Some applications of stochastic process techniques to statistics

by

Steffen Grønneberg

THESIS

Dissertation presented for the degree of

PHILOSOPHIÆ DOCTOR

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1133

ISSN 1501-7710

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Unipub.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Preface

This thesis is dedicated to my grandmother, Ada Madssen, who meant the world to me.

Contents

Chapter 1

Introduction to the Thesis

This thesis studies certain mathematical aspects of model selection, statistical estimation theory and probability using stochastic process tools. Except for the stochastic process tools that the our investigations use, it must be admitted up front that the papers of the this thesis really concerns three different problems. An introduction to a PhD thesis should summarize its papers through placing them in connection with each other and in a broader context, as well as discussing their interrelations in a wider perspective. As the enclosed papers are all of a somewhat separate character, I have chosen to decrease the focus typical for such introductions.

The introduction begins with Section 1 that describes the start of probability, both in the ancient rhetorical sense and in the mathematical sense starting around 1660. I will use this description as an anchor to connect the thesis' papers through a somewhat speculative discussion constituting the remainder of the introduction. I hope the trained philosopher will forgive my amateur efforts in using philosophical considerations as a tool to try to connect the papers.

Section 2 introduces the enclosed paper "On the errors committed by sequences of estimator functionals", which is accepted for publication in the international journal Mathematical Methods of Statistics. We will look at how to calculate probabilities related to the most fundamental law of probability: The weak and strong laws of large numbers and their uniform extensions.

Section 3 introduces the papers "The Copula Information Criterion and its implications for the Maximum Pseudo Likelihood Estimator" and "The Copula Information Criteria". The first paper is published in the book "Dependence modeling – Vine Copula Handbook" and was written by invitation. In many ways, it serves as an introduction to the more technical paper "The Copula Information Criteria", which is submitted for publication. To avoid repetition, we will introduce the concepts involved in model selection in general – rather than focusing solely on the copula information criterion. Section 3.3 provides a perhaps surprising connection

between "On the errors committed by sequences of estimator functionals" and the AIC-heuristics used in "The Copula Information Criteria".

Section 4 introduces the paper "Estimation and inference for jump regression models". This paper deals with a somewhat non-standard regression problem from both the Bayesian and frequentist perspective. Our basic set-up is observations $y_1, \ldots, y_n$ of the form

$$y_i = m(x_i, \theta) + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n,$$

where m is a step function with steps specified by the covariates $x_1, x_2, \ldots, x_n$ and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is mean zero Gaussian noise. We derive asymptotics for estimates of the parameters specifying m, and show that Bayesian estimation is superior to ML estimation from a frequentist perspective through using theory from Ibragimov & Khasminskii (1981)

1. From the ancients to 1640

Statistics can be widely described as decision making under uncertainty. Uncertainty is lack of information, and uncertain knowledge has been considered to be second class knowledge almost throughout western history. However, as the ancient golden age pre-Socratic Greek Sophist Gorgias puts it in his controversial essay, Encomium of Helen,

> For if all men on all subjects had memory of the past, (understanding) of the present, and foresight into the future, *logos* would not be the same in the same way; but as it is, to remember the past, to examine the present, or to prophesy the future is not easy; and so most men on most subjects make opinion (*doxa*) an adviser to their minds. But opinion is perilous and uncertain, and brings those who use it to perilous and uncertain good fortune.

This is just one part of an argument on how Helen of the Illiad is not to blame for escaping to Troy. The style of the encomium is such that it could be the lines of an Oscar Wildean dandy. However, the above pragmatic description of certainty and understanding is in clear contrast to Plato – whose main criticism against the sophists is precisely against their use of probabilistic statements (Gagarin, 1994). In the dialogue Phaedrus (267a), Plato – perhaps the greatest rhetorician of the western civilisation – ridicules the rhetoricians' by "We will let Tisias and Gorgias rest in peace, who saw that probabilities should be more honoured than truths, and who make small things appear great and great things small by the power of speech." Concerning Plato's critique of the Sophists, Gagarin (1994) says

> Plato provides no evidence to support his statement about the value of probability; none the less, critics ever since have largely accepted his views. [...] In sum, there is no evidence to support

> Plato's claim, echoed nearly unanimously by modern scholars, that Greek orators and rhetoricians valued probability more highly than the truth. Undoubtedly probability was sometimes used to support a false case, but so too was direct evidence; and the surviving speeches, at least, indicate that orators at this time had a clear and reasonable understanding of the value of probability arguments and considered them valid only to the extent that direct evidence for the truth was absent or inconclusive. Plato's criticisms on this point reflect his own concern with the overriding primacy of an absolute standard of truth, which is tied to and validated by his Forms; for him anything less than absolute truth was no truth at all.

The probabilities of the Sophists were what we would consider intuitive subjective probabilities. As an example, Aristotle attributes the following two arguments to the Corax of Syracuse (who, along with Tisias, is seen as the founder of ancient Greek rhetoric) in his Rhetoric 1402a17-28: Suppose that a fight has broken out between a weak and a strong man. The weak man uses the following probabilistic argument for his innocence: It is not likely that he, a weak man, assaulted a strong man. The other counters with more sophisticated probabilistic reasoning: He is not likely to have assaulted a weak man, since he, a strong man, would immediately be suspected of the crime. This argument is quite far away from our mathematically formalized probabilistic reasoning, but as a rhetorical technique, it is part of a strain of ideas that has been in continual use ever since.

Garber & Zabell (1979) summarizes the development of probabilistic arguments in the rhetorical tradition until the emergence of mathematical probability around 1640. And while it is true that some ur-concept of probability is traceable to the sophists, Ian Hacking argues in the preface of the second edition of Hacking (1975) that the network of ideas containing the rhetoricians probabilistic arguments are quite separate from the developments leading to the mathematical formalization of probability around 1640. Mathematical statistics is concerned with the study of statistics using a formalized concept of probability. With the greatest ease, modern statistics rely on advanced mathematical constructs such as abstract Brownian motion processes, whose existence and properties rely specifically on our axiomatization of probability and modern mathematics. In the next section, we will study a very basic problem of probability using these advanced tools, and surprisingly meet the limitations of the currently accepted framework quite easily in the form of non-measurable "random variables".

A strange and surprising feature of mathematical probability is that it is a fundamentally dual concept: Probability concerns both subjective and frequentist phenomena. Hacking (1975, p 43) says "any theory on the emergence of probability must

try to explain why the concept that emerged was dual in just this way." Hacking (1975, p 12) describes this as follows

> It is notable that the probability that emerged so suddenly is Janus-faced. On the one side it is statistical, concerning itself with stochastic laws of chance processes. On the other side it is epistemological, dedicated to assessing reasonable degrees of belief in propositions quite devoid of statistical background. (...) Pascal himself is representative. His famous correspondence with Fermat discusses the division problem, a question about dividing stakes in a game of chance that has been interrupted. The problem is entirely aleatory in nature. His decision-theoretic argument for belief in the existence of God is not. It is no matter of chance whether or not God exists, but it is still a question of reasonable belief and action to which the new probable reasoning can be applied.

Hacking (1975) develops a now famous thesis on this development. He describes his program on page 16 as

> I am inviting the reader to imagine, first of all, that there is a space of possible theories about probability that has been rather constant from 1660 to the present. Secondly, this space resulted from a transformation upon some quite different conceptual structure. Thirdly, some characteristics of that prior structure, themselves quite forgotten, have impressed themselves on our present scheme of thought. Fourth: perhaps an understanding of our space and its preconditions can liberate us from the cycle of probability theories that has trapped us for so long. This last picture has a familiar ring. The picture is, formally, the same as the one used by the psychoanalysts and by the English philosophers of language. "Events preserved in memory only below the level of consciousness", "rules of language that lie deep below the surface" and "a conceptual space determined by forgotten preconditions": all three have, of course, a common ancestor in Hegel.

The basis for his theory is the French philosopher Foucault's discussion on the sign in his book Foucault (1966). Foucault's historical programme in Foucault (1966) can be described as taking the problem of anachronisms seriously. His central concept is that of *epistemes*, which is the conscious and subconscious assumptions and requirements a time and culture demands of knowledge-claims. Many epistemes can coexist, they can change and develop and have complex interplay. His central program, which he calls the archaeological method, is to follow the origins and developments of epistemes by analysing primary sources from the time under study.

The thesis of Hacking (1975) is summarized as follows. In addition to the Great Chain of Being – which describes an hierarchical structure of nature from the lowliest stones, up to plants, to man and up to angels and finally God – a central part of late medieval thought was the understanding that the world was connected through similitudes, analogies and *signs*. Everything is connected, and each part of the world is connected with everything else through these connections. This was not a poetic image, but literal one: If an herb looked like a human organ, one should interpret this as a sign: this herb might have healing powers for the similar organ. In occult Christianity, this was extended to the thought that such signs were not only similarities – but God's writing in nature. If one could decipher these signs one could *read the thoughts of God*. The alchemists and physicians had intricate systems of interpretation to reach nature's secrets. And while the alchemists wanted to read God's thoughts to come closer to Him, the physicians wanted to establish a science based on these signs. The problem with such signs is that some signs are always valid, while others are only valid sometimes. The signs that are not always regular were called "signs with probability" (Hacking, 1975, p. 43). These signs had two types of uncertainty: A subjective uncertainty – one did not always manage to read God's signs correctly – and frequentist uncertainty – the sign's power does not necessarily come into force; the herbal medicine does not always work. Hacking (1975, p. 44-45) describes the connection between these thoughts and the emergence of probability as follows.

> The sign-as-evidence indicates with probability, but I do not claim that the authors who employed it where an "influence" on the founding fathers of probability. Some historians of ideas are much concerned with the way in which work A can influence his successor B. Two kinds of influence are considered. B may deliberately choose to employ central concepts or techniques of A, or else B may unwittingly pursue a programme initiated by A. Such talk of "Influence" is part of the historian's language of precursors and anticipations. It would be amazing if Paracelsus [An alchemist physician discussed in the connection of reading the thoughts of God, and an inspiration to the Faust-myth] were an "influence" on a Pascal or a Leibniz. The mathematicians despised what they knew of the occult. Yet their contempt for those earlier hermetical figures does not preclude the possibility that whenever these geometers thought about opinion, they thought in a conceptual space that was the legacy of the very empirics whom they scorned. The intellectual objects about which, and *in* which, the new mathematicians thought had been formed in the crucibles of the alchemists and the vials of the physicians.

After this discussion, Hacking (1975) continues to extend Foucault's theories, stated on page 70 of Foucault (1966) as follows.

> If we question Classical[1] thought at the level of what, archaeologically, made it possible, we perceive that the dissociation of the sign and resemblance in the early seventeenth century caused these new forms – probability, analysis, combination and universal language system – to emerge, not as successive themes engendering one another or driving one another out, but as a single network of necessities. And it was this network that made possible the individuals we term Hobbes, Berkeley, Hume or Condillac.

This places the emergence of probability as a crucial ingredient of seventeenth and eighteenth century thought.

While the above may quotations seem somewhat wild, and it may seem very unscientific to rely on similitudes in the study of medicine, these old medieval categories of inference are still very much in use today as the basis for discovery. This is the case, even in pure mathematics, as discussed thoroughly in Pólya (1945, 1954). Mathematical exploration and discovery very much rests on these types of inferences, and learning to do advanced mathematics may in some sense be thought of as learning how to use the medieval categories of similitudes, analogies and signs – while checking the resulting uncertain inferences through stringent deductions. It is most unfortunate that this very important final step is unavailable for inference regarding the real world.

The first major work of mathematical probability theory is Jacques Bernoulli's *Ars conjectandi*. Chapter 17 of Hacking (1975) describes its main mathematical content as follows.

> Chapter 5 of Part IV of *Ars cojectandi* proves the first limit theorem of probability theory. The intended interpretation of this result is still a matter of controversy, but there is no dispute about what Bernoulli actually proved. He takes for granted a chance set-up on which he can make repeated trials. There is a constant unknown chance p of "success" S on any given trial. When n trials are made a proportion s_n of successes is observed. Bernoulli proves what is now called the weak law of large numbers: the probability of an n-fold sequence in which $|p - s_n| < \varepsilon$ increases to 1 as n grows without bound. Moreover, for any given error ε, he shows how to compute a number n such that the probability of getting s_n in the interval $[p - \varepsilon, p + \varepsilon]$, itself exceeds any given probability $1 - \delta$. In particular, if $(1 - \delta) = 0.999$, we have a moral certainty that s_n will fall in the assigned interval. For example

[1]That is, the time between around 1750 to 1830, not the classical period of the ancients.

> if p is $3/5$ then a moral certainty of error less than $1/50$ is guaranteed by an n in excess of 25 550.

Frequentist probability is fundamentally thought of through the law of large numbers. Stability of long term frequencies is in our backbone when it comes to probability, and yet any real world connection is clearly a theoretical postulate. Proving the law of large numbers is in some sense circular: It must be valid, otherwise the frequentist probability formalism does not make sense almost by definition.

The law of large numbers is in it self a rather empty result. In contrast, an error bound is much more directly connected with the real world. We will discuss such error bounds rather thoroughly in the following section, and here we will only mention that we can do much better than the bound of Bernoulli: We get that $n = 6773$ is the exact uniform bound, reached precisely when $p = 1/2$. It would seem that a simple test of this claimed connection between the probability model and the real world by throwing a fair coin $n = 6773$ times. However, we would need to do this many times to assess the claim that $|p - s_n| > \varepsilon$ in no more than 0.1% of the time. How many times must we perform this experiment in order to formally test this hypothesis? We regress into an infinite loop which strictly speaking cannot be resolved without some leap of faith.

