tuned by minimising the training error, $\hat{\theta}(\hat{\lambda})$ can be written as the components of a Z-estimator. Because of this the results of the previous section follow more or less directly from standard theory. In practice, however, λ is seldom tuned with respect to TE, but instead set to the minimizer of $\text{CV}(\lambda)$, the estimate of the risk function obtained by cross-validation. In this section, we will show that the difference between tuning with respect to TE and CV is negligible asymptotically and that all of the previously derived results can be applied also when $\hat{\lambda}$ is the minimizer of CV.

First, note that for each λ the vector $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T$ is the solution to equation $n^{-1} \sum_{i=1}^n \xi(Z_i, \theta, \lambda, D) = 0$ with $\xi(z, \theta, \lambda, D) = (\xi_1(z, \theta, \lambda)^T, \xi_2(z, \theta, \lambda, D)^T)^T$ where $\xi_1 = \varphi$ and $\xi_2(z, \theta, \lambda, D) = \partial_{\theta} \varphi(z, \theta, \lambda) D + \partial_{\lambda} \varphi(z, \theta, \lambda)$. This follows from definition of $\hat{\theta}(\lambda)$ and the implicit function theorem. Because of this, the vector $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T$ is a Z-estimator, and under weak conditions $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T$ is consistent for some vector $(\theta_0(\lambda)^T, D_0(\lambda)^T)^T$ and $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T \approx (\theta_0(\lambda)^T, D_0(\lambda)^T)^T + V(\lambda)^{-1} n^{-1} \sum_{i=1}^n \xi\{Z_i, \theta_0(\lambda), D_0(\lambda), \lambda\}$ where $V(\lambda) = -\partial_{\theta_0(\lambda), D_0(\lambda)} E \xi(Z, \theta, \theta', \lambda)$. See e.g. chapter 5 of Van der Vaart (2000) for proofs and sufficient conditions. Because of this $(\hat{\theta}(\lambda)^T - \hat{\theta}_{(-i)}(\lambda)^T, \hat{\theta}'(\lambda)^T - \hat{\theta}'_{(-i)}(\lambda)^T) \approx n^{-1} V(\lambda)^{-1} \xi\{Z_i, \theta_0(\lambda), \theta'_0(\lambda)\} = O_p(1/n)$. We formalize this idea in the following lemma.

Lemma 1. *Let the parameter sets Θ for θ , $\mathcal{D}$ for $\hat{\theta}'(\lambda)$ and Λ for λ be compact and $\xi(z, \theta, D, \lambda) = (\varphi(z, \theta, \lambda)^T, \xi_2(z, \theta, \lambda, D)^T)$ where $\xi_2(z, \theta, \lambda, D) = \partial_\theta \varphi(z, \theta, \lambda)D + \partial_\lambda \varphi(z, \theta, \lambda)$. Assume the conditions of Lemma S1.3 in the supplementary material hold true for this ξ , and in particular, that $m_0(z)$ bounds $\xi(z, \theta, D, \lambda)$ for all θ , D and λ . Then $\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\|, \|\hat{\theta}'_{(-i)}(\lambda) - \hat{\theta}'(\lambda)\| \leq a_{0,n}m_0(Z_i)$ where $a_{0,n} = O_p(1/n)$ uniformly in λ and does not depend on i .*

The proof is given in the supplementary material. With Lemma 1, we are ready to quantify the difference between $CV'(\lambda)$ and $TE'(\lambda)$.

Lemma 2. *Assume all conditions of Lemma 1 hold true with $E m_0(Z)^4 < \infty$. Then*

$$\frac{1}{n} \sum_{i=1}^n \hat{\theta}'_{(-i)}(\lambda)^T \nabla_{\hat{\theta}_{(-i)}(\lambda)} \psi(Z_i, \theta) = \frac{1}{n} \sum_{i=1}^n \hat{\theta}'(\lambda)^T \nabla_{\hat{\theta}(\lambda)} \psi(Z_i, \theta) + \delta_n(\lambda), \quad (4)$$

where $\|\delta_n(\lambda)\| = O_p(1/n)$ uniformly in λ , provided there exist square integrable functions p_1 and p_2 not depending on θ that bound respectively all first and second order partial derivatives of ψ with respect to θ .

Proof. By Lemma 1, $\hat{\theta}_{(-i)}(\lambda) = \hat{\theta}(\lambda) + R_n^i(\lambda)$ and $\hat{\theta}'_{(-i)}(\lambda) = \hat{\theta}'(\lambda) + S_n^i(\lambda)$ with $\|R_n^i(\lambda)\|, \|S_n^i(\lambda)\| \leq a_{0,n}m_0(Z_i)$. A Taylor expansion reveals that the left hand side of (4) is equal to $n^{-1} \sum_{i=1}^n \{\hat{\theta}'(\lambda) + S_n^i(\lambda)\}^T \{\nabla_{\hat{\theta}(\lambda)} \psi(Z_i, \theta) + H_{\theta_i^*} \psi(Z_i, \theta) R_n^i(\lambda)\}$, for some θ_i^* s on the line segments between $\hat{\theta}(\lambda)$ and $\hat{\theta}_{(-i)}(\lambda)$. Utilizing the existence of p_1 and p_2 , we get that $\|CV'(\lambda) - TE'(\lambda)\|$ is bounded by $a_{0,n} \hat{\theta}'(\lambda)^T n^{-1} \sum_{i=1}^n p_2(Z_i) m_0(Z_i) + a_{n,0} n^{-1} \sum_{i=1}^n p_1(Z_i) m_0(Z_i) + a_{0,n}^2 n^{-1} \sum_{i=1}^n p_2(Z_i) m_0(Z_i)^2$. Since $E p_1(Z)^2$, $E p_2(Z)^2$ and $E m_0(Z)^4$ exist, the means converge. This ensures $\|CV'(\lambda) - TE'(\lambda)\| = O_p(1/n)$ uniformly in λ as $\mathcal{D}$ is compact. $\blacksquare$

Since $CV'(\lambda)$ and $TE'(\lambda)$ are uniformly close as functions of λ by the above lemma, we would expect $\hat{\theta}(\hat{\lambda})$ to asymptotically behave similarly when tuning with respect to CV and TE. The following theorem confirms the intuition.

Theorem 5. Let λ_{TE} be the minimizer of TE and λ_{CV} the minimizer of CV and assume that the conditions of Theorem 2 and Lemma 2 hold true. Then $\sqrt{n}\{\hat{\theta}(\lambda_{\text{TE}}) - \theta_0\}$ and $\sqrt{n}\{\hat{\theta}(\lambda_{\text{CV}}) - \theta_0\}$ are asymptotically equivalent. In particular, Theorem 1, Corollary 1 and Theorem 2 hold true for $\hat{\theta}(\hat{\lambda}_{\text{CV}})$, and the limiting distribution of $\sqrt{n}\{\hat{\theta}(\lambda_{\text{CV}}) - \theta_0\}$ can be estimated as in Theorem 4.

Proof. Let $\hat{\alpha}_{\text{CV}} = (\hat{\theta}(\hat{\lambda}_{\text{CV}})^T, \hat{\lambda}_{\text{CV}}^T, \hat{\theta}'(\hat{\lambda}_{\text{CV}})^T)^T$. Since $\text{TE}'(\lambda) = \text{CV}'(\lambda) + O_p(1/n)$ uniformly in λ , $\text{TE}'(\hat{\lambda}_{\text{CV}}) = O_p(1/n)$. Hence, $n^{-1} \sum_{i=1}^n \eta(Z_i, \hat{\alpha}_{\text{CV}}) = O_p(1/n)$. Since $n^{-1} \sum_{i=1}^n \eta(Z_i, \hat{\alpha}_{\text{CV}}) = o_p(1/\sqrt{n})$ is all that is needed for Van der Vaart (2000, Th. 5.21), the conclusion of Theorem 1 follows for $\hat{\alpha}_{\text{CV}}$ as well. This ensures that $\sqrt{n}(\hat{\alpha} - \alpha_0)$ and $\sqrt{n}(\hat{\alpha}_{\text{CV}} - \alpha_0)$ are asymptotically equivalent. Theorem 5 follows from this fact. ■

4.1 A more precise approximation

Theorem 5 is a somewhat surprising result, as it tells us that asymptotically speaking it makes no difference if we tune with respect to TE or CV. Yet, in most applications, minimising TE and CV gives very different results. We will now take a closer look at what the difference between CV and TE really is. To make the derivations more readable, we will only show pointwise results for $\hat{\theta}(\lambda)$, and drop λ from the notation.

Arguing similarly as in the paragraph preceding Lemma 1 we can show $\hat{\theta}_{(-i)} - \hat{\theta} \approx -n^{-1} J^{-1} \varphi(Z_i, \theta_0)$ where θ_0 is the solution to $E \varphi(Z, \theta) = 0$ and $J = -\partial_{\theta_0} E \varphi(Z, \theta)$. This approximation is not particularly surprising. The function $z \mapsto J^{-1} \varphi\{z, \theta_0\}$ is called the influence function of $\hat{\theta}$ as it in some sense tells how much a single data point affects estimation of θ . See e.g. Huber (2009) for more details on influence functions. Since $\hat{\theta}_{(-i)}$ is the estimator you get instead of $\hat{\theta}$ if you remove the i -th data point, it is natural that the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$, is precisely the “influence” of Z_i .

Assuming ψ is sufficiently smooth in θ , a Taylor expansion shows $n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) + n^{-1} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta})^T \nabla_{\hat{\theta}} \psi(Z_i, \theta) \approx \text{CV}$. A further Taylor expansion of $\nabla_{\hat{\theta}} \psi(Z_i, \theta)$ around θ_0 , reveals $\text{CV} \approx n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) - n^{-1} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta}) \nabla \psi(Z_i, \theta_0) \approx n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) -$

$n^{-2} \sum_{i=1}^n \varphi(Z_i, \theta_0)^T (J^{-1})^T \nabla_{\theta_0} \psi(Z_i, \theta)$. By the law of large numbers and properties of the trace operator, we hence get $CV \approx TE - n^{-1} \text{Tr}(J^{-1}C)$ where $C = E \varphi(Z_i, \theta_0) \nabla_{\theta_0} \psi(Z, \theta_0)^T$. The following theorem confirms these informal arguments. A full proof is given in the supplementary material.

Theorem 6. *Let $\hat{\theta}$ and $\hat{\theta}_{(-i)}$ the solutions to $\sum_{j=1}^n \varphi(Z_j, \theta) = 0$ and $\sum_{j \neq i}^n \varphi(Z_j, \theta)$ respectively. Assume further that $\hat{\theta}, \hat{\theta}_{(-i)} \in \Theta \subseteq \mathbb{R}^p$ where Θ is compact and that there exists functions m_0, m_1, p_1 and p_2 bounding the norms of φ , all partial derivatives of φ with respect to θ , ψ and all partial derivatives of ψ with respect to θ , respectively. Then, for a given function $\psi(z, \theta)$, we have*

$$n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) - n^{-1} \text{Tr}(J^{-1}C) + o_p(1/n) \quad (5)$$

where $J = -\partial_{\theta_0} E \varphi(Z, \theta)$ and $C = E \varphi(Z, \theta_0) \nabla_{\theta_0} \psi(Z, \theta)^T$, with θ_0 being the solution to $E \varphi(Z, \theta) = 0$, provided the eigenvalues of $\{\partial_{\theta} E \varphi(Z, \theta)\}^{-1}$ are bounded in Θ and $E m_0(Z)^4, E m_1(Z)^4, E p_1(Z)^4, E p_2(Z)^2 < \infty$.

When $\varphi(z, \theta) = \nabla_{\theta} \log f(z, \theta)$ for some density or probability mass function f , J is the Fisher matrix in the model. Furthermore, $C = K$ if $\psi = \log f$, where $K = \text{Var} \varphi(Z, \theta_0)$. Hence, Theorem 6 simplifies to $\sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = \text{TIC} + o_p(1)$, where TIC is Takeuchi's information criterion (TIC), defined in Section 3. Additionally, if the model f is specified correctly, $K = J$ and $\sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = \ell_n(\hat{\theta}) - p + o_p(1)$. The right hand side of this equation is easily recognized as Akaike's information criterion (AIC), leading Theorem 6 to simplify to the results of Stone (1977) when $\varphi(z, \theta) = \nabla_{\theta} \log f(z, \theta)$ and the parametric model is specified correctly.

Theorem 6 shows that CV and TE are more similar than one might initially think. In fact, the only thing separating the two quantities is a term of order $O_p(1/n)$, which is dominated by the size of TE itself, asymptotically speaking. In smaller samples, however, $n^{-1} \text{Tr}(J^{-1}C)$ might be quite large and will ensure that tuning with respect to TE and CV give different results. That being said, there are cases where $n^{-1} \text{Tr}(J^{-1}C)$

is not negligible and the difference between TE and CV will matter also in the limit. We will now turn our attention towards such a case.