In the case of a coin, we can be highly convinced of its long term frequency distribution by the several laborious experiments performed by various people lacking any strong sense of their own mortality and limited time as corporeal beings. For more complex phenomena, such as non-repeatable stochastic processes like the stock market, we cannot even in theory check the various probability statements we casually make in the statistics literature. And, to take this line of thought to its limit: we cannot ever repeat the exact conditions of an experiment. Probability models depend crucially on our modelling assumptions, and the model specification is in part a subjective process.

2. The first limit theorem and the variable N_ε

We now move on to present the paper "On the errors committed by sequences of estimator functionals", which is a work in probability theory motivated by statistical concerns. Our basis will be the Bernoulli bound presented in the previous section.

2.1. Improvements on the Bernoulli bound. A modern reader will not be impressed by Bernoulli's error-bound of $n \geq 25\,550$. His proof is based upon a detailed analysis of the binomial coefficients, and he would be shocked to learn how easily his result can be improved by the use of the Chebyshev-inequality. As it is clear that for any random variable X, we have that

$$X = X \times 1 = XI\{X \geq \varepsilon\} + XI\{X < \varepsilon\} \geq XI\{X \geq \varepsilon\} \geq \varepsilon I\{X \geq \varepsilon\},$$

the linearity and monotonicity of expectation shows the Chebyshev inequality $P(X \geq \varepsilon) \leq \mathbb{E}X/\varepsilon$. Hence, sub-additivity and the Chebyshev inequality gives

$$\begin{aligned} P\{|X| \geq \varepsilon\} \leq & P(X \geq \varepsilon) + P(X < -\varepsilon) = P(X \geq \varepsilon) + P(-X \geq -\varepsilon) \\ = & P(e^{\lambda_1 X} \geq e^{\lambda_1 \varepsilon}) + P(e^{\lambda_2 X} \geq e^{\lambda_2 \varepsilon}) \leq \mathbb{E}e^{\lambda_1 (X-\varepsilon)} + \mathbb{E}e^{-\lambda_2 (X-\varepsilon)} \end{aligned}$$

for any $\lambda_1, \lambda_2 > 0$. Now let $S_n = \sum_{i=1}^n X_i$, where X_i are independent with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. Calculus enables us to further bound the above inequality, see e.g. Chapter 1.6 of Shiryaev (1995), which gives

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}. \tag{2.1}$$

Note that in contrast to Bernoulli's statement, this inequality is uniform in p. Hence, for any p, we are guaranteed that

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \geq 1 - \delta \tag{2.2}$$

is satisfied when

$$n_{\text{Exponential}} = \left\lceil \frac{\log(2/\delta)}{2\varepsilon^2} \right\rceil.$$

When $1-\delta = 0.999$ and $\varepsilon = 1/50$, we get $n = 9502$. This is still a crude bound. Any modern computer can easily calculate the exact solution, resulting in the comparison between the exponential bound and the exact uniform bound in Figure 1(a). The exact uniform bound is 6773.

These finite sample calculations may seem strange to the typical statistician: For sufficiently small ε and δ, it is clear that the Central Limit Theorem yields very good approximations. Such an approach would be based on the approximation

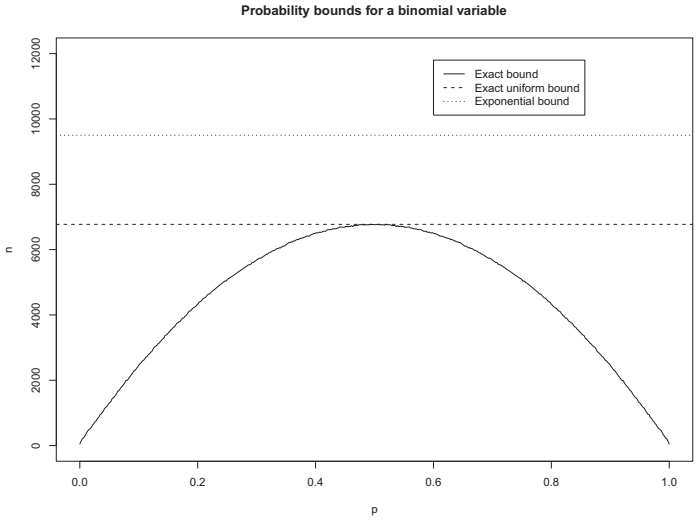
$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \approx P\left(|N\left(0, p(1-p)\right)| \geq \sqrt{n}\varepsilon\right) = 1 - 2\Phi\left(-\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right),$$

so that

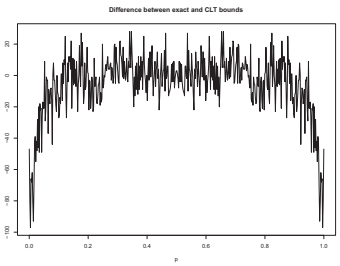
$$n_{\text{Normal}} = \left\lceil \frac{p(1-p)}{\varepsilon^2} \Phi^{-1}\left(\frac{\delta}{2}\right)^2 \right\rceil \tag{2.3}$$

suffices. On the scale of Figure 1(a), the exact solution and the solution based on the normal approximation are indistinguishable. The normal-approximated uniform bound is $\lceil \varepsilon^{-2}/4\Phi^{-1}(\delta/2)^2 \rceil = 6767$, impressively close to the exact solution 6773 – but slightly underestimated. Figure 1(b) shows the difference between the exact solution and n_{Normal}, while Figure 1(c) shows their relative error. These errors can be bounded by results such as the Berry–Esseen Theorem, but they differ in character from the exponential bound, in that they both overestimate and underestimate n.

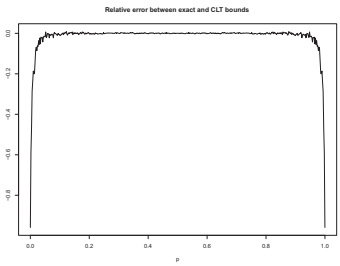
The above set-up is so simple that we can find exact solutions fairly easily. As illustrated by the example we will momentarily study, we often cannot. What the



(a) The solution of $P(|S_n/n - p| \geq 1/50) \geq 0.999$ with respect to n.



(b) The difference between the exact solution of eq. (2.2) and the approximation of eq. (2.3)



(c) Relative error between the exact solution of eq. (2.2) and the approximation of eq. (2.3)

FIGURE 1. Plots related to eq. (2.2).

above set-up does do though, is illustrate fundamental behavior of three types of calculations in statistics:

(1) Exact, or approximately exact calculations – which are often impossible or very difficult to find.
(2) Finite sample bounds – which are often skewed in a known direction.

(3) The asymptotic approach. That is, solving the problem when $n \to \infty$ or some other control variable approaching a limit. Typically, such approximations are skewed in some unknown direction, which varies according to the exact probabilistic law of the variables involved.

The choice of which of the above three computational methods to use is of fundamental practical importance in most areas of statistics. This problem is perhaps especially clear in the field of model selection, as we will see in the next section.

Reaching better bounds than the above exponential bound of eq. (2.1) has been a subject of intense research, summarized e.g. in Chapter 11.1 of Shorack & Wellner (1986). The reason for this great interest in the simple binomial case is that for an *iid* sequence $Y_1, Y_2, \ldots$, the variable $S_n/n = \sum_{i=1}^{n} X_i/n$ with $X_i = I\{Y_i \leq x\}$ is the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\}.$$

For a given x, $S_n = nF_n(x)$ is binomially distributed. As

$$P(\lim_{n\to\infty} S_n/n = p) = P\left(\bigcap_{\varepsilon>0} \bigcup_{n=1}^{\infty} \left\{\sup_{k\geq n} |S_k/k - p| < \varepsilon\right\}\right),$$

the convergence $S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} p$ is equivalent to

$$\lim_{n\to\infty} P(\sup_{k\geq n} |S_k/k - p| > \varepsilon) = 0 \tag{2.4}$$

for a given $\varepsilon > 0$ by continuity of probability measures. Sub-additivity and inequality (2.1), gives

(2.5)

$$\lim_{n\to\infty} P(\sup_{k\geq n} |S_k/k - p| > \varepsilon) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} P(|S_k/k - p| > \varepsilon) \leq \lim_{n\to\infty} \frac{2}{1 - e^{-2\varepsilon^2}} e^{-2n\varepsilon^2} = 0.$$

Hence,

$$F_n(x) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}I\{X_i \leq x\} = F(x). \tag{2.6}$$

The sharper and more advanced bounds for the tail of a binomial variable lead to e.g. uniform laws of iterated logarithms for the empirical distribution function (Shorack & Wellner, 1986). However, inequality eq. (2.1) is strong enough to prove the result Pitman & Pitman (1979) call "the existence theorem for statistics as a branch of applied mathematics" and Love (1977) calls "the fundamental theorem of statistics", namely the Glivenko-Cantelli Theorem

$$\sup_x |F_n(x) - F(x)| = \sup_x |F_n(x) - P(X \leq x)| \xrightarrow[n\to\infty]{\text{a.s.}} 0$$

valid for any F. Indeed, the monotonicity of $x \mapsto F_n(x)$ implies that the point-wise convergence of eq. (2.6) implies the uniform result, see the proof of Lemma 11.4.3 of Dudley (2003).

The weak law of large numbers $S_n/n \xrightarrow[n\to\infty]{\mathscr{P}} p$ is a purely asymptotic result, and error-bounds for finite n – such as those of inequality (2.1) – must be given to show that the asymptotics are of practical interest. The same applies to the strong law $S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} p$. The validity of (2.4) shows that

$$n \mapsto P(\sup_{k\geq n} |S_k/k - p| > \varepsilon)$$

is decreasing. A natural question is how fast such a convergence takes place. A fruitful formulation of this question is to investigate the law of

$$N_\varepsilon = \sup\{n : |S_n/n - p| > \varepsilon\},$$

i.e. the last time the distance between S_n/n and p is larger than ε – or, the last time an error larger than ε occurs. Indeed, N_ε is finite almost surely for each $\varepsilon > 0$ if $S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} p$ by the definition of limits, and conversely, $S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} p$ if $N_\varepsilon > \infty$ almost surely for each $\varepsilon > 0$ by eq. (2.4). The relation

$$P(N_\varepsilon > n) = P(\sup\{n : |S_n/n - p| > \varepsilon\} > y) = P(\sup_{k\geq n} |S_k/k - p| > \varepsilon)$$

reveals that the law of N_ε is intimetely related to the speed at which the discrete time stochastic process

$$n \mapsto \sup_{k\geq n} |S_k/k - p|$$

converges to zero.

In the current iid case, the law of N_ε is defined in terms of the random variable

$$\tilde{U} = (X_1, X_2, \ldots)$$

defined on the product space $\mathbb{R}^\infty$. Under the typical construction of such a space, such as the elementary construction in Theorem 6.2.4 of Stroock (2005), the law of

$$U = \sum_{m=1}^{\infty} 2^{-m} X_i$$

is a uniform random variable on $[0, 1]$ when $p = 1/2$. Hence, for Lebesgue measure λ, we have

$$P(N_\varepsilon > n) = \int_0^1 \sup_{k\geq n} \left| \sum_{i=k}^{\infty} y_{(i)} \frac{2^i}{k} - 1/2 \right| \mathrm{d}\lambda(y)$$

where $y_{(i)}$ is the i'th binary expansion of y. In contrast to the discrete law of S_n for finite n, we cannot simply instruct a computer to calculate this probability.

Considering the above list of possible ways of calculating the law of N_ε, the first method is not in general feasible, except for certain special cases. The enclosed

paper "On the errors committed by sequences of estimator functionals" studies approximations of the third kind on the above list, for the limit law of $\varepsilon^2 N_\varepsilon$ as $\varepsilon \to 0^+$. This is already a well-studied problem, but we extend these results to a much wider class of estimators than that which was previously known. Parallelling the approximation leading to eq. (2.3), our method is based on an advanced version of the central limit theorem. Before we introduce the arguments leading to these approximations, let us consider non-asymptotic bounds. These bounds are reached from the simple sub-additivity argument of eq. (2.5), and are hence rather crude. However, in presenting these bounds, we will introduce the mathematical structures needed to present the CLT-based approximations.

For the binomial case, which includes the empirical distribution function for a fixed x, eq. (2.5) already implies the tail-bound

$$P(N_\varepsilon > n) = P(\sup_{k \geq n} |S_k/k - p| > \varepsilon) \leq \frac{2}{1 - e^{-2\varepsilon^2}} e^{-2n\varepsilon^2}. \tag{2.7}$$

By the above considerations, the variable of fundamental importance to the convergence secured by the Glivenko-Cantelli Theorem is

$$M_\varepsilon = \sup\{n : \|F_n - F\| > \varepsilon\}$$

where $\|\cdot\| = \sup_x |\cdot|$ is the uniform norm. Interestingly, for sufficiently large n, the very same bound as eq. (2.7) is valid also for the uniform M_ε.