4.2 When more precision is needed

For each parameter θ let $R(\theta)$ be some risk function, and let $\hat{\theta}$ be some estimator of θ based on data $Z_1, \dots, Z_n$. Assume further that $\hat{\theta} \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ for some matrix V . By Taylor expanding ψ around θ_0 and using the law of total expectation, one can show that, provided sufficient regularity,

$$E\{R(\hat{\theta})\} = R(\theta_0) + n^{-1}[\nabla_{\theta_0} R(\theta_0)c + (1/2) \text{Tr}\{H_{\theta_0} R(\theta_0)V\}] + o(1/n), \quad (6)$$

where c is equal to the limit of $n E(\hat{\theta} - \theta_0)$, see Dæhlen et al. (2024) for more on this parameter. If $\hat{\theta} \in \mathbb{R}$ and $R(\theta) = (\theta_0 - \theta)^2$, (6) reduces to the classical bias-variance decomposition of the MSE. With $R(\theta)$ equal to the error rate of certain classification procedures, (6) reduces to the expressions given in O'Neill (1980).

Note that (6) consists of three parts: a constant part, $R(\theta_0)$, a term of size $O(1/n)$ and a remainder term of smaller order. For two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ consistent for different values θ_0^1 and θ_0^2 respectively, the difference between their expected risk will asymptotically be dominated by how much $R(\theta_0^1)$ and $R(\theta_0^2)$ differ, making the $O(1/n)$ terms in (6) negligible asymptotically. If, on the other hand θ_0^1 and θ_0^2 are equal, the difference in expected risk for $\hat{\theta}_1$ and $\hat{\theta}_2$ will be dominated by the difference between the two corresponding terms of size $O(1/n)$ in (6). In this latter situation, we therefore need to estimate the $O(1/n)$ -effect from data in order to properly distinguish the risk of $\hat{\theta}_1$ from that of $\hat{\theta}_2$. Both TE and CV attempt to estimate $E R(\hat{\theta})$, but, as we will see, neither estimator is precise enough capture the $O(1/n)$ effects.

Let $R_n(\theta) = n^{-1} \sum_{i=1}^n \psi(Z_i, \theta)$ and $\hat{\theta}$ be some estimator consistent for θ_0 . Assuming

ψ is sufficiently smooth, a Taylor expansion shows

$$\text{TE} = R_n(\theta_0) + (\hat{\theta} - \theta_0)^T \nabla_{\theta_0} R_n(\theta) + (1/2)(\hat{\theta} - \theta_0)^T H_{\theta_0} R_n(\theta)(\hat{\theta} - \theta_0) + o_p(1/n). \quad (7)$$

Since $R_n(\theta_0)$ is a mean and $\hat{\theta} \xrightarrow{P} \theta_0$, we have $\text{TE} \xrightarrow{P} R(\theta_0)$, which means that TE does a good job, as far as estimating the constant part of (6) goes. Approximating the term of order $O(1/n)$ is, however, another story. The most obvious reason is perhaps that the remaining terms are nowhere near being consistent for $n^{-1}[\nabla_{\theta_0} R(\theta_0)c + (1/2) \text{Tr}\{H_{\theta_0} R(\theta)V\}]$, but even if they were, TE would not be precise enough to properly catch effects of size $O(1/n)$, as $R_n(\theta_0) = R(\theta_0) + O_p(1/\sqrt{n})$ by the central limit theorem. Since a term of order $O_p(1/\sqrt{n})$ will dominate $O(1/n)$ effects, the $O(1/n)$ -terms in (6) are too small to be captured by TE and will be washed out by the error of $R_n(\theta_0)$. As a result, TE is not well suited for evaluating estimators whose limiting risk differs only in the $O(1/n)$ -term. One might hope that this is a defect of the TE only, and that CV will be precise enough to capture the $O(1/n)$ effects. Sadly, Theorem 6 stops this idea dead in its tracks. Since the difference between CV and TE is a term of order $O_p(1/n)$, the added precision of CV is not sufficient to correct for the $O_p(1/\sqrt{n})$ error TE makes. This is pointed out in the context of model selection of linear regression models in Shao (1993).

The above might lead one to believe that the difference between CV and TE is negligible and that there is no reason to use CV rather than TE in analysis. This conclusion is, however, slightly too negative, as there is an important difference between CV and TE in expected value. To see this, assume that $E\psi(Z, \theta_0) = 0$ and that $\hat{\theta} = \theta_0 + J^{-1}n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta_0) + o_p(1/\sqrt{n})$. Then, by the central limit theorem, $\sqrt{n}(\hat{\theta} - \theta_0)$ and $\sqrt{n}\{R_n(\theta_0) - R(\theta_0)\}$ converge jointly in law to a central normal distribution with variance matrix

$$\Sigma = \begin{pmatrix} J^{-1}K(J^{-1})^T & J^{-1}C^T \\ C^T(J^{-1})^T & \text{Var} \psi(Z, \theta_0) \end{pmatrix},$$

where J and C are as in Theorem 6 and $K = \text{Var } \varphi(Z, \theta_0)$. Hence, provided all necessary quantities are uniformly integrable (see e.g. Billingsley (1999, p. 31)), $E(\text{TE}) = R(\theta_0) + \text{Tr}(CJ^{-1}) + (1/2) \text{Tr}\{H_{\theta_0}R(\theta)J^{-1}K(J^{-1})^T\} + o(1/n)$ which misses $E R(\hat{\theta})$ by a term $n^{-1} \text{Tr}(J^{-1}C) + o(1/n)$. Because of this, TE is a biased estimator of $E R(\hat{\theta})$. CV, on the other hand, satisfies $\text{CV} = \text{TE} - \text{Tr}(J^{-1}C) + o(1/n)$ by Theorem 6. Hence, arguing similarly as for TE, we get $E \text{CV} = E R(\hat{\theta}) + o(1/n)$, ensuring that in expected value, CV is able to separate between estimators whose risk differ only in the $O(1/n)$ -term. We stress that this holds in expected value only, and that the above shows nothing more than the fact that the bias of CV is of order $o(1/n)$. As noted before, the error of CV will be of order $O_p(1/\sqrt{n})$ for any single sample, a term which is too large to properly capture effects of size $O(1/n)$.

The discussion given in the previous paragraphs has consequences for our theorems concerning the limiting behaviour of $\hat{\theta}(\hat{\lambda})$. By the arguments above, neither TE nor CV is able to separate estimators whose risk differs only in the $O(1/n)$ term. Hence, when tuning λ with respect to CV or TE, we should not expect that the criteria are able to identify the optimal value of λ in this $O(1/n)$ -regime. This also explains why tuning with respect to TE is asymptotically equivalent to tuning with respect to CV.

The fact that the theory breaks down when $\theta_0(\lambda)$, the solution to $E \varphi(Z, \theta, \lambda) = 0$, is constant as a function of λ , can also be seen from Theorem 1. When $\theta_0(\lambda)$ is constant as a function of λ , the equation $\Psi(\alpha) = 0$ has multiple solutions and $\Psi'(\alpha_0)$ will be singular. Because of this, the conditions of Theorem 1 do not hold true when $\theta_0(\lambda)$ is constant, making the theorem inapplicable in this case. This does not pose a problem for regularized estimators like in ridge regression, as $\theta_0(\lambda)$ will rarely be constant in this case, but for hybrid estimators, this can indeed be quite problematic. Take for instance the hybrid generative-discriminative classification model where $\hat{\theta}(\lambda)$ is set to the maximizer of the following expression $n^{-1} \sum_{i=1}^n \{\lambda \log f_{Y|X}(Y_i, \theta | X = X_i) + (1 - \lambda) \log f_{(Y,X)}(Y_i, X_i, \theta)\}$ for some parametric model $f_{(Y,X)}$ for (Y, X) and where $f_{Y|X}$ is derived from $f_{(Y,X)}$. In this case, $\hat{\theta}(\lambda)$ aims for the maximizer of

$\Gamma_\lambda(\theta) = \lambda \mathbb{E} \log f_{(Y,X)}(Y, X, \theta) + (1 - \lambda) \mathbb{E} \log f_{Y|X}(Y, X, \theta)$. If the model is misspecified, $\theta_0(\lambda)$ will rarely be constant as a function of λ . If, on the other hand, the model does hold and the true underlying distribution is $f_{(Y,X)}(\cdot, \theta_0)$ for some $\theta_0 \in \mathbb{R}^p$, $\theta_0(\lambda)$ will be constantly equal to θ_0 . Because of this, the difference in expected limiting risk will be on the order of $O(1/n)$ when the model is specified correctly and $O(1)$ otherwise. Hence, Theorem 2 and Theorem 5 can be applied in this case only when the model is misspecified. See Section S2 in the supplementary material for more discussions.

5 When the effect of tuning disappears

We will now go through a situation in which the expressions in Corollary 1 simplify greatly and the effect of the tuning procedure becomes negligible asymptotically.

Assume there exists a λ_1 such that $\varphi(y, \theta, \lambda_1) = \nabla_\theta \psi(y, \theta)$ for all θ and F -almost all y . Then, $\theta_0(\lambda_1)$ solves $\mathbb{E} \varphi(Z, \theta, \lambda_1) = \nabla_\theta \mathbb{E} \psi(Z, \theta) = 0$. Because of this, λ_0 , b and Z_1 defined in Corollary 1 are equal to λ_1 , 0 and J , respectively. Entering this into the expressions in the corollary, shows that the right hand side of (2) is equal to $n^{-1} J^{-1} \varphi(z, \theta_0, \lambda_1)$. This is the influence function of $\hat{\theta}(\lambda_1)$. Hence, $\sqrt{n}(\hat{\theta}(\hat{\lambda}) - \theta_0) = \sqrt{n}\{\hat{\theta}(\lambda_1) - \theta_0\} + o_p(1)$ in this case, and the additional randomness introduced by tuning λ vanishes asymptotically. In practice, this means that one can ignore the effect of tuning λ and use the classic approximation $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\} \xrightarrow{d} N\{0, J^{-1} K (J^{-1})^T\}$, where $J = -\partial_{\theta_0} \mathbb{E} \varphi(Z, \theta, \lambda_1)$ and $K = \mathbb{E} \varphi(Z, \theta_0, \lambda_1) \varphi(Z, \theta_0, \lambda_1)^T$.

A λ_1 such that $\nabla_\theta \psi(z, \theta) = \varphi(z, \theta, \lambda_1)$ actually exists in quite a few situations. In for instance ridge regression where λ is set to the minimizer of estimated MSE, we have $\varphi(z, \beta, \lambda) = \nabla_\beta \{(y - \beta_1^T x - \beta_0)^2 + \lambda \|\beta_{1:}\|^2\}$, where $\beta = (\beta_0, \beta_{1:}^T)^T$, ensuring $\nabla_\beta \psi(z, \beta) = \varphi(z, \beta, 0)$. More generally, for any model $Y_i = g(X_i, \theta) + \epsilon_i$ fitted by minimising $\sum_{i=1}^n \{Y_i - g(X_i, \theta)\}^2 + \lambda \|\theta\|^2$ and tuned according to the MSE, $\nabla_\theta \psi(z, \theta) = \varphi(z, \theta, 0)$. Even more generally, for a model fitted by minimising $\sum_{i=1}^n h\{Y_i, g(X_i, \theta)\} + \lambda \|\theta\|^2$ for some function h , and where λ is tuned by minimising the CV or TE

estimate of $E h[Y, g\{X, \hat{\theta}(\lambda)\}]$, $\nabla_{\theta}\psi(z, \theta) = \varphi(z, \theta, 0)$, ensuring that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ and $\sqrt{n}\{\hat{\theta}(0) - \theta_0\}$ are asymptotically equivalent. This includes many regularized likelihood models when cross validated Kullback-Leibler divergence is used to tune λ .

The above arguments might lead one to believe that there rarely is need to use Theorem 2 or Theorem 5 in practice. This is, however, not correct. To illustrate the potentially large effect tuning of λ can have on the limit distribution of $\hat{\theta}(\hat{\lambda})$, we performed a small simulation experiment. For a selection of λ -values, we fitted logistic regression in two variables with a ridge penalty term. This corresponds to setting φ equal to the gradient of $y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) - \log(1 + \beta_0 + \beta_1 x_1 + \beta_2 x_2) - \lambda(\beta_1^2 + \beta_2^2)$ with respect to $\beta = (\beta_0, \beta_1, \beta_2)$. We chose $\hat{\lambda}$ by minimising $CV(\lambda)$ with $\psi(z, \theta) = [\exp(\beta_0 + \beta_1 x_1) / \{1 + \exp(\beta_0 + \beta_1 x_1)\} - y]^2$. For the true underlying distribution of the data, we chose the following. The binary variable Y was set to be Bernoulli distributed with equal probability for zero and one. Furthermore, for a selection of C -values, we used $X | Y = 0 \sim N(\mu, \Sigma)$ and $X | Y = 1 \sim N(-\mu, 2\Sigma)$ with $\mu = (1/2, 1/2)^T$ and Σ equal to the matrix with diagonal elements 2 and off-diagonal entries $-\sqrt{C/2}$.

For each selected C -value, we simulated $n = 100$ data points and computed $\hat{\lambda}_C$ and $\hat{\beta}_C(\hat{\lambda})$ based on the data set. We used cross-validation to tune λ . This procedure was repeated $B = 1000$ times for each C , and based on these B samples, we computed the empirical variance of $\sqrt{n}\hat{\beta}_C(\hat{\lambda}_C)$ for each C . We also estimated the variance using both the classic method and Theorem 5. The results are summarized in Figure 2, which shows the error of the two estimators as functions of C . From Figure 2 it is clear that the classic estimator is insufficient in this case. The variance of Theorem 5 is closer to the observed variability in almost all cases. Furthermore, we see that as $C \rightarrow 0$, the classic variance estimate decreases in effect. This is particularly clear for the estimate of the variance of the last component of $\sqrt{n}\hat{\beta}_C(\lambda_C)$, where the variance estimator provided by the classic method is a lot worse than the one proposed in this article and seems to increase even more as $C \rightarrow 0$. This simulation is of course only a toy example, but Figure 2 indicates that the error introduced by using the classic

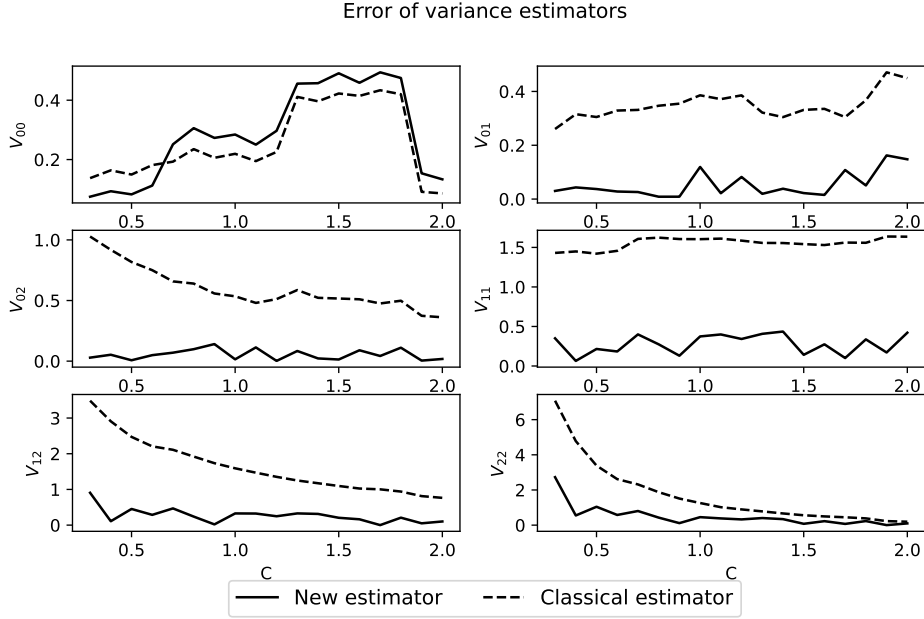


Figure 2: The plot shows the absolute error of the classic and new estimated covariance of $\sqrt{n}\hat{\beta}_C(\hat{\lambda}_C)$ compared to the observed variance based on a thousand simulated data sets with $n = 100$ data points. The variables V_{jk} refer to the covariance between the j -th and k -th component of $\sqrt{n}\hat{\beta}_C(\hat{\lambda}_C)$ for $j = 0, 1, 2$.

estimator of the variance potentially can be very large.

6 Examples

We will now go through a couple of applications illustrating our theoretical framework.

6.1 Diabetes in Pima Indians

The data set `pima.indians.diabetes`, publicly available in the R-package `mlbench` contains information about female Pima Indians and prevalence of diabetes. Denote the data set by $z_1, \dots, z_n$ where $z_i = (y_i, x_i^T)^T$ with x_i being a vector of covariates and y_i a binary variable encoding whether or not the individual has diabetes for $i = 1, \dots, n$. We fit a penalized logistic regression model to these data. For some

pre-specified value of λ , this corresponds to maximising the following expression

$$\ell_n(\beta, \lambda) = \sum_{i=1}^n [y_i \log p(x_i, \beta) + (1 - y_i) \log \{1 - p(x_i, \beta)\}] - n\lambda \|\beta_{1:}\|^2 \quad (8)$$

where $p(x, \beta) = [1 + \exp\{-\beta_0 - \beta_{1:}^T x\}]^{-1}$ when $\beta = (\beta_0, \beta_{1:}^T)^T$. To fit this model, we need to decide upon a value for the tuning parameter λ . One option is to choose λ minimising the Brier score defined as $BS = n^{-1} \sum_{i=1}^n [1 - p\{x_i, \hat{\beta}(\lambda)\}]^2$ where $\hat{\beta}(\lambda)$ is the minimizer of (8). This ensures that the model has good predictive ability. In most cases, however, BS will underestimate the true error. To mitigate this, we will instead choose λ minimising the cross-validated estimate of the Brier score, defined as $n^{-1} \sum_{i=1}^n [1 - p\{x_i, \hat{\beta}(\lambda)_{(-i)}\}]^2$.

With the procedure described above, we find that the minimizer of the cross-validated Brier score is equal to $\hat{\lambda} = 0.0085$. Fixing λ at this value, we can now fit a penalized logistic regression model to the data by minimising (8) evaluated at $\lambda = \hat{\lambda}$. Furthermore, by classic asymptotic likelihood theory, we know that for each fixed λ , $\hat{\beta}(\lambda)$ is consistent for some well-defined quantity $\beta(\lambda)$ and that $\sqrt{n}\{\hat{\beta}(\lambda) - \beta(\lambda)\}$ converges in law to a $N\{0, J(\lambda)^{-1}K(\lambda)J(\lambda)^{-1}\}$ distribution where $J(\lambda)$ and $K(\lambda)$ are well-defined matrices, see e.g. Van der Vaart (2000, Ch. 5). By considering λ as being fixed at $\hat{\lambda}$, we can therefore approximate the limiting distribution of $\hat{\beta}(\hat{\lambda})$ by a $N\{\beta(\hat{\lambda}), J(\hat{\lambda})^{-1}K(\hat{\lambda})J(\hat{\lambda})^{-1}/n\}$ distribution, and use it to make inference about $\beta(\lambda)$. The resulting estimated marginal densities of each component of $\hat{\beta}(\hat{\lambda})$ are shown in Figure 1. We remark that the very small value of $\hat{\lambda}$ is a consequence of the penalization term being scaled by n , leading the “actual” size of the penalization to be $n\hat{\lambda} \approx 3.342$.

The above approximate distribution is based on pointwise results which do not take the randomness introduced by tuning λ into account. This can, however, be done with the results of this paper. Assuming all regularity conditions are in place, we can apply the theory of this article to $\hat{\beta}(\hat{\lambda})$ by setting $\varphi(z, \beta, \lambda) = \nabla_{\beta}[y \log p(x, \beta) - (1 - y) \log \{1 - p(x, \beta)\} - \lambda \|\beta_{1:}\|^2]$ and $\psi(z, \beta) = \{1 - p(x, \beta)\}^2$. By Theorem 5,

$\sqrt{n}\{\hat{\beta}(\hat{\lambda}) - \beta_0\}$ converges in distribution to a central normal distribution with variance as described in Corollary 1 and which can be estimated using the approximations of Theorem 4. For each $j = 0, \dots, 8$, densities of the marginal approximated distributions of $\hat{\beta}(\hat{\lambda})_j$ together with the densities obtained using the classic, pointwise results as well as histograms based on $B = 2000$ bootstrap samples of $\hat{\beta}(\hat{\lambda})_j$ can be found in Figure 1 which was discussed briefly in the introduction.

From Figure 1 we note that when taking the added randomness of the tuning procedure into account, we get a larger estimated variances than when using the classic pointwise results. This is promising, as the estimated variance of $\hat{\beta}(\hat{\lambda})_j$ was closer to the empirical variance in the bootstrap sample for all but one value of j . For $\hat{\beta}(\hat{\lambda})_2$, the classic variance estimator performed better than what we get from Theorem 4. A quick glance at Figure 1, however, reveals that neither of the approximate distributions seem to fit well with the bootstrap distribution in this case. In fact, the histogram of this component is quite skewed, giving reason to believe that convergence has not been achieved for this particular parameter.

Although the results above are slightly in the favour of the theorems derived in the present article, the limiting distributions we get using our and the classic method are not astoundingly different. It is, however, important to remember that the classic result is a pointwise result concerning the limiting behaviour of $\hat{\beta}(\lambda)$ when λ is a fixed number. Because of this, the classic result really does not apply to $\hat{\beta}(\hat{\lambda})$ and cannot be used to say anything about the limiting behaviour when $\hat{\lambda}$ is set by tuning, as neither asymptotic normality nor consistency of $\hat{\beta}(\hat{\lambda})$ follows from standard theory. By using the classic pointwise method regardless of this fact, we are therefore really taking a gamble, hoping that the added randomness of the tuning procedure will be negligible. Theorem 5 and Theorem 4, on the other hand, do take the tuning procedure into account and can safely be applied to make valid asymptotic inference.

6.2 Hybrid likelihood

In the book “The better angels of our nature”, Pinker (2011) argues that the world is changing for the better and that violence is on the decline. This claim quickly stirred up heated debate in the academic community and since its publication a large number of articles have either attempted to support or refute Pinker’s claim, see e.g. Cirillo and Taleb (2016a,b) or Cunen et al. (2020). Recently, Dæhlen and Hjort (2024) investigated whether conflicts have become less violent over time. We will now use the theory derived in the present article to improve the analysis done in that paper.

Dæhlen and Hjort (2024) uses data from the Correlates of War data base (Sarkees and Wayman (2010)) concerning the number of battle deaths in the 95 most recent and concluded inter-state wars. The analysis is done by first splitting the data into “older” and “newer” wars with the Korean war as the cut-off point. To each of these data sets, they fit shifted log-normal models and use the result to make inference about the difference in the median number of battle deaths for older and newer wars. They also repeat the analysis for the difference in upper quartiles.

Further, to fit the parametric models, the framework of hybrid likelihood is used. We will not go into details concerning this method, but roughly speaking, they fit parametric models f_θ by minimising a convex combination of the log-likelihood function and a nonparametric counterpart called the log-empirical likelihood function, see e.g. Owen (2001). How much weight is put on either of these components is decided by a parameter called the balance parameter a taking values in $[0, 1)$.

Moreover, the FIC is used to select values for the balance parameters a_1 and a_2 . When investigating change in the median number of battle deaths, the median is used as focus parameter and the empirical median as the consistent estimator $\hat{\mu}_0$. When studying the difference in upper quartile, the 0.75-quantile and the empirical 0.75-quantile are used in place of the median and empirical median. By standard results, the empirical p -quantile satisfies $o_p(1/\sqrt{n}) = n^{-1} \sum_{i=1}^n \xi(Y_i, \mu)$ where $\xi(y, \mu) = \mu - \mu_p - \{p - I(y \leq \mu_p)\}/f(\mu_p)$ and μ_p is the p -quantile in the underlying distribution

of the data. Furthermore, by the results of Dæhlen and Hjort (2024), the estimators $\hat{\theta}_1(a_1)$ and $\hat{\theta}_2(a_2)$ can both be expressed as components of Z-estimators for each fixed value of a_1 and a_2 . Hence, by making modifications as explained in Section 3, we can use Corollary 1 to derive the limiting distribution of $\hat{\theta}_1(\hat{a}_1)$ and $\hat{\theta}_2(\hat{a}_2)$.

By combining the above with the estimators given in Theorem 4, we can approximate the variance of $g\{\hat{\theta}(\hat{a}_1)\} - g\{\hat{\theta}(\hat{a}_2)\}$, where $g(\theta)$ is the median in a log-normal distribution parametrized by θ , in a way taking the added randomness introduced by tuning a_1 and a_2 into account. Using this method, we get an estimated standard deviation of 18431. That is a lot higher than 3815, which is obtained using the pointwise results in Dæhlen and Hjort (2024). This is a consequence of Dæhlen and Hjort (2024) not incorporating the added randomness introduced by tuning a_1 and a_2 , leading to a much too small estimated standard deviation.