Indeed, Dvoretzky et al. (1956) proved the fundamental inequality

$$P(\sup_x |F_n(x) - F(x)| > \varepsilon) \leq C e^{-2n\varepsilon^2}$$

for some $C > 0$ independent of n, F and ε. Massart (1990) proves that $C = 2$ is the tight constant, as long as $\exp\{-2n\varepsilon^2\} < 1/2$. This is in fact the same bound as our fundamental inequality (2.1). Assuming n to be sufficiently large compared to ε, sub-additivity immediately shows

$$P(M_\varepsilon > n) = P(\sup_{k \geq n} \|F_n - F| > \varepsilon) \leq \frac{2}{1 - e^{-2\varepsilon^2}} e^{-2n\varepsilon^2}. \tag{2.8}$$

2.2. Uniformity and the Vapnik-Chervonenkis inequalities. The basic Bernoulli Binomial convergence Theorem shows that when $X_1, X_2, \ldots,$ is an *iid* sample, we can for any $\varepsilon, \eta > 0$ find a N so that

$$P\left(\left|\frac{\#\text{Number of } X_i \text{ in } A}{n} - P(X \in A)\right| > \varepsilon\right) < \eta \tag{2.9}$$

for all $n \geq N$. In contrast, the Glivenko-Cantelli Theorem can be read as

$$P\left(\lim_{n\to\infty} \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_i)\right|\right) = 1, \tag{2.10}$$

where

$$\mathcal{F} = \{f(x) = I\{x \leq r\} : r \in \mathbb{R}\}.$$

That is, we have convergence such as eq. (2.9) in a uniform sense over $\mathcal{F}$.

The convergence of eq. (2.9) is fundamentally different from eq. (2.9) in two ways. Firstly, the Bernoulli Theorem only works with finite combinations of variables $(X_1, X_2, \ldots, X_n)$, while the Glivenko-Cantelli Theorem deals with the whole sequence $(X_1, X_2, \ldots, X_n, \ldots)$. Secondly, the Glivenko-Cantelli Theorem does not deal with the convergence of one relative frequency, but the uniform convergence of relative frequencies over some space. In this most basic setting, the convergence is uniform over a set with a continuum cardinality.

The first point means we are here leaving the mathematical structures of the sixteenth century of finite repetitions. Questions when applied to the strong law of large numbers, such as the law of N_ε, are usually framed in the measure theoretic formalization of Kolmogorov. Interestingly, this measure theory formalization meets its limitation concerning questions of uniformity, as one often encounters non-measurable variables. Although we encounter this problem in the current section, we will wait until the next section before focusing on possible solutions to this problem.

From this perspective, it is natural to ask how large $\mathcal{F}$ can be. First of all, we note that it cannot be arbitrarily large while still maintaining convergence such as eq. (2.10). Let $X \sim U[0,1]$ and put

$$\mathcal{F} = \{f(x) = I\{x \in A\} : A \in \mathcal{A}\}$$

where $\mathcal{A}$ is the Borel σ-algebra. For any realization $X_1(\omega) = x_1$, $X_2(\omega) = x_2, \ldots X_n(\omega) = x_n$, the event $A = \{X_1(\omega) = x_1, X_2(\omega) = x_2, \ldots, X_n(\omega) = x_n\}$ is measurable so that $I\{x \in A\} \in \mathcal{F}$. As it is countable, we have $P(A) = 0$, but $\frac{1}{n}\sum_{i=1}^n I\{X_i \in A\} = 1$. Hence,

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i(\omega)) - \mathbb{E}f(X_i(\omega)) \right| = 1.$$

While it should come as no surprise that there are limits to how large $\mathcal{F}$ can be – and the above $\mathcal{F}$ is indeed extremely large – a more subtle problem is the following; still assuming $X \sim U[0,1]$, we now set $\mathcal{F}$ to be the singleton $\{I\{x \in A\}\}$ where A is a non-measurable set with respect to the Borel σ-algebra (implied by the continuum hypothesis). As $\sup_{f\in\mathcal{F}} |f(X_1)|$ is 1 if $X_1 \in A$ and zero otherwise, it is non-measurable. Indeed,

$$\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n f(X_i)|$$

is non-measurable for any n. There are also other settings for which the variable

$$\Gamma_n(\mathcal{F}, P) := \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i) \right|$$

may be non-measurable, see Chapter 5 of Dudley (1999). Hence, in investigating the types of $\mathcal{F}$ which secures the validity of eq. (2.10), we naturally meet mappings from

Ω to $\mathbb{R}$ which are not random variables. To overcome the problem of measurability, we will call function sets $\mathcal{F}$ a Glivenko-Cantelli set if there exists some measurable random variable $\bar{\Gamma}_n(\mathcal{F}, P)$ so that

$$\Gamma_n(\mathcal{F}, P) \leq \bar{\Gamma}_n(\mathcal{F}, P) \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{2.11}$$

Talagrand (1987) showed that if $\mathcal{F}$ is Glivenko-Cantelli, the set

$$\Omega^* = \{\omega \in \Omega : \lim_{n\to\infty} \Gamma_n(\mathcal{F}, P) = 0\}$$

is P-measurable. Hence, eq. (2.10) is indeed valid also when $\Gamma_n(\mathcal{F}, P)$ is non-measurable for finite n. The first general characterization of the Glivenko-Cantelli sets was found in Vapnik & Chervonenkis (1971). They also gave a very applicable sufficient condition for $\mathcal{F}$ to be Glivenko-Cantelli: $\mathcal{F}$ if Glivenko-Cantelli if it has finite so-called VC (Vapnik-Chervonenkis) index. Function spaces $\mathcal{F}$ with this property also follow a uniform extension of the central-limit theorem. This extended central limit theorem will be the basis for our approximations of the limit-law of N_ε in our paper "On the errors committed by sequences of estimator functionals". To later introduce these fundamental approximations without getting too technical, we will spend some time on VC-classes. We follow the exposition of van de Geer (2000).

Definition 1. *Let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$. For random variables $X_1, X_2, \ldots, X_n \in \mathcal{X}$, define the random variable*

$$\Delta^{\mathcal{D}}(X_1, \ldots, X_n) = card\{D \cap \{X_1, \ldots, X_n\} : D \in \mathcal{D}\},$$

the number of different subsets of the form $D \cap \{X_1, \ldots, X_n\}$. Define moreover the number

$$m^{\mathcal{D}}(n) = \sup\{\Delta^{\mathcal{D}}(X_1, \ldots, X_n) : X_1, X_2, \ldots, X_n \in \mathcal{X}\},$$

and

$$V(\mathcal{D}) = \inf\{n \geq 1 : m^{\mathcal{D}}(n) < 2^n\}.$$

We call $V(\mathcal{D})$ the index of the class $\mathcal{D}$, and $\mathcal{D}$ is a Vapnik-Chervonenkis *class if $V(\mathcal{D}) < \infty$.*

Definition 2. *The subgraph of a function $g : \mathcal{X} \mapsto \mathbb{R}$ is*

$$subgraph(f) = \{(x, y) \in \mathcal{X} \times \mathbb{R} : f(x) > y\}.$$

For a class of functions $\mathcal{F}$, let $V(\mathcal{F})$ be the index of the collection of subgraphs $\{subgraph(f) : f \in \mathcal{F}\}$. A collection of functions $\mathcal{F}$ is called a Vapnik-Chervonenkis subgraph class if $V(\mathcal{F}) < \infty$.

The following inequality is proved as Theorem 2.14.9 in van der Vaart & Wellner (1996) in a slightly more general case, and is originally proved in Vapnik & Chervonenkis (1971).

Theorem 1. *Suppose $\mathcal{F}$ has finite VC-index. There then exists a random variable $\bar{\Gamma}_n(\mathcal{F}, P)$ with*

$$\Gamma_n(\mathcal{F}, P) \leq \bar{\Gamma}_n(\mathcal{F}, P)$$

and constants $D, V > 0$ independent of P such that

$$P(\bar{\Gamma}_n(\mathcal{F}, P) > \varepsilon) \leq \left(\frac{D\sqrt{n}\varepsilon}{\sqrt{V}}\right)^V e^{-2n\varepsilon^2}.$$

Given a function space $\mathcal{F}$ with finite VC-index, define

$$\begin{aligned} N_\varepsilon &= \sup\left\{n : \sup_{f\in\mathcal{F}} \|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1)\|_{\mathcal{F}} > \varepsilon\right\} \\ &= \sup\left\{n : \sup_{f\in\mathcal{F}} \|P_n(f) - P(f)\|_{\mathcal{F}} > \varepsilon\right\} \\ &= \sup\left\{n : \|P_n - P\|_{\mathcal{F}} > \varepsilon\right\} \end{aligned}$$

where $\|K\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |K(f)|$ is the uniform norm on $\mathcal{F}$ and

$$P_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i), \qquad Pf = \mathbb{E}f(X_1).$$

Following eq.(2.8), Theorem 1 shows that for some $C, V > 0$, we have

$$\begin{aligned} P(N_\varepsilon > n) &= P(\sup_{k\geq n} \|P_k - P\|_{\mathcal{F}} > \varepsilon) \\ &\leq \sum_{k=n}^\infty P(\|P_k - P\|_{\mathcal{F}} > \varepsilon) \\ &\leq C\varepsilon^V \sum_{k=n}^\infty k^{V/2} e^{-2k\varepsilon^2} \\ &\leq C\varepsilon^V \int_n^\infty x^{V/2} e^{-2x\varepsilon^2}\, \mathrm{d}x \\ &= C\varepsilon^V \frac{\Gamma(V/2+1)}{(2\varepsilon)^{V/2+1}} P(\mathrm{Gamma}(V/2+1, 2\varepsilon^2) > n) \end{aligned}$$

When $V > 2$, the Gamma tail-bound inequality found in section 35.1.3 of DasGupta (2008) gives

$$P(N_\varepsilon > n) \leq C(V/4+1)\varepsilon^{V-2} x^{V/2} e^{-2n\varepsilon^2}, \tag{2.12}$$

and when $0 < V \leq 2$, we have

$$P(N_\varepsilon > n) \leq \frac{2C\varepsilon^V}{1 - e^{-2\varepsilon^2}} e^{-2n\varepsilon^2}. \tag{2.13}$$

Both of these inequalities are uniform in P. Section 6.4 of Dudley (1999) shows that the existence of some random variable $\bar{\Gamma}_n(\mathcal{F}, P)$ so that $\Gamma_n(\mathcal{F}, P) \leq \bar{\Gamma}_n(\mathcal{F}, P) \xrightarrow[n\to\infty]{\mathscr{P}} 0$ uniformly in P implies that $\mathcal{F}$ has finite VC-index. Hence, inequality (2.12) or inequality (2.13) is valid when $\mathcal{F}$ is Glivenko-Cantelli uniformly in P.

As mentioned above, these bounds for the distribution of N_ε are based solely on the subadditivity technique of (2.8), and are therefore rather crude. Although slightly sharper inequalities do exist (see Section 6.5 of Dudley (1999)), any general tail-bound that only uses VC-index of $\mathcal{F}$ has the potential of being improved in concrete situations. For example, the space of half-lines on $\mathbb{R}$ (that is, the classical empirical distribution case) has finite VC-index, but the above tail-inequality is clearly inferior to eq. (2.8).

2.3. CLT-based approximations for the tail of N_ε. Finite sample tail-bounds for N_ε which does not rely on the subadditivity step in eq. (2.8) can be found in special cases. For example, in the simple average case, martingale inequalities yield tail-bounds for N_ε directly, without using subadditivity. See Chapter IV.5 and Chapter VII.3.5 in Shiryaev (1995). However, there does not seem to be any known and generally applicable way to reach sharp tail-bounds for the N_ε variable for more general estimates than the simple average.

Returning to the list of the three basic ways of calculating a probability, the two first seem to be of little use except in special cases. We now investigate the third option.

Suppose we have some estimator $\hat{\theta}_n$ based on n observations, and that

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\text{a.s.}} \theta.$$

The estimator $\hat{\theta}$ may for example be $\frac{1}{n}\sum_{i=1}^n f(X_i)$, which typically estimates $\mathbb{E}f(X_1)$. We wish to approximate the law of

$$N_\varepsilon = \sup\{n \geq 1 : \|\hat{\theta}_n - \theta\| > \varepsilon\}$$

when ε is small. By definition of N_ε, we have the following series of equivalences:

$$\begin{aligned}\varepsilon^2 N_\varepsilon > y \iff \sup\{n \geq 1 : \|\hat{\theta}_n - \theta\| > \varepsilon\} > y/\varepsilon^2 \\ \iff \sup_{n \geq y/\varepsilon^2} \|\hat{\theta}_n - \theta\| > \varepsilon \iff \sup_{s \geq 1} \|\hat{\theta}_{\lfloor s\cdot y/\varepsilon^2 \rfloor} - \theta\| > \varepsilon.\end{aligned}$$

This means that

$$P(\varepsilon^2 N_\varepsilon > y) = P(\sup_{s\geq 1} \|\hat{\theta}_{\lfloor s\cdot y/\varepsilon^2 \rfloor} - \theta\| > \varepsilon).$$

Let us now define $m = \lfloor y/\varepsilon^2 \rfloor$, so that

$$P(\varepsilon^2 N_\varepsilon > y) = P(\sup_{s\geq 1} \|\sqrt{m}\left[\hat{\theta}_{sm} - \theta\right] \| > \sqrt{y_0}),$$

where $y_0 = \varepsilon^2\lfloor y/\varepsilon^2 \rfloor$. This shows that the variable $\varepsilon^2 N_\varepsilon$ is a functional of the stochastic process

$$s \mapsto \mathbb{X}_m(s) = \sqrt{m}\left[\hat{\theta}_{sm} - \theta\right].$$

So if we have process convergence

$$\mathbb{X}_m(s) \xrightarrow[m\to\infty]{\mathscr{W}} \mathbb{X}(s), \qquad s > 0$$

in an appropriate space, for some process $\mathbb{X}(s)$, we get

$$\varepsilon^2 N_\varepsilon \xrightarrow[\varepsilon\to 0^+]{\mathscr{W}} \sup_{s\geq 1} \|\mathbb{X}_s\|^2$$

by the continuous mapping theorem. This means that

$$P(N_\varepsilon > \lambda) = P(\varepsilon^2 N_\varepsilon > \varepsilon^2\lambda) \approx P(\sup_{s\geq 1} \|\mathbb{X}_s\|^2 > \varepsilon^2\lambda) \tag{2.14}$$

for small ε. For this to be useful, we need to describe the limit process $\mathbb{X}$. The paper "On the errors committed by sequences of estimator functionals" shows that for a large class of estimators, approximation in eq. (2.14) is valid and we identify the limit structure and show that it is quite simple.