For the difference in upper quartiles, $\hat{a}_1 = 0$ and $\hat{a}_2$ is chosen as the largest accepted value of a_2 (in this case 0.99). Because of this, $\text{FIC}'_j(\hat{a}_j) \neq 0$ for $j = 0, 1$ and Theorem 2 leads us to approximate the limiting distribution of $\hat{\theta}_j(\hat{a}_j)$ by the pointwise limit of $\hat{\theta}_j(\hat{a}_j)$ for $j = 0, 1$. This leads to the same approximate distribution as what is used in Dæhlen and Hjort (2024), and hence, all estimated standard deviations agree. As a result, ignoring the effect of tuning of a_1 and a_2 becomes unproblematic in this case.

7 Concluding remarks

We have studied the effect tuning procedures has on the limiting distribution of $\hat{\theta}(\hat{\lambda})$. We covered multiple ways of tuning, including minimization of information criteria, the training error and the cross-validated estimator of the risk. In addition, we have defined consistent estimators for the limiting variance of $\sqrt{n}\hat{\theta}(\hat{\lambda})$ and proved a result sharply characterizing the distance from CV to TE. Lastly, we went through a simulation setting and applied the theory on two real data sets.

Our work does have some limitations. The perhaps most pressing issue with our

current theory is the inability to handle non-smooth functions φ and ψ . This excludes multiple interesting settings. Since, ψ must be smoothly differentiable, tuning with respect to the absolute error or error rate is not covered by our theory. Furthermore, smoothness of φ excludes many interesting models, like the Lasso or other models penalized by an L_1 -norm. This does, of course, make our theory less applicable, but the smoothness assumptions are crucial for our proofs, and we believe that extending our theory to cover non-smooth settings is far from trivial.

Further, we have assumed that for each λ , $\hat{\theta}(\lambda)$ is a Z-estimator. This assumption is necessary for the proof of Theorem 5 and Theorem 6, but might be more strict than what is really needed at least for the latter theorem. By the informal arguments preceding Theorem 6, one would expect that Theorem 5 and Theorem 6 hold true as long as $\hat{\theta}(\lambda)$ has an influence function IF, such that $\hat{\theta}(\lambda) = \theta_0(\lambda) + \text{IF}(Y_i, \theta, \lambda) + \epsilon_n(\theta, \lambda)$ with “regular enough” remainder term. Formalizing this intuition in a proper way would certainly extend our framework, but will require proving a more general version of Lemma S1.3 in the supplementary material.

Finally, we have focused on Theorem 2 and Theorem 5 as our goal was to study the limiting behaviour of $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set by some tuning procedure. These results are, however, corollaries of Theorem 1, which gives the limiting distribution of the full vector $(\hat{\theta}(\hat{\lambda})^T, \hat{\lambda}^T, \hat{\theta}'(\hat{\lambda})^T)^T$. In particular, Theorem 1 can be used to study the limiting behaviour of $\hat{\lambda}$. This can, in turn, be applied to make inference about λ and create approximate confidence intervals and hypothesis tests for the tuning parameter. For instance, it might be used to formally test if a regularization is beneficial or if either of the two components in a hybrid model is significantly preferable.

Supplementary Materials

Proofs, lemmas and further discussions can be found in the article’s supplementary material.

Acknowledgments

This work is partially funded by the Norwegian Research council through the BigInsight center for innovation driven research (237718). Partial support was also given from Centre for Advanced Study, at the Academy of Science and Letters, Oslo, in connection with the 2022-2023 project Stability and Change, led by H. Hegre and N.L.

References

- Arcones, M. A. (2005). Convergence of the optimal m-estimator over a parametric family of m-estimators. *Test* 14, 281–315.
- Bachoc, F., D. Preinerstorfer, and L. Steinberger (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics* 48(1), 440–463.
- Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics* 41, 802–837.
- Billingsley, P. (1999). *Convergence of probability measures*. New York: John Wiley & Sons.
- Bouchard, G. and B. Triggs (2004). The tradeoff between generative and discriminative classifiers. In *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, Prague, Czech Republic, pp. 721–728.
- Cirillo, P. and N. N. Taleb (2016a). The decline of violent conflicts: What do the data really say? *Nobel Foundation Symposium 161: The Causes of Peace*, 1–26.
- Cirillo, P. and N. N. Taleb (2016b). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications* 452, 29–45.
- Claeskens, G. and N. L. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association* 98(464), 900–916.

- Claeskens, G., N. L. Hjort, et al. (2008). *Model selection and model averaging*. UK: Cambridge University Press.
- Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische mathematik* 31(4), 377–403.
- Cunen, C., N. L. Hjort, and H. M. Nygård (2020). Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research* 57(2), 221–234.
- Dæhlen, I. and N. L. Hjort (2024). Model robust hybrid likelihood. *arXiv preprint arXiv:2409.15975*.
- Dæhlen, I., N. L. Hjort, and I. Hobæk Haff (2024). Accurate bias estimation with applications to focused model selection. *Scandinavian Journal of Statistics* 5.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association* 78(382), 316–331.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association* 81(394), 461–470.
- Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 99(467), 619–632.
- Hall, A. R. and A. Inoue (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics* 114(2), 361–394.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society* 50, 1029–1054.
- Huber, P. J. (2009). *Robust statistics* (2nd ed.). Hoboken: Wiley.

- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of multivariate Analysis* 94, 401–419.
- Jullum, M. and N. L. Hjort (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica* 27(3), 951–981.
- Ko, V. and N. L. Hjort (2019). Model robust inference with two-stage maximum likelihood estimation for copulas. *Journal of Multivariate Analysis* 171, 362–381.
- Konishi, S. and G. Kitagawa (2008). *Information criteria and statistical modeling*. New York: Springer Science & Business Media.
- Kuchibhotla, A. K., J. E. Kolassa, and T. A. Kuffner (2022). Post-selection inference. *Annual Review of Statistics and Its Application* 9, 505–527.
- Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44, 907–927.
- Li, K.-C. (1985). From stein’s unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics* 13, 1352–1377.
- Li, K.-C. (1986). Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics* 14, 1101–1112.
- Li, K.-C. (1987). Asymptotic optimality for C_p , CL, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics* 15, 958–975.
- Mu, B., T. Chen, and L. Ljung (2018). On asymptotic properties of hyperparameter estimators for kernel-based regularization methods. *Automatica* 94, 381–395.
- Mu, B., T. Chen, and L. Ljung (2021). On the asymptotic optimality of cross-validation based hyper-parameter estimators for regularized least squares regression problems. *arXiv preprint*.

- O'Neill, T. J. (1980). The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. *Journal of the American Statistical Association* 75(369), 154–160.
- Owen, A. B. (2001). *Empirical likelihood*. Boca Raton, FL: CRC press.
- Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Toronto: Viking.
- Sarkees, M. R. and F. Wayman (2010). *Resort to war: 1816 - 2007*. Washington DC: CQ Press.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association* 88(422), 486–494.
- Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica sinica* 7, 221–242.
- Shao, J. (2003). *Mathematical statistics*. New York: Springer Science & Business Media.
- Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics* 13, 970–983.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1), 44–47.
- Tibshirani, R. and K. Knight (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61(3), 529–546.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.

Appendices

Appendix A

Supplementary material for Paper II

Model robust hybrid likelihood

Ingrid Dæhlen and Nils Lid Hjort

Department of Mathematics, University of Oslo

Supplementary Material

The supplementary material contains proofs of Theorem 1 and 2 in the paper.

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

Lemma 1. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and let $\psi: \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^q$ be a function with $\Theta \subseteq \mathbb{R}^p$. Assume there is a unique solution θ_n to the equation $0 = \Psi_n(\theta) = n^{-1} \sum_{i=1}^n \psi(Y_i, \theta)$ for all n and a unique solution θ_0 to the limit equation $0 = \Psi(\theta) = E[\psi(Y, \theta)]$ when $Y \sim F$. Then θ_n converges in probability to θ_0 , provided*

- (C1) *the function Ψ is continuously differentiable in a neighborhood of θ_0 and $\Psi'(\theta_0)$ is invertible;*
- (C2) *there exists a neighborhood $\mathcal{N}$ of θ_0 and an F -integrable function p_0 such that $\|\psi(y, \theta)\| \leq p_0(y)$ for all $\theta \in \mathcal{N}$ and F -almost all y .*

Proof. Let Ψ and Ψ_n be defined as in the lemma. Inspection of the proof of Theorem 5.42 in Van der Vaart (2000) shows that the conditions of the theorem are slightly too strict and that the only thing needed for the arguments to go through is that Ψ is continuously differentiable with non-singular derivative in a neighborhood of θ_0 and that there exists a closed ball around θ_0 inside which Ψ_n converges uniformly to Ψ . The first condition is ensured by (C1) and continuity of the determinant function. The second condition follows from (C2) and the uniform law of large numbers. Hence, there exists a sequence of estimators converging in probability to θ_0 . By assumption, $\Psi_n(\theta) = 0$ has a unique solution θ_n . Because of this, θ_n must be this consistent sequence, proving that θ_n converges in probability to θ_0 . $\square$

S1 The maximum empirical likelihood estimator

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

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

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

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

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

with ∂m as short hand for $\partial m / \partial \mu$. Lastly assume $\Gamma^{\text{EL}}(\alpha) = \mathbb{E} \psi(Y, \alpha)$ has a unique root $\alpha_0 = (\lambda_0, \mu_0)$. Then $\hat{\alpha}$ is consistent for α_0 and

$$\hat{\alpha} = \alpha_0 - n^{-1} J \Gamma^{\text{EL}}(\alpha_0)^{-1} \sum_{i=1}^n \psi_{\text{EL}}(Y_i, \alpha_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (\text{S1.1})$$

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

- (1) Γ^{EL} is continuously differentiable with non-singular Jacobian matrix at α_0 ;
- (2) $\Pr(1 + \lambda^T m(Y, \mu) > L) = 1$ for some $L > 0$ and all $(\lambda, \mu) \in \mathcal{N}$;
- (3) the norm of m and all first and second order partial derivatives of m are bounded by some functions p_j for $j = 0, 1, 2$ respectively, with $\mathbb{E}[p_0(Y)^4]$, $\mathbb{E}[p_1(Y)^4] < \infty$ and $\mathbb{E}[p_2(Y)^2] < \infty$.

In particular, $\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N[0, J \Gamma^{\text{EL}}(\alpha_0)^{-1} K_{\text{EL}} J \Gamma^{\text{EL}}(\alpha_0)^{-1}]$, with $K_{\text{EL}} = \text{Var}[\psi_{\text{EL}}(Y, \alpha_0)]$, $Y \sim F$.

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

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

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

We start by showing consistency towards α_0 . This is achieved by proving condition (C2) of Lemma 1. (C1) holds by condition (1) of this lemma. By assumption (2) there exists $L > 0$ such that $1 + \lambda^T M(\mu) > L$ with probability 1 for all $\alpha \in \mathcal{N}$. Combining this with condition (3) ensures $\|\psi_{\text{EL}}(y, \alpha)\| \leq C_1 p_0(y) + C_2 p_1(y)$ for some constants $C_1, C_2 \in \mathbb{R}$. The right hand side of this inequality is a sum of F -integrable functions. This shows condition (C2) of Lemma 1, and guarantees that $\hat{\alpha}$ converges in probability to α_0 .

We will use Theorem 5.21 in Van der Vaart (2000) to show (S1.1). We start with the Lipschitz condition. Direct computations show $\|\frac{\partial \psi}{\partial \alpha}(y, \alpha)\| \leq C_3 p_0(y)^2 + C_4 p_0(y) p_1(y) + C_5 p_1(y) + C_6 p_1(y)^2 + C_7 p_2(y)$, for some constants C_j for $j = 3, \dots, 7$. This is a consequence of condition (3) of the lemma. Since all functions on the right hand side of this equation are square integrable with respect to the measure F , the sum is square integrable as well. This in conjunction with the mean value theorem for multivariate functions, ensures the Lipschitz property of Theorem 5.21 in Van der Vaart (2000). Furthermore, $\mathbb{E}[\|\psi(Y, \alpha_0)\|^2] < \infty$ by a similar argument. The rest of the conditions of Theorem 5.21 in Van der Vaart (2000) hold by assumption. (S1.1) follows.

Strictly speaking, we have only showed that the sequence of solutions to $n^{-1} \sum_{i=1}^n \psi(Y_i, \alpha)$ satisfies the properties of the lemma. Since, $\log \text{EL}_n(\mu)$ is only equal to $-\sum_{i=1}^n \log[1 +$

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

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

Theorem 1. *Let all assumptions of Theorem 2 hold true and all quantities be as defined there. Furthermore, assume that $\mu \rightarrow \text{E} \log[1 + \lambda(\mu)^T m(Y, \mu)]$ where $\lambda(\mu)$ denotes the solution to $0 = \text{E}\{m(Y, \mu) / [1 + \lambda^T m(Y, \mu)]\}$ has a maximizer μ_0 . Then the MEL estimator $\hat{\mu}$ is consistent for μ_0 , and*

$$\hat{\mu} = \mu_0 + n^{-1} J_{\text{EL}}^{-1} \sum_{i=1}^n \psi_\mu(Y_i, \mu_0) + o_{\text{pr}}(1/\sqrt{n}), \quad (\text{S1.2})$$

with $J_{\text{EL}} = H_{\mu_0} \text{E}\{\log[1 + \lambda^T M(\mu)]\}$ and $\psi_\mu(y, \mu) = -\nabla_\mu \log[1 + \lambda(\mu)^T m(y, \mu)]$, ∇_μ being the gradient with respect to μ .

Proof. By Lemma 2, $\hat{\mu}$ converges in probability to the last s components of α_0 , the parameter solving $0 = \text{E}[\psi_{\text{EL}}(Y, \alpha)]$. By definition $\lambda(\mu)$ is the root of $\text{E}\{m(Y, \mu) / [1 + \lambda^T m(Y, \mu)]\}$, the first q component functions in $\text{E}[\psi_{\text{EL}}(Y, \alpha)]$. Hence, $\alpha_0 = (\lambda(\mu_0), \mu_0)$ where μ_0 is the root of $\text{E}\{\partial m(Y, \mu)^T \lambda(\mu) / [1 + \lambda(\mu)^T m(Y, \mu)]\}$. Furthermore, the maximizer of $-\text{E}\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$ satisfies $0 = \text{E}\{[\lambda'(\mu)^T m(Y, \mu) + \partial m(Y, \mu)^T \lambda(\mu)] / [1 + \lambda(\mu)^T m(Y, \mu)]\}$, since $\mathcal{M}$ is open and m is differentiable. As $\lambda(\mu)$ is the root of $\text{E}\{m(Y, \mu) / [1 + \lambda^T m(Y, \mu)]\}$, the right hand side of this equation is equal to $\text{E}\{\partial m(Y, \mu)^T \lambda(\mu) / [1 + \lambda(\mu)^T m(Y, \mu)]\}$. Since the root of $\Gamma^{\text{EL}}(\alpha)$ is unique, μ_0 must therefore be the the maximizer of $-\text{E}\{\log[1 + \lambda(\mu)^T m(Y, \mu)]\}$.

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

$$\hat{\mu} = (0, I) \hat{\alpha} = \mu_0 - (0, I) \Gamma^{\text{EL}}(\alpha_0)^{-1} n^{-1} \sum_{i=1}^n \psi(Y_i, \alpha_0) + o_{\text{pr}}(1/\sqrt{n}).$$

Let H_{xy} be the matrix obtained by first differentiating $\text{E}\{\log[1 + \lambda^T M(\mu)]\}$ with respect to y and afterwards with respect to x evaluated at (λ_0, μ_0) . Then $J \Gamma^{\text{EL}}(\alpha_0)$ can be written as the following block matrix

$$J \Gamma^{\text{EL}}(\alpha_0) = \begin{pmatrix} H_{\lambda\lambda} & H_{\mu\lambda} \\ H_{\lambda\mu} & H_{\mu\mu} \end{pmatrix}.$$

Furthermore, $\lambda'(\mu_0) = -H_{\lambda\lambda}^{-1} H_{\mu\lambda}$ by the implicit function theorem. Utilizing this and the formula for inverses of block matrices, we find $(0, I) \Gamma^{\text{EL}}(\alpha_0) = (A^{-1} \lambda'(\mu_0)^T, A^{-1}) = A^{-1} (\lambda'(\mu_0)^T, I)$, where $A = H_{\mu\mu} - H_{\lambda\mu} H_{\lambda\lambda}^{-1} H_{\mu\lambda}$. Direct computation shows that $A = J_{\text{EL}}$. Putting all of this together shows $-(0, I) \Gamma^{\text{EL}}(\alpha_0)^{-1} \psi(y, \alpha_0) = J_{\text{EL}}^{-1} (\lambda'(\mu_0)^T, I) \psi(y, \lambda_0, \mu_0) = J_{\text{EL}}^{-1} \psi_\mu(y, \mu_0)$, proving (S1.2). $\square$

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

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

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

and set

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

with $\partial_{\mu} m$ being shorthand for $\partial m / \partial \mu$. Then, if $\alpha_{P,0} = (\mu_0, \lambda_0, P_0)$ is the unique solution to $0 = \Gamma_P^{\text{EL}}(\mu, \lambda, P) = (\mathbb{E}[\psi_{\text{EL}}^P(Y, \alpha_P)]^T, \mathbb{E}[h(Y, P)]^T)^T$, the vector $\hat{\alpha}_P = (\hat{\mu}^T, \hat{\lambda}^T, \hat{P}^T)^T$ converges in probability to $\alpha_{P,0}$ and

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

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

- (1) the function Γ_P^{EL} is continuously differentiable with non-singular Jacobian matrix at $\alpha_{P,0}$;
- (2) $\Pr(1 + \lambda^T m(Y, \mu, P) > L) = 1$ for all $(\mu, \lambda, P) \in \mathcal{N}$ and some $L > 0$;
- (3) the norm of m and all first and second order partial derivatives of m with respect to μ and P are bounded by functions p_j for $j = 0, 1, 2$ respectively, with $\mathbb{E}[p_0(Y)^4]$, $\mathbb{E}[p_1(Y)^4] < \infty$ and $\mathbb{E}[p_2(y)^2] < \infty$.
- (4) the norm of $h(y, P)$ and all first order partial derivatives of h with respect to P is bounded by functions q_1 and q_2 not depending on P and with $\mathbb{E}[q_1(Y)]$, $\mathbb{E}[q_2(Y)^2] < \infty$.

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

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

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

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

Theorem 2. Assume the conditions of Theorem 2 hold true and let all quantities be as defined here. Furthermore, let $\lambda(\mu, P)$ and P_0 denote the solutions to $0 = \mathbb{E}\{m(Y, \mu, P)/[1 + \lambda^T m(Y, \mu, P)]\}$ and $\mathbb{E}h(Y, P) = 0$ respectively, and assume that the function $\mu \mapsto \mathbb{E} \log[1 +$

$\lambda(\mu, P)^T m(Y, \mu, P)]$ has a maximizer $\mu(P)$ for each P . Then the MEL estimator $\hat{\mu}(\hat{P})$ is consistent for $\mu_0 = \mu(P_0)$ and

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

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

Proof. Let $\alpha_{P,0} = (\lambda_0^T, \mu_0^T, P_0^T)^T$ be the root of $E[\Gamma_P^{\text{EL}}(\alpha_P)] = (E[\psi_{EL}^P(Y, \alpha_P)]^T, E[h(Y, P)]^T)^T$. Then, by definition, P_0 is the solution to $E[h(Y, P)] = 0$ and λ_0 the solution to

$$E\left(\frac{m(Y, P_0, \mu_0)}{1 + \lambda^T m(Y, P_0, \mu_0)}\right) = 0.$$

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

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

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

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

and ensuring that $\Gamma^{\text{EL}}(\mu_1, \lambda(\mu_1, P_0), P_0) = 0$, but by assumption this equation has a unique root at $\alpha_{P,0}$. Hence, $\mu_1 = \mu_0$ and μ_0 is the maximizer of $-E\{\log[1 + \lambda(\mu, P_0)^T m(Y, P_0, \mu)]\}$.

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

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

where $J \Gamma_P^{\text{EL}}(\lambda, \mu)$ is equal to $J \Gamma^{\text{EL}}(\lambda, \mu)$ defined in Lemma 3 where the estimating function used is $(y, \mu) \mapsto m(y, P, \mu)$ for each fixed P and $\partial_P f$ is short-hand for $\partial f / \partial P$ for a general function f . By inversion of block matrices and condition (1) in Lemma 2, this matrix is continuously differentiable in a neighborhood of $(\lambda_0^T, \mu_0^T, P_0^T)^T$ with

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

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

$$\begin{aligned} & - (0 \ I_s \ 0) (J \Gamma_P^{\text{EL}}(\alpha_{P,0}))^{-1} \psi_P(y, \alpha_{P,0}) = \\ & J_{\text{EL}}^{-1} \psi_{\mu}^{P_0}(y, \mu_0) - \mu'(P_0)[\partial_P E h(Y, P)]^{-1} h(y, P_0), \end{aligned}$$

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

S2 Proving Theorem 1 in the article

We first show a lemma.

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

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

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

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

Assume that $\beta_0 = (\lambda_0, \mu_0)$ is the unique root of $\Gamma^{\text{HL}}(\lambda, \mu) = E[\psi(Y, \beta)]$, where $Y \sim F$. Then $\hat{\beta}$ is consistent for β_0 and

$$\hat{\beta} = \beta_0 - n^{-1} J \Gamma^{\text{HL}}(\beta_0)^{-1} \sum_{i=1}^n \psi_{\text{HL}}(Y_i, \beta_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (\text{S2.7})$$

provided condition (2) and (3) of Lemma 2 hold true with $\alpha_0 = (\lambda_0^T, \mu(\theta_0)^T)^T$ and the following additional assumptions

- (4) the function $(\lambda, \theta) \mapsto E\{(1 - a)u(Y, \theta)^T - aM'(\theta)/[1 + \lambda^T M(\theta)], M^T/[1 + \lambda^T M(\theta)]\}$ is continuously differentiable in a neighbourhood of β_0 with non-singular derivative at β_0 .
- (5) the function $\mu(\theta)$ is twice continuously differentiable in a neighbourhood of θ_0 .
- (6) in a neighbourhood of θ_0 , all components of u and $\partial u/\partial \theta$ are bounded by some F -square integrable functions q_1 and q_2 respectively not depending on θ .

In particular, $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, J \Gamma^{\text{HL}}(\beta_0)^{-1} K_{\text{HL}} J \Gamma^{\text{HL}}(\beta_0)^{-1})$, with $K_{\text{HL}} = \text{Var}[\psi_{\text{HL}}(Y, \beta_0)]$, $Y \sim F$.

Proof. Notice that

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

where the last equality follows from the definition of $\lambda_n(\theta)$ and ∂m is short hand for $\partial m/\partial \mu$. Hence, since Θ is open and the maximizer of h_n in Θ is the unique critical point of the function, $\hat{\beta}$ is the unique solution to $0 = n^{-1} \sum_{i=1}^n \psi_{\text{HL}}(Y_i, \beta)$ where ψ_{HL} is as defined in the lemma. Because of this $\hat{\beta}$ is a Z-estimator and, just as for the MEL estimator, standard results can be used to study its limiting behavior.

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

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

is bounded some square F -integrable function depending on neither λ nor μ . Combined with assumption (5), this proves that the norm of $\psi(y, \beta)$ is bounded by some square integrable function not depending on β . In particular, (C2) of Lemma 1 holds true. As (C1) holds by assumption (4), consistency of $\hat{\beta}$ towards β_0 follows.

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

S2.1 Proof of Theorem 1 in the article

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

S3 Proving Theorem 2 in the article

We start with a lemma, yet again.

Lemma 5. *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. from some distribution F and $m: \mathbb{R}^{d+s+r} \rightarrow \mathbb{R}^q$ a function continuously differentiable in the last $s+r$ arguments. Let $\hat{P}$ be some estimator satisfying $n^{-1} \sum_{i=1}^n h(Y_i, \hat{P}) = o_p(1/\sqrt{n})$ and assume that $(y, \mu) \mapsto m(y, \mu, \hat{P})$ is used in construction of the empirical likelihood function EL_n . Let f_θ for $\theta \in \Theta \subseteq \mathbb{R}^p$, where Θ is an open set, be some parametric family with score function u and $\mu: \mathbb{R}^{p+r} \rightarrow \mathbb{R}^s$ such that $E\{m[Y, \mu(\theta, P), P]\} = 0$ when $Y \sim f_\theta$. Fix $a \in [0, 1)$ and assume that $\hat{\theta}(\hat{P})$ is the maximizer of h_n and unique critical value. Furthermore, let $\hat{\lambda}$ be the unique solution to*

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

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

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

Then if $\beta_{P,0} = (\theta_0, \lambda_0, P_0)$ is the unique root of $\Gamma_P^{\text{HL}}(\lambda, \mu) = E[\psi_{\text{HL}}^P(Y, \beta)]$ where $Y \sim F$, $\hat{\beta}_P$ is consistent for $\beta_{P,0}$ and

$$\hat{\beta} = \beta_0 - n^{-1} J \Gamma_P^{\text{HL}}(\beta_0)^{-1} \sum_{i=1}^n \psi_{\text{HL}}^P(Y_i, \beta_0) + o_{\text{pr}}(1/\sqrt{n}) \quad (\text{S3.9})$$

provided condition (2-4) of Lemma 3 hold true with $\alpha_{P,0} = (\mu(\theta_0, P_0)^T, \lambda_0^T, P_0^T)^T$ and the following additional assumptions

- (5) The function Γ_P^{HL} is continuously differentiable in a neighbourhood of $\beta_{P,0}$ with non-singular derivative at $\beta_{P,0}$;
- (6) For all P in an neighborhood of P_0 , the function $\mu(\theta, P)$ is twice continuously differentiable with respect to θ and P in a neighbourhood of $(\theta(P_0)^T, P_0^T)^T$.
- (7) There exists a neighborhood of θ_0 on which all components of u and $\partial u / \partial \theta$ are bounded by some F -square integrable functions q_1 and q_2 respectively not depending on θ .