So far, we have looked at the estimation of the set

$$\{\mathbb{E}f(X) : f \in \mathcal{F}\}$$

through simple averages. This can be seen as the estimation of the function

$$f \mapsto \mathbb{E}f(x). \tag{2.15}$$

"On the errors committed by sequences of estimator functionals" extends this study to estimators of the form

$$\theta_n = \phi(P_n f)$$

where $\phi : l^\infty(\mathcal{F}) \mapsto \mathbb{E}$ for some space $\mathbb{E}$. That is, ϕ takes the function $f \mapsto P_n f$ as an argument and returns a function. This is indeed a generalization of the case of averages, as this case is regained when ϕ is the identity mapping. We work with the assumption that ϕ is functionally differentiable in the Hadamard-sense with a differential denoted by $\dot{\phi}$. The technical definitions are given in the paper.

Under some additional technical constraints, which are s fulfilled if $\mathcal{F}$ has finite VC-index, we have

$$\varepsilon^2 N_\varepsilon \xrightarrow[n\to\infty]{\mathscr{W}} \sup_{0<s\leq 1} \sup_{e\in\mathcal{E}} |\dot{\phi}[\mathbb{Z}_s(f)](e)|^2,$$

as $\varepsilon \to 0$, where $N_\varepsilon = \sup\{n : \|\phi(P_n) - \phi(P)\|_{\mathcal{F}}\}$. Here, $(s, f) \mapsto \mathbb{Z}_s(f)$ is a continuous mean zero Gaussian process on $(0, \infty) \times \mathcal{F}$ with covariance function

$$\mathbb{E}\dot{\phi}\mathbb{Z}_{s_1}(e_1)\dot{\phi}\mathbb{Z}_{s_2}(e_2) = (s_1 \wedge s_2)\mathbb{E}\dot{\phi}W^\circ(e_1)\dot{\phi}W^\circ(e_2),$$

where W° is a P-Brownian bridge process on $\mathcal{F}$.

2.4. Full circle: Calculating the quantiles of the limiting distribution. In Section 2.1, we used the limiting result

$$P\left(\left|\frac{S_n}{n}-p\right|\geq\varepsilon\right)\approx P\left(\left|N\left(0,p(1-p)\right)\right|\geq\sqrt{n}\varepsilon\right)=1-2\Phi\left(-\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right),$$

to get

$$n_{\text{Normal}}=\left\lceil\frac{p(1-p)}{\varepsilon^2}\Phi^{-1}\left(\frac{\delta}{2}\right)^2\right\rceil.$$

For a given ε, and optimizing away p, we readily found the uniform bound $\lceil\varepsilon^{-2}/4\Phi^{-1}(\delta/2)^2\rceil=6767$. In our current problem, we wish to use the approximation

$$P(N_\varepsilon>\lambda)=P(\varepsilon^2N_\varepsilon>\varepsilon^2\lambda)\approx P(\sup_{0<s\leq1}\sup_{e\in\mathcal{E}}|\dot{\phi}[\mathbb{Z}_s(f)](e)|^2>\varepsilon^2\lambda)\tag{2.16}$$

to approximate the law of N_ε when ε is small. The law of N_ε is much more difficult to compute than the law of the limit variable. However, now even the limit variable is subject to a law that is difficult to compute.

Under mild regularity conditions, the limit process is the supremum of a Gaussian process. Although there is a extensive literature on approximating such probabilities, these investigations have mostly found upper bounds of exceedance probabilities given in terms of unspecified constants and are of little use in actual calculations. Simulation is always possible, but for complex functionals $\dot{\phi}$ and large spaces $\mathcal{F}$ this can be difficult. In some special cases of interest, good explicit bounds are known, or the exact distribution can be simulated with ease. One such case is when $e\mapsto\dot{\phi}[\mathbb{Z}_s(f)](e)$ is a Gaussian Martingale on $[0,\tau)$, as is the case for the Nelson–Aalen estimator. Then the limit variable of $\varepsilon^2N_\varepsilon$ is

$$\sigma^2\sup_{(s,t)\in[0,1]^2}\|\mathbb{S}(s,t)\|^2$$

where $\mathcal{S}$ is a Brownian Sheet on $[0,1]^2$ and where

$$\sigma^2=\inf\left\{s:\left\langle\dot{\phi}W^\circ,\dot{\phi}W^\circ\right\rangle_s>\tau\right\}$$

where $\left\langle\dot{\phi}W^\circ,\dot{\phi}W^\circ\right\rangle_s$ is the covariation process of $\dot{\phi}W^\circ$. This distribution can easily be found by simulation, and fairly good tail-bounds are known.

For the general case, however, we are returned to the list of possible ways of calculating probabilities. The left hand side of eq. (2.16) is clearly very much more difficult to calculate than the right hand side. However, even with such a reduction, this problem may still be difficult. Fatalov (2003) is a comprehensive survey of bounds for norms of Gaussian processes where the involved constants are specified. Only a very few seem to be useful for our current investigation.

2.5. A new type of sequential confidence bands for the Nelson–Aalen estimator. Let us look at a further application of the approximation given in eq. (2.16). Besides its theoretical interest, the limit law of $\varepsilon^2 N_\varepsilon$ can be used to derive approximate sequential confidence sets. Indeed, calculate or approximate the upper α quantile of the limit variable of $\varepsilon^2 N_\varepsilon$ and denote this quantile by λ_α. Fix the radius of the confidence set as ε_0 and compute

$$m = [\lambda_\alpha/\varepsilon_0^2]. \tag{2.17}$$

By the distributional convergence, we get that

$$\begin{aligned} P(\varepsilon^2 N_\varepsilon < \lambda_\alpha) &= P(\|\phi(P_n) - \phi(P)\|_{\mathcal{E}} \le \varepsilon_0 \text{ for all } n \ge m) \\ &= P(\phi(P) \in B\left(\varepsilon_0, \phi(P_n)\right) \text{ for all } n \ge m) \end{aligned}$$

is close to $1-\alpha$ where

$$B(\varepsilon, y) = \{x : \|x - y\|_{\mathcal{E}} \le \varepsilon\}$$

is an ε-ball in $l^\infty(\mathcal{E})$. This has intuitive appeal. Whereas confidence sets are usually of the form

$$P(\phi(P) \in C_n) \ge 1 - \alpha, \qquad \text{for all } n \ge m$$

and thus only give a probability statement for one $n \ge m$ at the time, a fixed-volume confidence set gives a simultaneous answer for all $n \ge m$.

Let us illustrate this for the Nelson–Aalen estimator. Suppose that we observe $X_i = (Z_i, \Delta_i) \sim F$, in which $Z_i = Y_i \wedge C_i$ and $\Delta_i = 1\{Y_i \le C_i\}$ are defined in terms of unobservable *iid* failure times $Y_i < \tau$. Here Y_i are distributed according to G and we will assume that the censoring times C_i are iid. The Nelson–Aalen estimator

$$\Lambda_n(t) = \int_{[0,t]} \frac{1}{\bar{\mathbb{H}}_n} \mathrm{d}\mathbb{H}_n^{uc},$$

where

$$\mathbb{H}_n^{uc}(t) = \frac{1}{n} \sum_{i=1}^n \Delta_i 1\{Z_i \le t\}$$

and

$$\bar{\mathbb{H}}_n(t) = \frac{1}{n} \sum_{i=1}^n 1\{Z_i \ge t\}$$

converges uniformly

$$\Lambda(t) := \int_{[0,t]} \frac{1}{1 - G(t)} \, \mathrm{d}G$$

almost surely under quite general conditions (Shorack & Wellner, 1986, see). That is,

$$P\left(\lim_{n\to\infty} \sup_{0<t<\tau} |\Lambda_n(t) - \Lambda(t)| = 0\right) = 1.$$

We are interested in finding the limit of

$$N_\varepsilon = \sup\{n \ge 1 : \sup_{0<t<\tau} |\Lambda_n(t) - \Lambda(t)| > \varepsilon\}.$$

This estimator fits into the framework of our paper. As is well-known, $\sqrt{n}(\Lambda_n - \Lambda)$ converges to a Gaussian Martingale. This means the limit of $\varepsilon^2 N_\varepsilon$ reduces to the study of the supremum of a Brownian Sheet. Let

$$\sigma^2 = \int_{[0,\tau]} \frac{1 - \Delta\Lambda(z)}{P\{Z \geq z\}} d\Lambda(z)$$

and suppose $\mathbb{S}$ is a Brownian Sheet on $[0,1]^2$. Then we get

$$\varepsilon^2 N_\varepsilon \xrightarrow[\varepsilon \to 0^+]{\mathscr{W}} \sigma^2 \left(\sup_{0 \leq s \leq 1} \sup_{0 \leq t \leq 1} |\mathbb{S}(s,t)| \right)^2 .$$

Hence, the m of eq. (2.17) can be calculated to arbitrary precision for any given σ^2. We also give an upper bound for m in our paper.

Let us also note that the exact distribution of the supremum of a Brownian Sheet do not seem to be known. The best known bound for its distribution seems to be Talagrand (1994), which gives bounds in terms of unspecified constants. Csáki et al. (2000) is almost useful, but works with $\sup \mathbb{S}$ and not the required $\sup |\mathbb{S}|$, and their results does not seem to be transferable to our case. Goodman (1976) provides good general lower bounds, but his upper bound – which is what we need to bound m – is worse than the one used in our paper.

3. Gorgias' revenge: Model selection and pragmatism

A statistical model is the specification of some general patterns of summaries of basic events $\omega \in \Omega$. The summaries of these events are given by a probability measure P. This measure is often unknown to the modeller, but is supposed known to be in a set of probability measures

$$\{P_\theta : \theta \in \Theta\}$$

That is, there exists some $\theta_0 \in \Theta$ such that $P = P_{\theta_0}$. Based on observations whose distribution is P, a fundamental problem of statistics is to regain θ_0. We will denote a generic estimator of θ_0 by $\hat{\theta}$. A good estimator is near θ_0 with high probability.

The most classical situation is the observation of a series of random variables $X_1, \ldots, X_n$ in some space such as $\mathbb{R}^d$. Let us denote the empirical estimator for θ_0 based on these observations by $\hat{\theta}_n$. Then, typical good estimators are consistent in the sense that

$$\hat{\theta} \xrightarrow[n \to \infty]{\mathscr{P}} \theta_0. \tag{3.1}$$

The parameter set Θ is usually a Euclidian space with dimensionality much lower than n.

Model building under uncertainty is in many ways one of the fundamental features of rational existence. Rationality is the ability to reason with abstractions

based on sense perception. The process of building and working with these abstractions can be understood as non-formalized model building. The belief in platonic forms is a belief in these abstractions' power to reach the actual objects of the world – Kant's "things in themselves". The view of most philosophers after Kant is that this is impossible. These general considerations can be translated into the statistical dictum that all models are wrong, but some are useful.

If the validity of the statistical model is uncertain, as it always is, the theory of model selection solves this problem by cutting the Gordian knot: Extend the size of Θ, while limiting the dimensionality of the parameter set. This is achieved by splitting Θ into smaller candidate models $\Theta_i \subset \Theta, i \in I$ where the cardinality of I is often small and the dimensionality of each Θ_i is moderate to small compared to n. Empirical estimates of θ_0, say $\hat{\theta}$ are then constrained to belong to some Θ_i.

While this problem is similar to basic statistical estimation described at the start of this section, the major difference is that although

$$\bigcup_{i \in I} \Theta_i \subset \Theta,$$

we also have

$$\bigcup_{i \in I} \Theta_i \neq \Theta.$$

Indeed, Θ is often a very large space compared to $\bigcup_{i \in I} \Theta_i$. The process of constraining estimates to be in $\bigcup_{i \in I} \Theta_i$ is an important distinction and has major practical and theoretical consequences. The problem is that even though $P \in \{P_\theta : \theta \in \Theta\}$, so that $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$, we may have

$$\theta_0 \notin \bigcup_{i \in I} \Theta_i.$$

So, any empirical estimator $\hat{\theta}_n$, based on n observations, can never fulfill the basic consistency demand of eq. (3.1). In most situations, we rather have

$$\hat{\theta}_n \xrightarrow[n \to \infty]{\mathscr{P}} \theta^\circ \tag{3.2}$$

where

$$\theta^\circ = \underset{\theta \in \bigcup_{i \in I} \Theta_i}{\operatorname{argmin}} \mathcal{D}(P, \theta) \tag{3.3}$$

is the least false parameter configuration with respect to some criteria function $\mathcal{D}$. If θ° and θ_0 are sufficiently close, the resulting model is hoped to provide good approximations to the real stochastic behavior of the observed system.

Let us here note that Θ cannot be chosen arbitrarily large. When Θ parametrizes the space of all probability distributions that the observed data-points can have, it is fundamentally impossible to estimate the parameters with respect to any non-trivial criteria. We will expand on the relation between $\bigcup \Theta_i$ and Θ in Section 4.

There is a two-fold irony surrounding the model selection problem in statistical modelling. Firstly, it is ironic that mathematics, which is can be seen as Plato's strongest illustration of man's wisdom into the hidden layers of the universe, is now used to model uncertainty. Secondly, it is ironic that the necessity of model selection and its typical non-consistency confirms Gorgias pragmatism and critique of certain knowledge, and hence can be read as an answer to Plato's ridicule of the Sophists – through applied mathematics[2].