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

S3.1 Proof of Theorem 2 in the article

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

Bibliography

- Imbens, G. W., R. H. Spady, and P. Johnson (1998). Information theoretic approaches to inference in moment condition models.
- Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics* 35, 634–672.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.

Appendix B

Supplementary material for Paper III

The asymptotic effect of tuning parameters - Supplementary material

Ingrid Dæhlen, Nils Lid Hjort and Ingrid Hobæk Haff

Department of Mathematics, University of Oslo

Address for correspondence: Ingrid Dæhlen, Department of Mathematics, University of Oslo, Moltke Moes vei 35, 0851 Oslo, Norway, email: ingrdae@math.uio.no

S1 Proof of Lemma 1 and Theorem 6

We start by showing a lemma regarding uniform consistency.

Lemma S1.1. *Let $Y \in \mathbb{R}^d$, $\theta \in \mathbb{R}^p$ and $\lambda \in \Lambda \subseteq \mathbb{R}^q$ where Λ is a compact set. With $\Psi_{\lambda,n}: \mathbb{R}^{d+p} \rightarrow \mathbb{R}^p$ and $\Psi_\lambda: \mathbb{R}^p \rightarrow \mathbb{R}^p$ for each $\lambda \in \Lambda$, let $\theta_0(\lambda)$ be the unique zero of $\Psi_\lambda(\theta)$ for each fixed λ and assume the following.*

(C1) *There exists $\epsilon > 0$ such that $\Psi_{\lambda,n} \xrightarrow{\text{Pr}} \Psi_\lambda$ uniformly on $\mathcal{S}_\epsilon = \{(\theta, \lambda) \mid \lambda \in \Lambda, \|\theta_0(\lambda) - \theta\| \leq \epsilon\}$*

(C2) *The function $(\theta, \lambda) \mapsto \Psi_\lambda(\theta)$ is continuously differentiable on $\mathcal{S}_\epsilon$ with $\Psi'_\lambda\{\theta(\lambda)\}$ non-singular for each $\lambda \in \Lambda$.*

Then, if $\Psi_{\lambda,n}(\theta)$ has at most a single zero $\hat{\theta}(\lambda)$ for each n and each λ ,

$$\sup_{\lambda} \|\hat{\theta}(\lambda) - \theta_0(\lambda)\| \xrightarrow{\text{Pr}} 0. \quad (\text{S1.1})$$

Proof. The following is a modified version of the proof of theorem 5.42 in Van der Vaart (2000).

By (C2) and the implicit function theorem, $\theta(\lambda)$ is continuous as a function of λ in Λ . Hence, $\{\theta(\lambda) \mid \lambda \in \Lambda\}$ is a compact set. This implies that $\mathcal{S}_\epsilon$ is compact. Because of this, the extreme value theorem and the fact that $\Psi'_\lambda\{\theta(\lambda)\}$ is non-singular for all $\lambda \in \Lambda$, we can choose ϵ to be small enough that $\Psi'_\lambda(\theta)$ is non-singular for all $(\theta, \lambda) \in \mathcal{S}_\epsilon$,

ensuring that $(\theta, \lambda) \mapsto \Psi'_\lambda(\theta)^{-1}$ is continuous on $\mathcal{S}_\epsilon$. By the extreme value theorem, we then have that the norm of each component of $\Psi'_\lambda(\theta)^{-1}$ is bounded by some $C < \infty$ in $\mathcal{S}_\epsilon$.

Now fix some $\delta > 0$ and choose δ_1 smaller than both ϵ and $\delta/2C$. Arguing as in Van der Vaart (2000), we see that there exists closed neighbourhoods $G_{\delta_1}^\lambda$ of $\theta_0(\lambda)$ such that Ψ^λ is a homeomorphism from $G_{\delta_1}^\lambda$ into the closed ball around zero with radius δ_1 : $\overline{B_{\delta_1}(0)}$ for each λ . By construction of δ_1 and the inverse function theorem, the partial derivatives of the inverse of Ψ_λ is bounded by C for each λ . Hence, the diameter of $G_{\delta_1}^\lambda$ is bounded by $2C\delta_1 < \delta$ for each λ implying $G_{\delta_1}^\lambda \subseteq \overline{B_\delta\{\theta_0(\lambda)\}}$ for each λ . We will show that $\Psi_{\lambda,n}(\theta)$ has a root in $G_{\delta_1}^\lambda$ for every λ with probability tending to one. Since $G_{\delta_1}^\lambda \subseteq \overline{B_\delta\{\theta_0(\lambda)\}}$ for each λ , this would guarantee (S1.1).

Since $\delta_1 < \epsilon$, we know

$$\sup_{\lambda \in \Lambda} \sup_{\theta \in G_{\delta_1}^\lambda} \|\Psi_{\lambda,n}(\theta) - \Psi_\lambda(\theta)\| \xrightarrow{\text{Pr}} 0.$$

Let $K_{\delta_1,n}$ be the event that $\|\Psi_{\lambda,n}(\theta) - \Psi_\lambda(\theta)\| < \delta_1$ for all $\theta \in G_{\delta_1}^\lambda$ and $\lambda \in \Lambda$. Then the above implies that $\Pr(K_{\delta_1,n}) \rightarrow 1$.

Now let $y \in \overline{B_{\delta_1}(0)}$. Since Ψ_λ is a homeomorphism from $G_{\delta_1}^\lambda$ into $\overline{B_{\delta_1}(0)}$ for each λ , $\Psi_\lambda^{-1}(y) \in G_{\delta_1}^\lambda$ for every $\lambda \in \Lambda$. Hence, on the event $K_{\delta_1,n}$, we have $\|\Psi_{\lambda,n}\{\Psi_\lambda^{-1}(y)\} - y\| \leq \delta_1$ for all $\lambda \in \Lambda$, which implies that $y \mapsto y - \Psi_{\lambda,n} \circ \Psi_\lambda^{-1}(y)$ maps $\overline{B_{\delta_1}(0)}$ into itself for every $\lambda \in \Lambda$. Since the maps are continuous, Brouwer's fixed point theorem ensures that $y \mapsto y - \Psi_{\lambda,n} \circ \Psi_\lambda^{-1}(y)$ has a fixed point in $\overline{B_{\delta_1}(0)}$ for every $\lambda \in \Lambda$. This is equivalent to $\Psi_{\lambda,n}$ having a zero in $G_{\delta_1}^\lambda \subseteq \overline{B_\delta\{\theta_0(\lambda)\}}$ for each λ . Because of this, the probability of all $\Psi_{n,\lambda}$ having a zero within δ of $\theta_0(\lambda)$ is bounded by $\Pr(K_{\delta_1,n})$. As this latter probability tends to one by the previous arguments, (S1.1) follows. This concludes the proof. ■

When $\Psi_\lambda(\theta) = E \phi(Z, \theta, \lambda)$ and $\Psi_{\lambda,n}(\theta) = n^{-1} \sum_{i=1}^n \phi(Z_i, \theta, \lambda)$ for some function $\phi: \mathbb{R}^{d+p+q} \rightarrow \mathbb{R}^p$, condition (C1) can be easily verified using the uniform law of large

numbers. We state this in the following corollary.

Corollary S1.2. *Let $\Psi_\lambda(\theta) = \mathbb{E} \phi(Z, \theta, \lambda)$ and $\Psi_{\lambda,n}(\theta, \lambda) = n^{-1} \sum_{i=1}^n \phi(Z_i, \theta, \lambda)$ for some function $\phi: \mathbb{R}^{d+p+q} \rightarrow \mathbb{R}^p$. Then (C1) follows from*

(C3) *There exists an F -integrable function $m_0: \mathbb{R}^d \rightarrow \mathbb{R}$ and an $\epsilon > 0$ such that $\|\phi(z, \theta, \lambda)\| \leq m_0(z)$ for all $(\theta, \lambda) \in S_\epsilon$ and F -almost all Z .*

We will prove a theorem characterizing the distance from $\hat{\theta}(\lambda)$ to $\hat{\theta}_{(-i)}(\lambda)$.

Lemma S1.3. *For each λ in Λ , a compact subset of $\mathbb{R}^q$ and $1 \leq i \leq n$ let $\hat{\theta}(\lambda)$ and $\hat{\theta}_{(-i)}(\lambda)$ be the solutions to $\sum_{j=1}^n \phi(Z_j, \theta, \lambda)$ and $\sum_{j \neq i} \phi(Z_j, \theta, \lambda)$ in Θ , a compact subset of $\mathbb{R}^p$. Assume there exists functions m_0 and m_1 such that $\|\phi(z, \theta, \lambda)\| \leq m_0(z)$ and $\|\partial_\theta \phi\{z, \theta, \lambda\}\| \leq m_1(z)$ for all θ, λ and almost all z . Then, if $\theta \mapsto \mathbb{E} \phi(Z, \theta, \lambda)$ satisfies (C2) of Lemma S1.1 and the eigenvalues of $J(\theta, \lambda)^{-1} = -\{\mathbb{E} \partial_\theta \phi(Z, \theta, \lambda)\}^{-1}$ are bounded in $\Theta \times \Lambda$, the following holds for all λ :*

$$\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\| \leq a_{0,n} m_0(Z_i) \quad (\text{S1.2})$$

where $a_{0,n} = O_p(1/n)$ and does not depend on i or λ . If, furthermore, the eigenvalues of $\partial_\theta J(\theta, \lambda)$ are bounded in $\Theta \times \Lambda$, we have

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = n^{-1} J\{\theta_0(\lambda), \lambda\}^{-1} \phi\{Z_i, \theta_0(\lambda), \lambda\} + E_n^i(\lambda) \quad (\text{S1.3})$$

where $E_n^i(\lambda)$ satisfies

$$\|E_n^i(\lambda)\| \leq n^{-1} \{a_{1,n} + a_{2,n} m_0(Z_i) + a_{3,n} m_0(Z_i)^2 + a_{4,n} m_0(Z_i)^2 m_1(Z_i)\} \quad (\text{S1.4})$$

with $a_{j,n} = o_p(1)$ uniformly in λ , not depending on i for $j = 1, 2, 3, 4$.

Proof. By definition we have

$$0 = \sum_{j=1}^n \phi\{Z_j, \hat{\theta}(\lambda), \lambda\} \quad \text{and} \quad 0 = \sum_{j \neq i}^n \phi\{Z_j, \hat{\theta}_{(-i)}(\lambda), \lambda\}. \quad (\text{S1.5})$$

Hence, a Taylor expansion of $\sum_{j=1}^n \phi\{Z_j, \hat{\theta}_{(-i)}(\lambda), \lambda\}$ around $\hat{\theta}(\lambda)$ reveals

$$\begin{aligned}\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\} &= \sum_{j=1}^n \phi\{Z_j, \hat{\theta}_{(-i)}(\lambda), \lambda\} \\ &= \sum_{j=1}^n \phi\{Z_j, \hat{\theta}(\lambda), \lambda\} + \left(\sum_{j=1}^n \partial_{\theta_i^*} \phi\{Z_j, \theta, \lambda\} \right) \{\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\} \\ &= -n J_n\{\theta_i^*(\lambda), \lambda\} \{\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\}\end{aligned}$$

for some $\theta_i^*(\lambda)$ s on the line segment between $\hat{\theta}_{(-i)}(\lambda)$ and $\hat{\theta}(\lambda)$, and with $J_n\{\theta, \lambda\} = -n^{-1} \sum_{j=1}^n \partial_{\theta} \phi(Z_j, \theta, \lambda)$. Because of the above, we can write

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1} J_n\{\theta_i^*(\lambda), \lambda\}^{-1} \phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}.$$

Since there exists a function $m_1(z)$ bounding all partial derivatives of $\phi(z, \theta, \lambda)$ with respect to θ , $J_n(\theta, \lambda)$ converges uniformly to $J(\theta, \lambda) = -E \partial_{\theta} \phi(Z, \theta, \lambda)$ over the compact parameter space $\Theta \times \Lambda$. Furthermore matrix inversion is uniformly continuous over compact sets not containing non-invertible matrices, ensuring that also $-J_n(\theta, \lambda)^{-1}$ converges to $-J(\theta, \lambda)^{-1} = E \partial_{\theta} \phi(Z, \theta, \lambda)^{-1}$ uniformly in λ and θ . Then $-J_n\{\theta_i^*(\lambda), \lambda\}^{-1} = -J\{\theta_i^*(\lambda), \lambda\}^{-1} + \epsilon_n^i$ where $\epsilon_n^i = o_p(1)$ uniformly in λ and i . By assumption, the eigenvalues of $J(\theta, \lambda)^{-1}$ are bounded by a number with absolute value C_1 . Then $\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\| \leq n^{-1}(C_1 + \delta_n) \|\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}\|$, where $\delta_n = o_p(1)$ does not depend on λ or i . Lastly by assumption $\|\phi(z, \theta, \lambda)\| \leq m_0(z)$. The above then guarantees

$$\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\| \leq n^{-1}(C_1 + \delta_n) m_0(Z_i) \quad (\text{S1.6})$$

for all λ . This proves (S1.2).

By the previous arguments

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1} [J\{\theta_i^*(\lambda), \lambda\}^{-1} + \epsilon_n^i] \phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}. \quad (\text{S1.7})$$

where $\epsilon_n^i = o_p(1)$ uniformly in i and λ . We will again use Taylor expansions to arrive at a result. We start with the second factor on the right hand side: $J\{\theta_i^*(\lambda), \lambda\}^{-1} = J\{\theta_0(\lambda), \lambda\}^{-1} + \left(\partial_{\theta_i^{**}(\lambda)} J\{\theta, \lambda\}^{-1}\right)\{\theta_i^*(\lambda) - \theta_0(\lambda)\}$, where $\theta_i^{**}(\lambda)$ is some vector on the line segment between $\theta_0(\lambda)$ and $\theta_i^*(\lambda)$. Now since $\theta_i^*(\lambda)$ lies on the line segment between $\hat{\theta}(\lambda)$ and $\hat{\theta}_{(-i)}(\lambda)$, $\|\theta_i^*(\lambda) - \theta_0(\lambda)\| \leq \|\hat{\theta}(\lambda) - \theta_0(\lambda)\| + \|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\|$. Combining this with Lemma S1.1 and (S1.6) ensures $\|\theta_i^*(\lambda) - \theta_0(\lambda)\| \leq R_n + n^{-1}(C_1 + \delta_n)m_0(Z_i)$, where $R_n = o_p(1)$ uniformly in λ and does not depend on i . In addition, the eigenvalues of $\partial_\theta J(\theta, \lambda)^{-1}$ are bounded by C_2 . Hence, $J\{\theta_i^*(\lambda), \lambda\}^{-1} = J\{\theta_0(\lambda), \lambda\}^{-1} + r_n^i(\lambda)$ where $\|r_n^i(\lambda)\| \leq n^{-1}(C_1 + \delta_n)C_2m_0(Z_i)$ uniformly in λ .

Secondly, we will work with the third factor on the right hand side of (S1.7). Taylor expanding the function $\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\}$ around $\theta_0(\lambda)$, reveals

$$\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\} = \phi\{Z_i, \theta_0(\lambda), \lambda\} + \partial_{\theta_i^{***}}\phi(Z_i, \theta, \lambda)\{\hat{\theta}_{(-i)}(\lambda) - \theta_0(\lambda)\},$$

for some $\theta_i^{***}(\lambda)$ on the line segment between $\theta_{(-i)}(\lambda)$ and $\theta_0(\lambda)$. Since there exists $m_1(z)$ bounding the norm of each component of $\partial_\theta \phi(z, \theta, \lambda)$ and depending only on z , we can use arguments similar to those above to show $\phi\{Z_i, \hat{\theta}_{(-i)}(\lambda), \lambda\} = \phi\{Z_i, \theta_0(\lambda), \lambda\} + s_n^i(\lambda)$ where $\|s_n^i(\lambda)\| \leq n^{-1}(C_1 + \delta_n)m_1(Z_i)m_0(Z_i)$ uniformly in λ and for all i .

We are now ready to complete the proof. Combining all of the previous arguments shows

$$\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1}[J\{\theta_0(\lambda), \lambda\}^{-1} + r_n^i(\lambda) + \epsilon_n^i][\phi\{Z_i, \theta_0(\lambda), \lambda\} + s_n^i(\lambda)],$$

which imply $\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda) = -n^{-1}J\{\theta_0(\lambda), \lambda\}\phi\{Z_i, \theta_0(\lambda), \lambda\} + E_n^i(\lambda)$ where

$$\|E_n^i(\lambda)\| \leq n^{-1}\{a_{1,n} + a_{2,n}m_0(Z_i) + a_{3,n}m_0(Z_i)^2 + a_{4,n}m_0(Z_i)^2m_1(Z_i)\}$$

and $a_{j,n} = o_p(1)$ uniformly in λ and does not depend on i for $j = 1, 2, 3, 4$. This concludes the proof.

■

The first part of Lemma S1.3 shows Lemma 1 in the article. We will now use the second part to show Theorem 6.

S1.1 Proof of Theorem 6

Taylor's theorem ensures, $\sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = \sum_{i=1}^n \psi(Z_i, \hat{\theta}) + \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T (\hat{\theta}_{(-i)} - \hat{\theta})$ for some θ_i^* s on the line segments between $\hat{\theta}$ and $\hat{\theta}_{(-i)}$. By assumption, all conditions of Lemma S1.3 hold true, and hence,

$$\sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T (\hat{\theta}_{(-i)} - \hat{\theta}) = -n^{-1} \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T J^{-1} \varphi(Z_i, \theta_0) + \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T E_N^i. \quad (\text{S1.8})$$

By the same lemma and existence of p_1 , we have

$$\begin{aligned} \left\| \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T E_N^i \right\| &\leq a_{1,n} n^{-1} \sum_{i=1}^n p_1(Z_i) + a_{2,n} n^{-1} \sum_{i=1}^n m_0(Z_i) p_1(Z_i) \\ &\quad + a_{3,n} n^{-1} \sum_{i=1}^n m_0(Z_i)^2 p_1(Z_i) + a_{4,n} n^{-1} \sum_{i=1}^n m_0(Z_i)^2 m_1(Z_i) p_1(Z_i). \end{aligned}$$

Since the moments of all terms in the above equation exists by assumption and $a_{j,n} = o_p(1)$ for $j = 1, 2, 3, 4$, the second term on the right hand side of (S1.8) is $o_p(1)$.

For the first term of (S1.8), note that by Taylor's theorem $\nabla_{\theta_i^*} \psi(Z_i, \theta) = \nabla_{\theta_0} \psi(Z_i, \theta) + H_{\theta_i^{**}} \psi(Z_i, \theta) (\theta_i^* - \theta_0)$ for some θ_i^{**} on the line segment between θ_i^* and θ_0 . By the triangle inequality and the fact that θ_i^* lies on the line segment between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$, $\|\theta_i^* - \theta_0\| \leq \|\hat{\theta} - \theta_0\| + \|\hat{\theta} - \hat{\theta}_{(-i)}\|$. Hence, by Lemma S1.3, we have

$$\begin{aligned} &\left\| -n^{-1} \sum_{i=1}^n \nabla_{\theta_i^*} \psi(Z_i, \theta)^T J(\theta_0)^{-1} \varphi(Z_i, \theta_0) - n^{-1} \sum_{i=1}^n \nabla_{\theta_0} \psi(Z_i, \theta)^T J(\theta_0)^{-1} \varphi(Z_i, \theta_0) \right\| \leq \\ &\|\hat{\theta} - \theta_0\| n^{-1} C \sum_{i=1}^n p_2(Z_i) m_0(Z_i) + a_{0,n} n^{-1} C \sum_{i=1}^n p_2(Z_i) m_0(Z_i)^2, \end{aligned}$$

for some C . This expression is $o_p(1)$ since the expected value of both means are finite

and $\|\hat{\theta} - \theta_0\|$, $a_{0,n} = o_p(1)$ by Lemma S1.3. This concludes the proof.

S2 Appendix - When the limit function is flat

In Theorem 1, we assumed that λ_0 was the unique minimizer of the limiting risk $R(\lambda) = E \psi\{Z, \theta_0(\lambda)\}$. In certain cases, however, $\theta_0(\lambda)$, the solution to $E \varphi\{Z, \theta, \lambda\} = 0$, is constant as a function of λ , and as a result, $R(\lambda)$ is flat. This is the case for hybrid models on the form $\varphi(z, \theta, \lambda) = \lambda \varphi_1(z, \theta) + (1 - \lambda) \varphi_2(z, \theta)$ when $E \varphi_1(Z, \theta)$ and $E \varphi_2(Z, \theta)$ share roots, e.g. in the hybrid generative-discriminative model when the generative model is specified correctly or for hybrid combinations of empirical and hybrid likelihood functions (see Dæhlen and Hjort (2024)) under model conditions. In fact, when $R(\lambda)$ is flat, the limit of both $TE(\lambda)$ and $CV(\lambda)$ are constant, and there is no obvious “best” value for λ , making it unclear what $\hat{\lambda}$ even should be aiming for were it consistent. We will now use informal mathematical arguments and simulations to illustrate why a clear and informative limit result for $\hat{\theta}(\hat{\lambda})$ is hard to find when $\theta_0(\lambda)$ is constant.

Assume that $\theta_0(\lambda)$ is constant and equal to θ_0 . To investigate the behaviour of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ we will first take closer look at the function $CV(\lambda)$. By Theorem 6, the following holds pointwise,

$$CV(\lambda) = n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}(\lambda), \lambda\} - n^{-1} \text{Tr}\{J(\theta_0, \lambda)^{-1} V(\theta_0, \lambda)\} + o_p(1/n).$$

Taylor expanding the first term on the right hand side in the above equation, reveals

$$\begin{aligned} n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}(\lambda), \lambda\} &= n^{-1} \sum_{i=1}^n \psi(Z_i, \theta_0, \lambda) + \{\hat{\theta}(\lambda) - \theta_0\}^T n^{-1} \sum_{i=1}^n \nabla_{\theta_0} \psi(Z_i, \theta, \lambda) \\ &\quad + \{\hat{\theta}(\lambda) - \theta_0\}^T \frac{1}{2n} \sum_{i=1}^n H_{\theta^*} \psi(Z_i, \theta, \lambda) \{\hat{\theta}(\lambda) - \theta_0\}, \end{aligned}$$

for some θ^* on the line segment between $\hat{\theta}(\lambda)$ and θ_0 . Assume θ_0 is the true minimizer of $E\psi(Z, \theta)$, i.e. that $E\nabla_{\theta_0}\psi(Z, \theta) = 0$. Then, arguments similar to those given in the proof of Lemma 4.2 and Theorem 6 show

$$\begin{aligned} S_n(\lambda) &= n \left(\text{CV}(\lambda) - \frac{1}{n} \sum_{i=1}^n \psi(Z_i, \theta_0) \right) \\ &= U_n(\lambda)^T V_n - \text{Tr}\{J(\theta_0, \lambda)^{-1} V(\theta_0)\} + \frac{1}{2} U_n(\lambda)^T W_n U_n(\lambda) + o_p(1), \end{aligned}$$

where $U_n(\lambda) = \sqrt{n}\{\hat{\theta}(\lambda) - \theta_0\}$, $V_n = n^{-1} \sum_{i=1}^n \nabla_{\theta_0}\psi(Z_i, \theta)$, and $W_n = n^{-1} \sum_{i=1}^n H_{\theta_0}\psi(Z_i, \theta)$. Since θ_0 is constant with respect to λ , we have $\hat{\lambda} = \text{argmin}_{\lambda} S_n(\lambda)$.

Note that W_n is consistent for $W = E H_{\theta_0}\psi(Z, \theta)$ by the assumptions of Theorem 6. Furthermore, it does not depend on λ . If, in addition, the conditions of Van der Vaart (2000, Ex. 19.7) are fulfilled for $(z, \lambda) \mapsto (J(\theta_0, \lambda)^{-1}\varphi(z, \theta_0, \lambda), \psi(z, \theta_0))$, the process $\lambda \mapsto (U_n(\lambda), V_n)$ converges as a process to a Gaussian process $(U(\lambda), V)$ with covariance structure as described in the beginning of Van der Vaart (2000, Sec. 19.2). This, in combination with Slutsky's lemma and the continuous mapping theorem, is sufficient for S_n to converge in distribution in Skorokhod space to the following limit process

$$S(\lambda) = U(\lambda)^T V - T(\lambda) + \frac{1}{2} U(\lambda)^T W U(\lambda), \quad (\text{S2.9})$$

where $T(\lambda) = \text{Tr}\{J(\theta_0, \lambda)^{-1} V(\theta_0)\}$.