The above paragraph describes the interpretation of the model selection. However, the mathematical distinction between the statistical estimation of the parameter θ° and the model selection problem is also subtle. Once the parameter sets $\Theta_i, i \in I$ is fixed, the least false parameter of eq. (3.3) is just a number defined in terms of the true model – known to be in Θ. There are many general strategies for this set-up, such as those surveyed in Bickel et al. (1993). Using this theory, one could perceive the model set as Θ and estimate the actual true parameter θ_0 based on this model, resulting in an empirical estimator $\tilde{\theta}_n$. Then, θ° could be estimated as a plug-in estimator

$$\underset{\theta \in \bigcup_{i \in I} \Theta_i}{\operatorname{argmin}} \mathcal{D}(P, \tilde{\theta}_n). \tag{3.4}$$

However, this would mean we would have to estimate the very complex parameter θ_0 to estimate the much simpler parameter θ°. In this process, we would break what Vapnik (1995, page 30) calls "The main principle for solving problems using a restricted amount of information", given by

> When solving a given problem, try to avoid solving a more general problem as an intermediate step.

Also, estimating θ_0 may only be possible to achieve with very slowly converging estimators, and this would make the plug-in estimator of eq. (3.4) a very poor estimator.

Another take on the problem would be to estimate θ° as a minimum distance estimator parametrized by the set $\bigcup_{i \in I} \Theta_i$. That is, one could study estimators of

[2]However, it should be noted that Plato's Socrates clearly was well aware of the limitations and uncertainties of logical deductions. The nature of the Platonic dialogues limits any easy interpretation on what Plato "meant" and his texts are far too interesting for such a simple interpretation. The limitations and uncertain applicability of the seemingly primitive deduction-rules used in several the dialogues seems to function as an illustration of such limitations. An example of this type of awareness is found in the Phaedo, where Socrates is given the lines " For if what I say is true, then I do well to be persuaded of the truth, but if there be nothing after death, still, during the short time that remains, I shall not distress my friends with lamentations, and my ignorance will not last, but will die with me, and therefore no harm will be done.". However, when the forms are discussed, their presentation seems more religious than the critical effort Plato puts into shaping logic.

the form

$$\tilde{\theta}_n = \underset{\theta \in \bigcup_{i \in I} \Theta_i}{\operatorname{argmin}} \mathcal{D}_n(\theta) \tag{3.5}$$

where $\mathcal{D}_n$ is calculated on the basis of observed data. However, $\bigcup_{i \in I} \Theta_i$ is topologically not connected, and does not adhere to any of the typical regularity conditions associated with parametric statistics. Even so, there exist results capable of giving non-asymptotic error bounds for such estimators in Spokoiny (2009) and Golubev & Spokoiny (2009). These results can be used to derive both consistency and rates of convergence for minimal distance estimators under highly non-standard parametrizations such as $\bigcup_{i \in I} \Theta_i$. However, their probability bounds are defined in terms of the true, unknown measure P_{θ_0} and therefore cannot be directly used to give uncertainty estimates of the resulting estimator. While the direct estimation of θ° as a minimum distance estimator parametrized by $\bigcup_{i \in I} \Theta_i$ seems to be the ideal perspective to work from, I do not know of any applied work that has used such a program. Generally applicable uncertainty estimates are mostly unknown and seem to be extremely difficult to find under general assumptions. This point is taken somewhat further at the end of Section 3.2.

3.1. Two-stage model selection procedures. By far the most common model selection technique is to split the estimation of θ°into two stages. First, an index $\hat{i}_n \in I$ is chosen by a model selection formula such as the AIC, that tries to reach i° – the index for which

$$\theta^\circ \in \Theta_{i^\circ}$$

is achieved, under the assumption that this i is unique. Note the difference between this technique and the plug-in estimator of eq. (3.4): the estimation of i° is constrained to the discrete set I with low cardinality.

After the calculation of $\hat{i}$, one typically discards the probabilistic consequences of model selection, and estimate the least false estimator

$$\tilde{\theta}^\circ(i) = \operatorname{argmin} \mathcal{D}(P, \theta),$$

where the argmin is over $\theta \in \Theta(\hat{i})$ – as if we know that $\hat{i}$ is the index actually known to contain the least false parameter configuration.

As mentioned in the first section of the introduction, the paper "The Copula Information Criterion and its implications for the Maximum Pseudo Likelihood Estimator" introduces the more technical paper "The Copula Information Criteria" quite thoroughly. We will therefore not spend much time on the technical setting for the copula information criteria, but will rather directly motivate the CIC as an extension of the AIC formula. In order to do this, let us briefly motivate the AIC formula.

The MLE of a parametric model Θ based on n iid observations with cumulative distribution function F° and density f° is

$$\hat{\theta}_n(i) := \operatorname*{argmax}_{\theta\in\Theta_i} \int \log f_\theta(x)\,\mathrm{d}F_n(x), \tag{3.6}$$

where f_θ is the density of P_θ with respect to Lebesgue measure and F_n is the empirical distribution function. Under mild regularity assumptions on the parametrization, we have

$$\begin{aligned}\hat{\theta}_n(i) \xrightarrow[n\to\infty]{\mathscr{P}} \theta^\circ &:= \operatorname*{argmax}_{\theta\in\Theta} \int \log f_\theta(x)\,\mathrm{d}F^\circ(x)\\ &= \operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}(f^\circ, f_\theta).\end{aligned}$$

Here,

$$\mathrm{KL}(f^\circ, f_\theta) = \int \log \frac{f_\theta}{f^\circ}\,\mathrm{d}F^\circ$$

is the Kullback–Leibler divergence between f° and f_θ. It is zero if and only if $f^\circ = f_\theta$ almost surely. While Kullback–Leibler divergence is not a metric, it dominates the Hellinger metric, defined by

$$h(f^\circ, f_\theta) = \left(\frac{1}{2}\int (\sqrt{f^\circ} - \sqrt{f_\theta})^2\,\mu\right)^2.$$

In fact,

$$h(f^\circ, f_\theta) \le \frac{1}{2}\mathrm{KL}(f^\circ, f_\theta).$$

A simple proof is given in Lemma 1.3 of van de Geer (2000). There are also other motivations for using Kullback–Leibler divergence, see the general treatment of Claeskens & Hjort (2008) and the next subsection.

As elaborated in "The Copula Information Criterion and its implications for the Maximum Pseudo Likelihood Estimator", the AIC formula tries to estimate

$$\operatorname*{argmin}_{\theta\in\bigcup_{i\in I}\Theta_i} \mathrm{KL}(f^\circ, f_\theta)$$

for model sets $\Theta_i, i \in I$ through first forming an estimator $\hat{i}$, estimating the index $i^\circ \in I$ that the argmax in the above display achieves, and then estimate the parameter configuration in $\Theta_{\hat{i}}$ through Maximum Likelihood.

The estimation of i° is done through an estimator of the form

$$\hat{i} = \operatorname*{argmin}_{i\in\mathcal{I}} \left\{\int \log f_{\hat{\theta}_n(i)}(x)\,\mathrm{d}F_n(x) - p_n\right\}.$$

Here p_n is a first order asymptotic approximation of the expectation of

$$\int \log f_{\hat{\theta}_n(i)}(x)\,\mathrm{d}[F_n - F^\circ](x),$$

so that

$$\mathbb{E}\int \log f_{\hat\theta_n(i)}(x)\,\mathrm{d}F_n(x) - p_n \approx 0 \tag{3.7}$$

when the model Θ_i is assumed to contain the true data-generating parameter. From this perspective, the AIC is a natural generalization of the interpretation of the MLE as a minimizer of KL-divergence: As the parameter estimation is estimated through the MLE, the AIC methodology tries to reach the parameter configuration in $\bigcup_i \Theta_i$ attaining the least KL-divergence from the true density f_{θ_0} to the set $\{f_\theta : \theta \in \bigcup_i \Theta_i\}$.

However, eq. (3.7) is admittedly a somewhat weak motivation for using p_n. But as we all know, we end up with the extremely simple formula $p_n = n\text{length}(\theta)$. A slightly more motivated version of the problem is to require eq. (3.7) to hold also when the models under consideration are wrong. This generalization leads to the so-called TIC-formula, with a p_n that depends on the data.

The TIC-formula is also motivated through its first order equivalence with a certain version of cross-validation. Indeed, we have that for the TIC-choice of p_n (more involved than the AIC choice), we have

$$\mathrm{TIC}_n = 2n\widehat{\mathrm{xv}}_n + o_P(1), \tag{3.8}$$

where

$$\mathrm{TIC}_n = 2n\left(\int \log f_{\hat\theta_n(i)}(x)\,\mathrm{d}F_n(x) - p_n\right)$$

and we work with the cross-validation sum

$$\widehat{\mathrm{xv}}_n = n^{-1}\sum_{k=1}^{n} \log f(X_k, \hat\theta_{(k)})$$

in which $\hat\theta_{(i)}$ is the ML estimate

$$\hat\theta_{(i)} = \operatorname*{argmax}_\theta \sum_{j\neq i} \log f(X_j, \theta)$$

based on the sample without the i'th observation. While this is a stronger motivation than eq. (3.7), there does not seem to be any applicable finite sample bounds on the $o_p(1)$-term available.

In the CIC paper, we work not with the maximum likelihood estimator, but the so-called maximum pseudo-likelihood estimator, given by

$$\hat\theta_n(i) = \operatorname*{argmax}_{\theta\in\Theta_i} \int_{u\in[0,1]^d} \log c_\theta(u)\,\mathrm{d}\hat C_n(u)$$

where $\hat C_n$ is the empirical copula, given by

$$\hat C_n(u) := \frac{1}{n}\sum_{i=1}^{n} I\{F_{n,\perp}(X_i) \le u\} = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} I\{F_{n,j}(X_{i,j}) \le u_j\}.$$

In the above display, $F_{n,j}$ is the j'th marginal empirical distribution function.

In our paper, we find a p_n that solves

$$\mathbb{E}\int \log f_{\hat{\theta}_{n(i)}}(x)\,\mathrm{d}\hat{C}_n(x) - p_n \approx 0,$$

and call it the CIC. However, we find that such a p_n is rarely finite. Because of this infinitude, we find a p_n that solves an analogue of eq. (3.8). We call this formula the xv-CIC. It is of general applicability, and is simple and fast to calculate. Unfortunately, we discovered the xv-CIC after "The Copula Information Criterion and its implications for the Maximum Pseudo Likelihood Estimator" was published. While I would have rewritten this papers concluding remarks if I had found the xv-CIC before, the xv-CIC formula does have a completely different motivation than the CIC formula. If this scope is preserved, the concluding remarks continue to hold.

3.2. The way ahead: Non-asymptotic model-selection. The CIC and xv-CIC solve the problems defined in eq. (3.7) and eq. (3.8), and are hence susceptible to any critiques one may present against the AIC and TIC. As described in "The Copula Information Criteria", our investigation was not initiated in search for the most optimal model selection criteria, but to investigate the consequences of using the simple "AIC formula" when using the MPLE. Although the AIC formula is by far the most used model selection formula, its use has two main problems.

Firstly, it is often hard to take the uncertainty introduced in the model selection properly into account when giving confidence intervals for the resulting parameter estimates. This is difficult for the same reasons it is difficult to approximate the distribution of eq. (3.5). However, Claeskens & Hjort (2003) and Hjort & Claeskens (2003) works out the effects of model selection under local misspecification assumptions. However, it currently seems out of reach to do this without local misspecification assumptions.

Secondly, the AIC formula sets out to solve an asymptotic problem from a very specific point of view. While it is not at all sure that eq. (3.7) is fulfilled for a given n, a more serious problem seems to be that eq. (3.7) does not give a correction term p_n directly connected to argmax-based the estimator

$$\hat{\theta}_n(\hat{i}),$$

but with the criteria function

$$\int \log f_{\hat{\theta}_{n(i)}}(x)\,\mathrm{d}[F_n - F^\circ](x).$$

A more optimal – and clearly more challenging – perspective would be to solve the problem of finding a p_n that makes

$$\hat{\theta}_n(\hat{i}) = \operatorname*{argmin}_{\theta\in\Theta(\hat{i})} \mathcal{D}_n(\theta)$$

as good an estimator as possible – according to some specified criteria. Here p_n enters through the definition of $\hat{i}$. That is,

$$\hat{i} = \underset{i \in \mathcal{I}}{\operatorname{argmin}} \left\{ \mathcal{D}_n(\theta) - p_n \right\}.$$

This non-asymptotic problem can indeed be solved from several perspectives and in several settings. We mention the book-length treatments of Massart (2007) and Tsybakov (2009), and the paper Akakpo & Durot (2010) that provides a general framework for working with censored data and concentrates on an example concerning histogram selection. Selecting the optimal histogram for observed data is a seemingly simple problem even in the presence of censoring, but the solution to the above problem is mathematically very complex. Optimality of the p_n term is often given through so-called Oracle inequalities, such as

$$\mathbb{E}\|\hat{\theta}_n(\hat{i}) - \theta^\circ\|^2 \leq C \inf_{i \in \mathcal{I}} \left\{ \mathbb{E}\|\hat{\theta}_n(i) - \theta^\circ\|^2 \right\} + r_n$$

where r_n is showed to be small. However, even in the histogram case, the constants involved are difficult to calculate, usually because such inequalities are constructed through concentration inequalities based on the chaining technique (Talagrand, 2005).

Finally, let us note that Oracle inequalities such as above, do not show that direct estimators such as that of eq. (3.5) are inferior to selecting a wisely chosen p_n-term. Indeed, one can rather work directly through finding an optimal $\mathcal{D}_n$-function relative to $\bigcup \Theta_i$ and P. This most direct method is the most optimal way of doing "model selection" – by returning the problem to perhaps the most classical problem of statistical estimation: finding the optimal estimator to a given problem. The claimed optimality is obvious, as the specification of $\mathcal{D}_n$ includes the problem of specifying a p_n. The choice of $\mathcal{D}_n$ is the problem posed and solved by Fisher (1922), under different assumptions – where the main difference is a different topology for the parameter set.

3.3. A connection between the AIC and N_ε. It may seem surprising, but there is a deep connection between the N_ε variable and the AIC methodology. While this connection is probably already known, I have not seen it in the literature. The connection provides what I consider a very good motivation for using AIC-like model selection procedures and also provides a surprising connection between two of the thesis' papers.

Suppose $X_1, X_2, \ldots, X_n, \ldots$ are independent with distribution P_{θ_0}. Many test-statistics for $H_0 : \theta \in \Theta_0 \subset \Theta$ rejects H_0 if

$$T_n = T_n(X_1, X_2, \ldots, X_n)$$

is above a certain limit. Let

$$p_n(t, \theta) = P_{\theta_0}(T_n \geq t)$$

be the exceedance probabilities of T_n under H_0. Then,

$$L_n = L_N(X_1, X_2, \ldots, X_n) = p_n(T_n)$$

is the p-value actually obtained. Now introduce the familiar variable

$$N_\varepsilon = \sup\{n \geq 1 : L_n > \varepsilon\},$$

which may be infinite as we have not assumed $L_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$. Here, N_ε is the minimum sample size required for the T_n-test to become and stay significant. The following Theorem is due to Bahadur (1971), and is given in Chapter 24 of Shorack & Wellner (1986).

Theorem 2. *Suppose*

$$\frac{1}{n} \log L_n \xrightarrow[n\to\infty]{\text{a.s.}} -\frac{1}{2} c(\theta_0)$$

with $0 < c(\theta_0) < \infty$. *Then,*

$$P\left(\lim_{\varepsilon \to 0} \frac{N_\varepsilon}{2 \log(1/\varepsilon)} = \frac{1}{c(\theta_0)} \right) = 1.$$

The function $c(\theta_0)$ is called the Bahadur-slope. Note that as $[2 \log(1/\varepsilon)]^{-1}$ goes much faster to zero than ε^2 when $\varepsilon \to 0^+$, the above limit theorem can be seen as a Law of Large Numbers for the N_ε variable, while the weak convergence of $\varepsilon^2 N_\varepsilon$ is a Central Limit Theorem-like result.

The following Theorem uses the N_ε variable to motivate the desire to work with the probability model which minimize the Kullback–Leibler divergence to the true model. It also establishes the optimality of the likelihood ratio test statistic in certain settings. It appears as Theorem 22.5 in DasGupta (2008), where it is called the Stein-Bahadur-Brown Theorem. For extensions where Θ is of continuum cardinality, see Arcones (2005).

Theorem 3. *Suppose* $X_1, \ldots, X_n$ *are independent with distribution* P_θ, *where* $\theta \in \Theta$. *Assume* Θ *is finite, and consider testing* $H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta \setminus \Theta_0$. *Then the following are true.*

(1) *For any sequence of test statistics* T_n, *the Bahadur slope* $c_T(\theta)$ *satisfies*

$$c_T(\theta_0) \leq 2 \inf_{\theta \in \Theta_0} KL(P_{\theta_0}, P_\theta).$$

(2) *The likelihood ratio test statistic* Λ_n *satisfies* $c_\Lambda(\theta) = 2 \inf_{\theta \in \Theta_0} KL(P_{\theta_0}, P_\theta)$.

That is, for any test statistic T_n conforming to the regularity condition of Theorem 2 and 3, we have

$$\frac{N_\varepsilon}{2\log(1/\varepsilon)} \xrightarrow[\varepsilon\to 0^+]{\text{a.s.}} \frac{1}{c(\theta_0)} \geq \frac{1}{\inf_{\theta\in\Theta_0} \mathrm{KL}(P_{\theta_0}, P_\theta)},$$

and for the best possible test statistic, we have

$$\frac{N_\varepsilon}{2\log(1/\varepsilon)} \xrightarrow[\varepsilon\to 0^+]{\text{a.s.}} \frac{1}{\inf_{\theta\in\Theta_0} \mathrm{KL}(P_{\theta_0}, P_\theta)}.$$

This means that the larger the Kullback–Leibler divergence between two models, the easier it is to distinguish between them through hypothesis testing. Conversely, if the Kullback–Leibler divergence between two densities is small, it is very difficult to distinguish between them using any test what so ever. As the test statistics are arbitrary, this can be interpreted such that any *testable feature* of the two models are similar. This line of thought is analogous to the discussion regarding pseudo-random numbers in Brands & Gill (1995, 1996)

4. Non-standard alternative models: Regression with jumps

When we introduced the estimator defined in eq. (3.6), we worked with iid random vectors $X_1, X_2, \ldots, X_n$, and fitted their common distribution function. The "model selection" step took into consideration that the classes of common distribution functions could be misspecified, by embedding the finite dimensional model originating from the assumption

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) \in \mathcal{M}_{1,p} := \left\{ \prod_{i=1}^n P_{\theta_i}(X_i \in A_i) : (\theta_1, \ldots, \theta_p) \in \Theta \subseteq \mathbb{R}^{\sum \text{length}(\theta_i)} \right\}$$

into the infinite dimensional model originating from the assumption

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) \in \mathcal{M}_{1,\infty} := \left\{ \prod_{i=1}^n P_\psi(X_i \in A_i) : \psi \in \Psi \subseteq \mathbb{R}^\infty \right\},$$

where $\psi \mapsto P_\psi$ parametrizes all probability measures on $\mathbb{R}^d$ with respect to a given σ-algebra. This is the set-up typically described as "non-parametric", but it is clearly incapable of supporting time-series models et cetera.

Let us here mention that while one of these models has finite dimensionality, the other has infinite dimensionality. However, this is not really what distinguishes them. The dimensionality of their parametrization is a highly algebraic concept, and its infinitude is not very illuminating. While the dimensionality of Θ makes a surprise visit in the AIC formula, a more useful measure of size may be the behaviour of variables such as N_ε described in Section 2. This is connected to various efforts to measure the size of parameter sets through results such as Theorem 1, and these measures usually have names connected to the somewhat vague concept of entropy.

Now suppose that we have indices $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots < \tau_k = n$ so that $X_{\tau_0+1}, \ldots, X_{\tau_1}$ are iid, $X_{\tau_1+1}, \ldots, X_{\tau_2}$ are iid, et cetera until $X_{\tau_{k-1}}, \ldots, X_{\tau_k}$ which are also iid. Suppose our model is

$$\mathcal{M}_{i,p_i} = \Bigg\{ \prod_{0 \le j \le k-1} \prod_{\tau_{j-1} < i \le \tau_j} P_{\theta_{i,j}}(X_i \in A_i) : 0 < \tau_1 < \tau_2 < \cdots < \tau_{k-1} < n, \\ \theta \in \Theta \subseteq \mathbb{R}^{\sum_{i,j} \text{length}(p_{i,j})} \Bigg\},$$

parametrized *both* by the τ_i's and θ. However, how large is k? By analogy with the relation between $\mathcal{M}_1$ and $\mathcal{M}_2$, one could imagine that one develops a model selection strategy where

$$\mathcal{N}_p = \bigcup_{i=1}^{n} \mathcal{M}_{i,p}$$

is studied in relation to

$$\mathcal{N}_\infty = \left\{ P_\psi(X_i \in A_i) : \psi_i \in \Psi \right\},$$

where $\psi \mapsto P_\psi$ parametrizes all probability measures on $\mathbb{R}^d$ with respect to a given σ-algebra.

However, this is impossible – as we only have a single observation of each random variable – and any useful concept of the entropy of $\mathcal{N}_\infty$ must be infinite. It is impossible to estimate any model inside $\mathcal{N}_\infty$ as the model is simply too large and does not make any assumptions other than the randomness of the observations.

This can be interpreted as lacking *information* in the model set $\mathcal{N}_\infty$. We often justify the iid assumption through experimental design[3]. However, with structural changes such as those above, or more complex dependence structures found in time series, the modelling assumptions are almost always subjective and based on experience. Let us also mention in this regard that the contrast between parametric and non-parametric statistics is often artificially drawn between studying finite-dimensional or infinite-dimensional subsets of $\mathcal{M}_{1,\infty}$. When dealing with non-iid observations, the relationship between what is the model, and what is the "super-model" considered to be true is much more difficult – as exemplified by the impossibility of dealing with $\mathcal{N}_\infty$.

In our paper "Estimation and inference for jump regression models", we work with a mathematical structure similar to fitting models such as $\mathcal{M}_{d,p}$, with d/n small. In this paper, we do not constrain ourselves to the above multiple change-point set-up, but work in regression-type models of the form

$$y_i = m(x_i, \theta) + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n,$$

[3]However, there are many well-known examples of serial correlation in designed experiments (such as machines behaving differently through time etc). See Box et al. (2005).

where m is a step function with steps specified by the covariates $x_1, x_2, \ldots, x_n$ and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is mean zero Gaussian noise. This set-up extends to more involved m functions, and more involved relationships between the observations and the covariates such as a GLM-like framework. However, let us limit this discussion to the simpler set-up with $\mathcal{M}_{d,p}$. An important technical point is that the asymptotical approximations we use are based on letting $\tau_1, \ldots, \tau_{d-1}$ depend on n and all grow to infinity slower than n, as $n \to \infty$. This work was originally part of a larger project, leading to a model selection formula for studying

$$\bigcup_{i=1}^{D} \mathcal{M}_{i,p}$$

in relation to

$$\mathcal{N}_D = \Big\{ \prod_{0 \le j \le D-1} \prod_{\tau_{j-1} < i \le \tau_j} P_{\psi_i}(X_i \in A_i) : 0 < \tau_1 < \tau_2 < \cdots < \tau_{k-1} < n, \\ \psi_i \in \Psi_i \text{ for each } i \Big\},$$

where D is significantly smaller than n and specified a-priori. However, we realized that the preparatory material needed in such a model-selection paper would contain enough material for a separate paper. The result was "Estimation and inference for jump regression models".

One of the interesting features of such models is that the topology of the parameter space is drastically different from those topologies typically leading to asymptotic $\sqrt{n}$-normality of ML estimators. In fact, the ML estimators $\hat{\tau}_i$ for (the analogue of) τ_i are such that $n(\hat{\tau}_i - \tau_i)$ have a non-trivial limit distribution, while the remaining parameteres are $\sqrt{n}$-normal. Our underlying approximations work with the likelihood function as a stochastic process, following the techniques described in for example Ibragimov & Khasminskii (1981) and van der Vaart & Wellner (1996). Through these techniques, we also derive the frequentist asymptotic behavior of the Bayesian estimators for the parameters defining m and show that the Bayesian estimator is in fact superior to the ML estimator in terms of asymptotic mean squared error.

The theory surrounding these types of models is treated extensively in the litterature, but we seem to be the first to study how these estimators behave when the model is misspecified. In the current paper, we work with the following simple misspecification. Assume that m is not a step function with d jumps, but a continuous function plus a step-function with d jumps. We only work out the details for $d = 1$, but similar – though more complex – developments lead to model selection formulas to select the number of jumps in an AIC-like manner.

Bibliography

Akakpo, N. & Durot, C. (2010). Histogram selection for possibly censored data. *Mathematical Methods of Statistics* **19**, 189–218.

Arcones, M. (2005). Bahadur efficiency of the likelihood ratio test. *Mathematical Methods of Statistics* **14**, 163.

Bahadur, R. (1971). *Some limit theorems in statistics*, vol. 4 of *Regional Conference Series on Applied Mathematics*. SIAM, Philadelphia, Pennsylvania.

Bickel, P., Klaassen, A., Ritov, Y. & Wellner, J. (1993). *Efficient and adaptive inference in semi-parametric models*. Johns Hopkins University Press, Baltimore.

Box, G., Hunter, J. & Hunter, W. (2005). *Statistics for experimenters: design, innovation, and discovery*. Wiley-Interscience New York, 2nd ed.

Brands, S. & Gill, R. (1995). Cryptography, statistics and pseudo-randomness (part I). *Probability and Mathematical Statistics* **15**, 101–114.

Brands, S. & Gill, R. (1996). Cryptography, statistics and pseudo-randomness (Part 2). *Probability and Mathematical Statistics* **16**, 1–17.

Claeskens, G. & Hjort, N. (2003). The focused information criterion. *Journal of the American Statistical Association* **98**.

Claeskens, G. & Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Csáki, E., Khoshnevisan, D. & Shi, Z. (2000). Boundary crossings and the distribution function of the maximum of Brownian sheet. *Stochastic processes and their applications* **90**, 1–18.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer Verlag.

Dudley, R. (1999). *Uniform central limit theorems*. Cambridge university press.

Dudley, R. (2003). *Real Analysis and Probability*. Cambridge studies in advanced mathematics. Cambridge, 2nd ed.

Dvoretzky, A., Kiefer, J. & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial

estimator. *The Annals of Mathematical Statistics* **27**, 642–669.

Fatalov, V. (2003). Constants in the asymptotics of small deviation probabilities for Gaussian processes and fields. *Russian Mathematical Surveys* **58**, 725.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* , 309–368.

Foucault, M. (1966). *The order of things: An archaeology of the human sciences*. Editions Gallimard.

Gagarin, M. (1994). Probability and persuasion: Plato and early Greek rhetoric. *Persuasion: Greek rhetoric in action* , 46–68.

Garber, D. & Zabell, S. (1979). On the emergence of probability. *Archive for History of Exact Sciences* **21**, 33–53.

Golubev, Y. & Spokoiny, V. (2009). Exponential bounds for minimum contrast estimators. *Electronic Journal of Statistics* **3**, 712–746.

Goodman, V. (1976). Distribution estimates for functionals of the two-parameter Wiener process. *The Annals of Probability* **4**, 977–982.

Hacking, I. (1975). *The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference*. Cambridge Univ Press.

Hjort, N. & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**.

Ibragimov, I. & Khasminskii, R. (1981). *Statistical estimation*. Springer.

Love, M. (1977). *Probability Theory I*. Springer Verlag.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability* **18**, 1269–1283.

Massart, P. (2007). *Concentration inequalities and model selection*. Citeseer.

Oliver, J. (1981). Marcus Aurelius and the Philosophical Schools at Athens. *The American Journal of Philology* **102**, 213–225.

Pitman, E. & Pitman, E. (1979). *Some basic theory for statistical inference*, vol. 216. Chapman and Hall.

Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Pólya, G. (1954). *Mathematics and Plausible Reasoning: Induction and analogy in mathematics*. Princeton Univ Pr.

Shiryaev, A. (1995). *Probability*. Springer, 2nd ed.

Shorack, G. & Wellner, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley.

Spokoiny, V. (2009). A penalized exponential risk bound in parametric estimation. *Arxiv preprint arXiv:0903.1721* .

Stroock, D. (2005). *An introduction to Markov processes*. Springer Verlag.

Talagrand, M. (1987). The Glivenko-Cantelli problem. *The Annals of Probability* **15**, 837–870.

Talagrand, M. (1994). The small ball problem for the Brownian sheet. *The Annals of Probability* **22**, 1331–1354.

Talagrand, M. (2005). *The generic chaining*. Springer.

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Verlag.

van de Geer, S. (2000). *Empirical processes in m-estimation*. Cambridge university press Cambridge.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* **16**, 264.

I

ON THE ERRORS COMMITTED BY SEQUENCES OF ESTIMATOR FUNCTIONALS

STEFFEN GRØNNEBERG AND NILS LID HJORT

ABSTRACT. Consider a sequence of estimators $\hat{\theta}_n$ which converges almost surely to θ_0 as the sample size n tends to infinity. Under weak smoothness conditions, we identify the asymptotic limit of the last time $\hat{\theta}_n$ is further than ε away from θ_0 when $\varepsilon \to 0^+$. These limits lead to the construction of sequentially fixed width confidence regions for which we find analytic approximations. The smoothness conditions we impose is that $\hat{\theta}_n$ is to be close to a Hadamard-differentiable functional of the empirical distribution, an assumption valid for a large class of widely used statistical estimators. Similar results were derived in Hjort and Fenstad (1992, Annals of Statistics) for the case of Euclidean parameter spaces; part of the present contribution is to lift these results to situations involving parameter functionals. The apparatus we develop is also used to derive appropriate limit distributions of other quantities related to the far tail of an almost surely convergent sequence of estimators, like the number of times the estimator is more than ε away from its target. We illustrate our results by giving a new sequential simultaneous confidence set for the cumulative hazard function based on the Nelson–Aalen estimator and investigate a problem in stochastic programming related to computational complexity.

1. Introduction and summary

Let $(\Omega, \mathcal{A}, P)$ be a probability space and P_n be the empirical distribution based on the first n observations from an infinite *iid* sample $X_1, X_2, \ldots$ from P living on some space $\mathcal{X}$. That is, let

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

be the seemingly naïve estimator of the distribution function P – which puts a point mass $1/n$ on every observed value in $\mathcal{X}$. Although P_n can never converge as a measure to P uniformly over the whole of $\mathcal{X}$ unless P is discrete, one can measure closeness between P_n and P relative to a set of mappings $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$ by perceiving P_n as an element of $l^\infty(\mathcal{F})$ evaluated as

$$P_n(f) := \int f \, \mathrm{d}P_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Key words and phrases. The last n, Hadamard-differentiable statistical functionals, Sequential confidence regions, Gaussian processes, the Nelson-Aalen estimator.

Likewise, one perceives P as an element of $l^\infty(\mathcal{F})$ evaluated as

$$P(f) := \int f \, \mathrm{d}P = \mathbb{E} f(X),$$

and ask how large can $\mathcal{F}$ be in order for P_n to be very close to P as $n \to \infty$.

A natural measure of closeness is the size of

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|. \tag{1}$$

As $\|P_n - P\|_{\mathcal{F}}$ may not be measurable, one can work with outer almost sure convergence and ask when

$$P^*\Big(\lim_{n\to\infty} \|P_n - P\|_{\mathcal{F}} = 0\Big) = 1,$$

defined in terms of the *outer measure* $P^*(B) = \inf\{P(A) : A \supset B, A \in \mathcal{A}\}$ for any $A \subseteq \Omega$. If this convergence takes place, $\mathcal{F}$ has the so-called Glivenko–Cantelli property. Characterizations of how large $\mathcal{F}$ may be relative to the structure of P is dealt with in the now classical expositions of Dudley (1999) and van der Vaart & Wellner (1996).

Supposing that $\mathcal{F}$ is Glivenko–Cantelli (that is, has the Glivenko–Cantelli property), it is natural to ask by which rate this convergence takes place. One way to approach this is to ask how rapidly a function $r(n) \nearrow \infty$ may grow in order to keep the size of

$$r(n)\|P_n - P\|_{\mathcal{F}}$$

stable in some appropriate sense. This leads us to discover that under reasonable conditions on $\mathcal{F}$, the rate $r(n) = \sqrt{n}$ gives

$$\sqrt{n}\|P_n - P\|_{\mathcal{F}} = O_{P^*}(1).$$

These developments are described in van der Vaart & Wellner (1996) and Dudley (1999), which gives conditions on $\mathcal{F}$ to be a so-called Donsker class – that is, conditions for $\sqrt{n}[P_n - P]$ to converge weakly in $l^\infty(\mathcal{F})$ to a P-Brownian Bridge in the Hoffman-Jørgensen sense.

These two levels of accuracy are of fundamental importance in asymptotic statistics and are connected in non-trivial ways. The present investigation concerns one such connection. Talagrand (1987)'s deep study of the Glivenko–Cantelli property of $\mathcal{F}$ shows (in his Theorem 22, see also Theorem 6.6.A of Dudley, 1999) that if $\mathcal{F}$ is Glivenko-Cantelli and made up of P-integrable measurable functions, then

$$\tilde{\Omega} := \Big\{\omega \in \Omega : \lim_{n\to\infty} \|P_n - P\|_{\mathcal{F}}(\omega) = 0\Big\} \tag{2}$$

is measurable (even though $\|P_n - P\|_{\mathcal{F}}$ need not be) and $P(\tilde{\Omega}) = 1$. This implies that on all of $\tilde{\Omega}$, there exists *a last time* an error larger than any prescribed $\varepsilon > 0$ is ever committed. Let

$$N_\varepsilon = \sup\{n : \|P_n - P\|_{\mathcal{F}} > \varepsilon\}$$

be the last time an error larger than $\varepsilon > 0$ is ever committed. Notice that by the definition of almost sure convergence,

$$\{N_\varepsilon < \infty \text{ for each } \varepsilon > 0\} = \tilde{\Omega}.$$

Hence, N_ε is finite with probability one even though N_ε may not be measurable. It natural to inquire into the size N_ε, and this question connects the two precision levels above in the following manner. Define $m = [y/\varepsilon^2]$ and $y_0 = \varepsilon^2[y/\varepsilon^2]$ so that

(3)
$$P(\varepsilon^2 N_\varepsilon > y) = P\left(\sup_{n \geq m} \|P_n - P\|_{\mathcal{F}} > \varepsilon\right) = P\left(\sup_{s \geq 1} \sqrt{m}\|P_{[ms]} - P\|_{\mathcal{F}} > \sqrt{y_0}\right).$$

So if $\sup_{s\geq 1} \sqrt{m}\|P_{[ms]} - P\|_{\mathcal{F}}$ has a non-trivial weak limit, we can use this to find distributional approximations of N_ε. What is needed is that the partial sum process

$$\mathbb{X}_n := \sqrt{n}(P_{[ns]} - P) \tag{4}$$

converges weakly on $l^\infty([1,\infty) \times \mathcal{F})$ to some non-trivial variable $\mathbb{X}$. This shows that

$$\sup_{s \geq 1} \sqrt{m}\|P_{[ms]} - P\|_{\mathcal{F}} = \|\mathbb{X}_n\|_{[1,\infty]\times\mathcal{F}} \xrightarrow[n\to\infty]{\mathscr{W}*} \|\mathbb{X}\|_{[1,\infty]\times\mathcal{F}}$$

by the continuous mapping theorem, which together with eq. (3) shows that

$$\varepsilon^2 N_\varepsilon \xrightarrow[\varepsilon\to 0^+]{\mathscr{W}*} \|\mathbb{X}\|^2_{[1,\infty]\times\mathcal{F}}. \tag{5}$$

The class $\mathcal{F}$ is called functional Donsker if the so-called sequential empirical process $\mathbb{Z}_n(s,f) = s\mathbb{X}_m(s,f)$ converges weakly on $[0,1] \times \mathcal{F}$ to a mean zero Gaussian process $\mathbb{Z}$ on $(0,1] \times \mathcal{F}$ with covariance structure

$$\text{Cov}\,(\mathbb{Z}(s,f), \mathbb{Z}(t,g)) = (s \wedge t)\,(Pfg - PfPg)\,, \tag{6}$$

called a Kiefer-Müller process. The set of functional Donsker classes and Donsker classes are in fact the same (see Chapter 12.2 of van der Vaart & Wellner, 1996), and the seemingly stronger statement of full $l^\infty([1,\infty) \times \mathcal{F})$ convergence of $\mathbb{X}_n$ to $s^{-1}\mathbb{Z}_s$ actually follows when $\mathcal{F}$ is functionally Donsker (Exercise 2.12.5 van der Vaart & Wellner, 1996). Time reversal of the Kiefer-Müller process (exercise 2.12.4 van der Vaart & Wellner, 1996) implies that $\mathbb{Z}(s,f) := \mathbb{X}_{1/s}(f)$ is a Kiefer-Müller process on $(0,1] \times \mathcal{F}$. Hence,

$$\varepsilon^2 N_\varepsilon \xrightarrow[n\to\infty]{\mathscr{W}*} \|\mathbb{X}\|^2_{[1,\infty]\times\mathcal{F}} = \|\mathbb{Z}\|^2_{(0,1]\times\mathcal{F}}$$

for a Kiefer-Müller process $\mathbb{Z}$ on $l^\infty((0,1] \times \mathcal{F})$ as long as $\mathcal{F}$ is Donsker. Thus, while the mere almost sure existence of N_ε is secured through the Glivenko–Cantelli property of $\mathcal{F}$, we get distributional approximations of N_ε from the Donsker property of $\mathcal{F}$.

The above questions are natural for any statistical estimator, and not just for the empirical distribution function. For a sequence of estimators $\{\hat{\theta}_n\}_{n=1}^\infty$ for which

$\hat{\theta}_n \xrightarrow[n\to\infty]{\text{a.s.*}} \theta$, we can define

$$N_\varepsilon = \sup\{n : \|\hat{\theta}_n - \theta\| > \varepsilon\}$$

where $\|\cdot\|$ is an appropriate norm. The present paper shows that the above connection between the Glivenko–Cantelli and Donsker properties of $\mathcal{F}$ is transferred from the empirical distribution function P_n over $\mathcal{F}$ to all estimators $\hat{\theta}$ which are (in an appropriate sense) close to being so-called Hadamard-differentiable statistical functionals of P_n over $\mathcal{F}$. The class of Hadamard-differentiable statistical functionals includes a fair portion of statistical estimators in use – for example Z-estimators with classical regularity conditions.

The investigation of N_ε for various estimators has a long history in probability and statistics, and goes back at least to Bahadur (1967). A steady stream of papers has worked with the subject, and we mention Robbins et al. (1968), Kao (1978), Stute (1983) and Hjort & Fenstad (1992). The theory contained in the present paper generalizes these investigations and puts them in a general framework.

The perhaps most obvious motivation for studying N_ε is to identify the probabilistic aspects that influence its limit distribution as $\varepsilon \to 0^+$. We will see that for Hadamard-differentiable statistical functionals, only the Hadamard-differential and the choice of norm in defining N_ε matters, besides the factors influencing the limiting distribution of the last time an error larger than ε is committed by the empirical distribution function itself. This gives a fresh and statistically motivated interpretation of the Hadamard-differential as a measure of variance.

We note that practically all statistical estimators can in principle be studied by only focusing on the empirical distribution. That is, for practically every possible estimator $\hat{\theta}_n$ taking values on some space $\mathbb{E}$, we can find a class $\mathcal{F}$ and nonrandom mapping $\phi_n : \mathbb{D}_n \subseteq l^\infty(\mathcal{F}) \mapsto \mathbb{E}$ so that

$$\hat{\theta}_n = \phi_n(P_n(f))$$

in which $\phi_n(P_n(f))$ is ϕ_n evaluated at the mapping $f \mapsto P_n(f)$. Clearly, the class of all estimators written as $\phi_n(P_n(f))$ is far too vast for a unified study, and we need to impose some restrictions on ϕ_n. Such a study was initiated in Hjort & Fenstad (1992) which identified the limit of $\varepsilon^2 N_\varepsilon$ when $\hat{\theta}_n = \bar{X}_n + R_n$ where $\bar{X}_n = P_n(\iota)$ is an *iid* average and equal to the empirical distribution evaluated at the identity functional, and R_n is small in the sense that $\sqrt{m}\sup_{n\geq m}|R_n| = o_P(1)$. They also worked with estimators of the form $\hat{\theta}_n = \phi(F_n)$ defined in terms of the classical empirical distribution function F_n and where ϕ was assumed to be so-called locally Lipschitz differentiable – a rather strong functional differentiation concept which implies Hadamard-differentiability. Such estimators can be written as $\phi(P_n(f))$ where f ranges over identity functions over $(-\infty, t)$ for $t \in \mathbb{R}$.

This paper studies maps $\phi_n = \phi$ which for a Donsker class $\mathcal{F}$ are Hadamard-differentiable and estimators $\hat{\theta}_n$ which are close to Hadamard-differentiable functionals in the sense that

$$\hat{\theta}_n = \phi_n(P_n(f)) = \phi(P_n(f)) + R_n$$

where again $\sqrt{m}\sup_{n\geq m}|R_n| = o_{P^*}(1)$. We then apply these limit theorems to provide new sequential fixed width confidence intervals for such estimators, and use tail approximations for Gaussian processes to provide approximations for the sizes involved in computing such confidence sets.

Hadamard-differentiability (henceforth H-differentiability) is a quite weak differentiability concept, which means that a very large class of statistical estimators can be written as H-differentiable statistical functionals of the empirical distribution. Examples include the Nelson–Aalen and Kaplan–Meier estimators, the empirical copula process and a large class of Z-estimators (see Section 3.9.4 of van der Vaart & Wellner, 1996). We say that a map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ defined on topological vector spaces $\mathbb{D}$ and $\mathbb{E}$ is H-differentiable tangentially to a set $\mathbb{D}_0 \subseteq \mathbb{D}$ if there is a continuous linear map $\dot{\phi}_\theta : \mathbb{D}_0 \mapsto \mathbb{E}$, such that

$$\lim_{n\to\infty} \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = \dot{\phi}_\theta(h) \tag{7}$$

for all converging sequences $t_n \to 0$ and $h_n \to h$ such that $h \in \mathbb{D}_0$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for every n. Let $\Delta_h(t) = \phi(\theta + th)$. If ϕ is H-differentiable at P, its H-differential is given by $\Delta_h'(0)$ where Δ' is the classical derivative. As we will deal with functionals of empirical distributions, we will work exclusively with $\mathbb{D} \subseteq l^\infty(\mathcal{F})$ and $\mathcal{E} = l^\infty(\mathcal{E})$ both equipped with the supremum norm. We will suppress the dependence which ϕ has on $\mathcal{F}$ and the use of the uniform norm, and write $\phi(P_n)$ instead of $\phi(P_n(f))$. However, whether or not ϕ is Hadamard-differentiable is clearly dependent on both $\mathcal{F}$ and the use of the uniform norm. See Remark 4 for further comments on this interplay.

H-differentiability is one of many possible functional generalizations of ordinary differentiation. The mathematical mathematical significance of H-differentiability is that it is the weakest functional differentiability concept which respects a chain-rule (Section A.5 Bickel et al., 1993). Its statistical significance is that it is the weakest differentiability concept which allows a generally applicable functional extension of the classical delta method of asymptotic statistics, called the functional delta method (see van der Vaart & Wellner, 1996). We note that the above definition we explicitly assumes that the H-differential is linear. This assumption can be avoided at the cost of a somewhat more involved theory. As the main results of this paper valid also under such a weakening, we follow the text of van der Vaart & Wellner (1996) by assuming that the differential is linear as it simplifies our presentation.

However, see Remark 2 for further discussion on the consequences of estimators with non-linear H-differential for our investigation.

As a concrete example of an H-differentiable estimator, consider the Nelson–Aalen estimator on $[0,\tau]$. Suppose that we observe $X_i = (Z_i, \Delta_i) \sim F$ where $Z_i = Y_i \wedge C_i$ and $\Delta_i = 1\{Y_i \leq C_i\}$ are defined in terms of unobservable *iid* failure times $Y_i < \tau$ distributed according to G and observable *iid* censoring times C_i. Under fairly general conditions, given e.g. in Shorack & Wellner (1986), the Nelson–Aalen estimator $\Lambda_n(t)$ converges almost surely to its limit, and we have

$$\Lambda_n(t) = \int_{[0,t]} \frac{1}{\bar{\mathbb{H}}_n} \mathrm{d}\mathbb{H}_n^{uc} \xrightarrow[n\to\infty]{\text{a.s.}} \Lambda(t) := \int_{[0,t]} \frac{1}{1-G(t)} \,\mathrm{d}G$$

where

$$\mathbb{H}_n^{uc}(t) = \frac{1}{n}\sum_{i=1}^n \Delta_i 1\{Z_i \leq t\} \qquad \text{and} \qquad \bar{\mathbb{H}}_n(t) = \frac{1}{n}\sum_{i=1}^n 1\{Z_i \geq t\}.$$

Let F_n be the bivariate empirical distribution of the observations $X_i = (Z_i, C_i)$. By van der Vaart & Wellner (1996, example 3.9.19), we can write

$$\Lambda_n(t) = \phi(F_n)$$

for an H-differentiable functional ϕ. This H-differentiability structure now leads to the famous process convergence of the Nelson–Aalen estimator

$$\sqrt{n}\,(\Lambda_n(t) - \Lambda(t)) \xrightarrow[n\to\infty]{\mathscr{W}_*} \dot\phi(\mathbb{Z})(t)$$

through a simple application of the functional delta method (see van der Vaart & Wellner, 1996, section 3.9), where $\mathbb{Z}$ is a P-Brownian Bridge on $[0,\tau) \times \{0,1\}$. In the same manner, our paper shows that if we let

$$N_\varepsilon = \sup\left\{ n \in \mathbb{N} : \sup_{0\leq t\leq\tau} |\Lambda_n(t) - \Lambda(t)| \geq \varepsilon \right\} = \sup\left\{ n \in \mathbb{N} : \|\Lambda_n - \Lambda\|_{[0,\tau]} \geq \varepsilon \right\},$$

the H-differentiability structure implies that

$$\varepsilon^2 N_\varepsilon \xrightarrow[n\to\infty]{\mathscr{W}_*} \left(\sup_{0\leq s\leq 1} \sup_{0\leq t\leq\tau} |\dot\phi(\mathbb{Z}_s)(t)| \right)^2 = \|\dot\phi\mathbb{Z}_s\|^2_{[0,1]\times[0,\tau]} \tag{8}$$

as an immediate consequence of our main result in Section 2, where $\mathbb{Z}_s(z,c)$ is a Kiefer-Müller process on $(0,1] \times [0,\tau) \times \{0,1\}$. In this case, $\dot\phi(\mathbb{Z}_s))(t)$ is also a martingale in t for each s. This allows the application of the theorem of Section 3.2, which simplifies the limit result of eq (8) to

$$\varepsilon^2 N_\varepsilon \xrightarrow[\varepsilon\to 0^+]{\mathscr{W}} \sigma^2 \left(\sup_{0\leq s\leq 1} \sup_{0\leq t\leq 1} |\mathbb{S}(s,t)| \right)^2 = \sigma^2 \|\mathbb{S}\|^2_{[0,1]^2}$$

for a Brownian Sheet $\mathbb{S}$ on $[0,1]^2$ where

$$\sigma^2 = \int_{[0,\tau]} \frac{1-\Delta\Lambda(z)}{P\{Z \geq z\}} d\Lambda(z).$$

We give an application of our limit results to sequential confidence sets in Section 3. The variable N_ε is the last passage time of an ε-ball in the uniform norm, and its limiting distribution can be used to construct sequential confidence sets. The limit distribution of $\varepsilon^2 N_\varepsilon$ is defined in terms of a supremum of a Gaussian mean zero process, and we utilize known tail-bounds for Gaussian processes to find closed form approximations to the fixed-width confidence sets.

This martingale structure simplifies the construction of sequential confidence sets, and Section 3.2 gives very tight approximations for the sizes needed to construct such sets when the limit distribution of $\sqrt{n}[\phi(P_n) - \phi(P)]$ is a martingale. This results in a new and easily calculated sequential confidence set for the Nelson–Aalen estimator. Indeed, let A^{-1} be the inverse of (the rapidly converging) sum

$$A(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \left[\Phi((2k+1)\lambda) - \Phi((2k-1)\lambda)\right] \tag{9}$$

in which Φ is the cumulative distribution function of a standard Gaussian random variable. We will show that for some $m \in [\sigma^2 A^{-1}(\sqrt{\alpha})^2/\varepsilon_0^2, \sigma^2 A^{-1}(\sqrt{\alpha}/2)^2/\varepsilon_0^2 + 1]$, we have that

$$P\left(\Lambda \in \left\{ f : \sup_{t\in[0,\tau]} |f(t) - \Lambda_n(t)| \le \varepsilon_0 \right\} \text{ for all } n \ge m\right)$$

is close to $1-\alpha$. In particular, the choice $m = \sigma^2 A^{-1}(\sqrt{\alpha}/2)^2/\varepsilon_0^2 + 1$ works.

Section 3.3 deals with related a problem arising in stochastic programming. Shapiro & Ruszczynski (2008) gives several practical applications in operations research where interest is in the value of $\min_{x\in X} g(x)$ where $g(x) = \mathbb{E}G(x,\xi)$ is the expected loss of a loss-function G defined in terms on a random vector ξ which has a known distribution. Often $g(x)$ is difficult to compute, but $G(x,\xi)$ is simpler to compute, while ξ is possible to simulate. This motivates approximating $\min g(x)$ by $\min \hat{g}(x)$ where $\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^n G(x,\xi_i)$ in which $\xi_1, \xi_2, \ldots, \xi_n$ are *iid* realizations of ξ. A natural question is how to choose n. Our general theory provides a well-motivated answer in a large class of cases, and we work out the details for a risk averse stochastic problem using a so-called absolute semideviation risk measure.

We conclude the paper with surveying other statistically relevant results connected or implied by our main result in Theorem 1. We propose two new measures of asymptotic relative efficiency and also prove convergence of variables related to N_ε. These variables are the number of errors larger than ε, the ratio of errors of sizes contained in $[a\varepsilon, b\varepsilon]$ relative to all errors larger than ε and the mean size of errors larger than ε. The two last variables have not been studied in the literature previously.

2. Limit Theorems

We will work under the following set of assumptions.

(1) (*Probability structure and spaces*) Assume given a sequence of *iid* observations $\{X_n\}_{n=1}^{\infty}$ living on a metric space space $\mathcal{X}$ and distributed according to P. Suppose that $\mathcal{F}$ is made up of real-valued measurable square-integrable functions from $\mathcal{X}$ to $\mathbb{R}$.
(2) (*Donsker structure*) Assume that $\mathcal{F}$ is Donsker (and hence Glivenko–Cantelli) with respect to P, and is bounded with respect to P in the sense that $\sup_x \sup_{f\in\mathcal{F}} |f(x) - Pf| < \infty$.
(3) (*Differentiability structure*) Assume that $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} = l^\infty(\mathcal{F}) \mapsto l^\infty(\mathcal{E}) =: \mathbb{E}$ is H-differentiable at P tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$. Denote the H-differential at P by $\dot\phi$.

Assumptions 1 and 2 are the basic assumptions of van der Vaart & Wellner (1996), while assumption 3 is the weakest form of H-differentiability used in the literature and assumes only differentiability at the single point P tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$.

H-differentiability at P implies that ϕ is continuous at P (Proposition A.5.1, Bickel et al., 1993), and secures that $\phi(P_n)$ converges outer almost surely to $\phi(P)$. In fact, the measurability of $\tilde\Omega$ of eq. (2) shows that $\phi(P_n)$ even converge almost surely to $\phi(P)$ and that

$$\tilde\Omega = \{P_n \to P\} = \{\phi(P_n) \to \phi(P)\} = \{N_\varepsilon < \infty \text{ for each } \varepsilon > 0\} \tag{10}$$

where

$$N_\varepsilon = \sup\{n : \|\phi(P_n) - \phi(P)\|_{\mathcal{E}} > \varepsilon\}.$$

Hence, $N_\varepsilon < \infty$ with probability one, even though neither N_ε nor $\phi(P_n)$ needs to be measurable.

Most of the work in deriving the limit behaviour of N_ε is done in the following lemma. It states that weak convergence of the partial sum process

$$(s, f) \mapsto \sqrt{n}\left[P_{[sn]} - P\right](f)$$

in $l^\infty([1,\infty) \times \mathcal{F})$ implies weak convergence of the partial "sum" (or "partial functional") process

$$(s, e) \mapsto \sqrt{n}\left[\phi(P_{[sn]}) - \phi(P)\right](e) \xrightarrow[n\to\infty]{\mathscr{W}_*} \dot\phi(s^{-1}\mathbb{Z}_s).$$

in $l^\infty([1,\infty) \times \mathcal{F})$ if ϕ is H-differentiable. In a certain sense, the lemma is a generalized version of the functional delta method. However, we will make use of the measurability of

$$\{\phi(P_n) \to \phi(P)\}$$

which is difficult to prove for other types of estimators. And so if such measurability conditions are in place also for other weakly converging sequences having a separable and Borel-measurable limit variable, the transference of weak convergence from partial sums to "partial functionals" is valid. However, we state the Lemma specifically for $\phi(P_n)$ for concreteness.

Lemma 1. *Under assumptions 1-3, we have that*

$$\sqrt{n}\left[\phi(P_{[sn]})(e) - \phi(P)(e)\right] \xrightarrow[n\to\infty]{\mathscr{W}*} \dot{\phi}(s^{-1}\mathbb{