Direct computations show $E S(\lambda) = (1/2) \text{Tr}\{W J(\lambda)^{-1} K(\lambda) J(\lambda)^{-1}\}$. Hence, if there exists a λ_0 such that $\hat{\theta}(\lambda_0)$ is equal to the maximum likelihood estimator, the Cramer-Rao theorem guarantees that λ_0 minimizes $E S(\lambda)$. Since such a λ_0 often exists in the case where $\theta_0(\lambda)$ is constant, it is tempting to think that $\hat{\lambda}$ converges in probability to this λ_0 . Sadly this conclusion cannot be drawn, as S is a random process and there is no guarantee that minimizing the expected value of S minimizes the process itself. Furthermore, the convergence of S_n towards S is in distribution. Because of this, convergence in probability of $\hat{\lambda}$ towards λ_L , the minimizer of S , does

not follow. We can only deduce $\hat{\lambda} \xrightarrow{d} \lambda_L$, and since S is a stochastic process and not a nonrandom function, its minimizer is not guaranteed to be a constant. Because of this, λ_L is a random variable and $\hat{\lambda} \xrightarrow{d} \lambda_L$ does not imply convergence in probability. The above therefore does not show consistency of $\hat{\lambda}$ towards λ_0 or any fixed value for that matter, but rather that the distribution of $\hat{\lambda}$ tends towards that of the minimizer of S .

To illustrate the above, we performed a small simulation experiment for a hybrid LDA-logistic regression model with one covariate and binary response variable. We simulated from an LDA model, under which the true parameter minimizes both the limit of the generative and logistic likelihood. Hence $\theta_0(\lambda)$ is constantly equal to the true parameter in this simulation. We used a $\psi(y, x, \theta) = (y - \exp\{\beta(\theta)^T x\} / [1 + \exp\{\beta(\theta)^T x\}])^2$, a smoothed version of the 0 - 1 loss function. With $n = 2000$ we computed $B = 1000$ different realizations of S_n . A selection of sample paths is displayed in Figure 1 together with a histogram of the corresponding minima. From the figure we see that the sample paths are indeed not constant and that although the minimum is most often one, other values are repeatedly chosen, implying that λ_L might really not be constant in this case.

Although the above arguments and simulations seem to indicate that there is reasonable doubt that $\hat{\lambda}$ is consistent for any one value, this does not necessarily rule out all hope that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ has a normal limiting distribution. In the simulation experiment described above, we also computed $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ in each iteration. Histograms of each component of this vector are displayed in Figure 2. From the plots we note that the distributions at least look quite normal. This intuition is supported by the QQ-plots shown in Figure 3. Looking at these plots, we see no obvious deviances from the corresponding normal distributions.

One possible solution is to work with $A_n(\hat{\lambda}) = \sqrt{n}K(\theta_0, \lambda_0)^{-1/2}J(\theta_0, \hat{\lambda})\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ rather than $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$. Pointwise, we have $A_n(\lambda) \xrightarrow{d} N(0, I_p)$, and hence, we would expect $A_n(\hat{\lambda}) \xrightarrow{d} N(0, I_p)$ as well. To check whether this idea seems reasonable, we computed $A_n(\hat{\lambda})$ in each iteration of the simulation setting described above. A

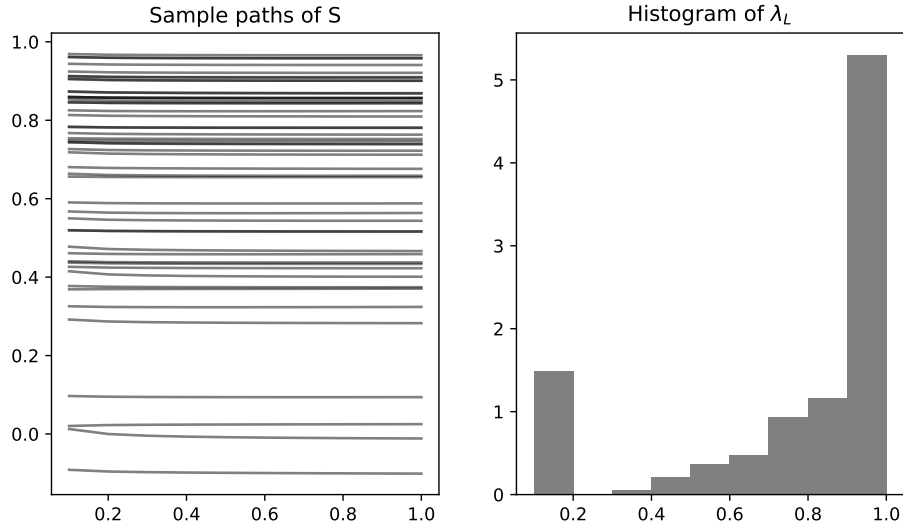


Figure 1: The figure shows two plots. To the left $B = 1000$ realizations of S_n with $n = 2000$ are displayed. On the right hand side a histogram of the corresponding minima is shown.

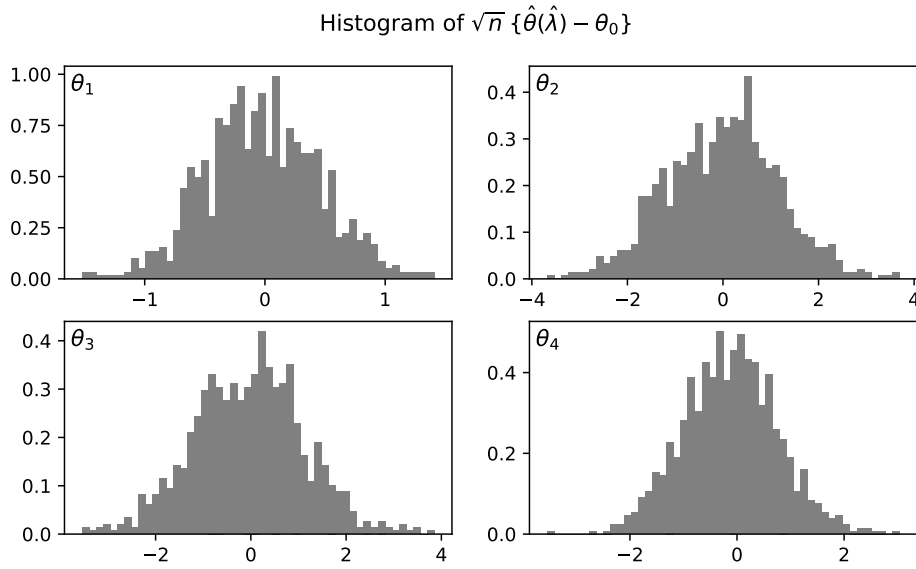


Figure 2: The figure shows histograms of each component of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$.

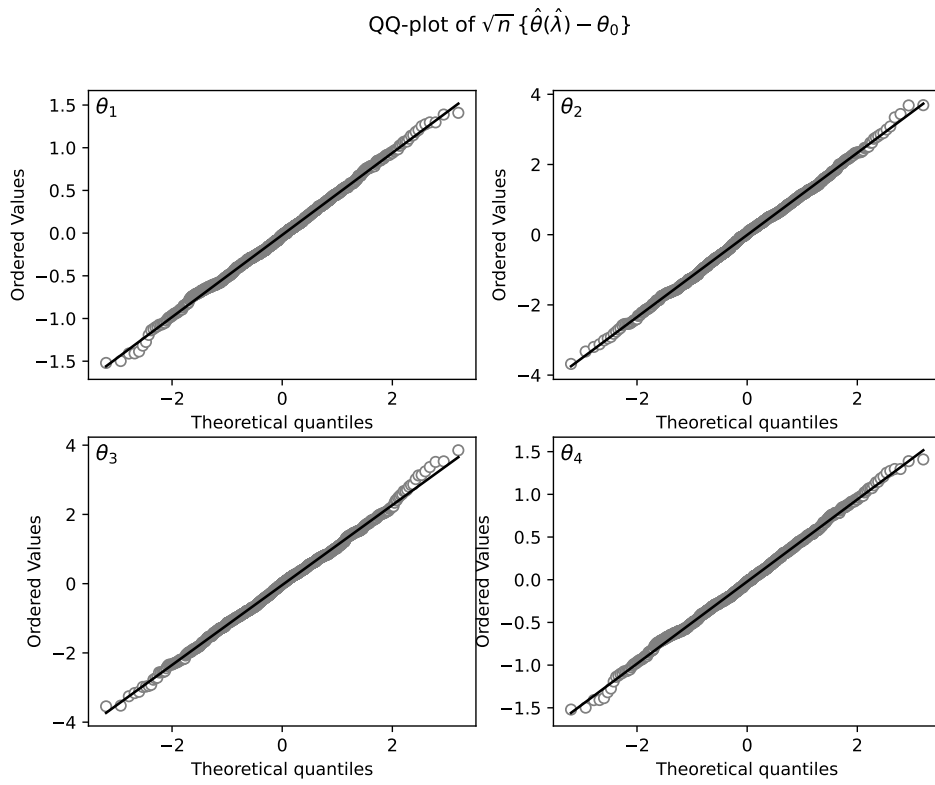


Figure 3: The figure shows QQ-plots for each component of $\sqrt{n} \{ \hat{\theta}(\hat{\lambda}) - \theta_0 \}$. The empirical quantiles are compared to those of fitted normal distributions.

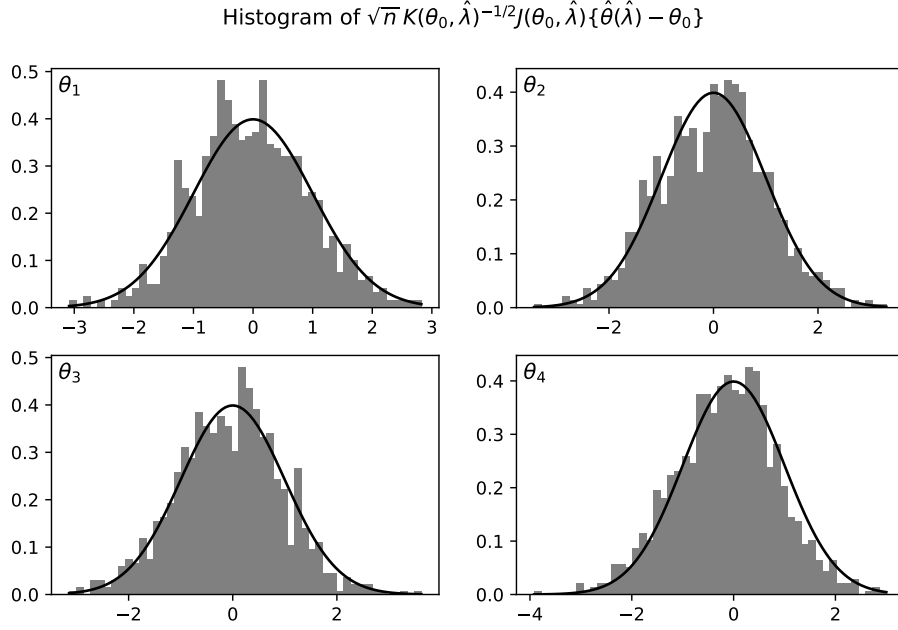


Figure 4: The figure shows histograms of each component of $\sqrt{n} K(\theta_0, \hat{\lambda})^{-1/2} J(\theta_0, \hat{\lambda}) \{\hat{\theta}(\hat{\lambda}) - \theta_0\}$.

histogram of each component of $A_n(\hat{\lambda})$ is shown in Figure 4 together with the graph of density function for the standard normal distribution. The plot seem to indicate that $A_n(\hat{\lambda})$ behaves more or less like a standard normally distributed variable.

One could argue that Corollary 1 and Theorem 5 failing to cover situations where the limiting risk function is flat, makes the result and the estimators of Theorem 4 less useful in practical situations. Say that we for instance, have fitted a generative hybrid-discriminative model to some data and want to apply the theory in the main part of this paper to construct confidence intervals and hypothesis tests for the parameters in the model. If the generative model fitted to the data is wrongly specified, this can be done by applying Corollary 1 or Theorem 5 in conjunction with Theorem 4. If the model conditions do hold true, however, the arguments made in this section shows that the results cannot be applied. Hence, we cannot say that Corollary 1, Theorem 5 or the estimators in Theorem 4 are truly model-agnostic, since they fail to apply when the limiting risk function is flat. In theory, this is of course quite problematic. In practice, however, this is less of a problem. When $\theta'_0(\lambda) = 0$ for all values of λ , $E \varphi(Z, \theta_0, \lambda) = 0$

for all λ . Because of this, $\hat{\theta}'(\hat{\lambda})$ of Theorem 3 is consistent for 0 and $\hat{A}^*$ in Theorem 4 is asymptotically equivalent with $(J^{-1}(\hat{\lambda}), 0, 0)$. Hence, $\hat{A}^* \hat{K}(\hat{\lambda})^* (\hat{A}^*)^T$ is asymptotically equivalent with $J^{-1}(\hat{\lambda}) K(\hat{\lambda}) J^{-1}(\hat{\lambda})$ which is what the simulations of this section seem to indicate is the limiting variance of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$.

References

- Dæhlen, I. and N. L. Hjort (2024). Model robust hybrid likelihood. *arXiv preprint arXiv:2409.15975*.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.

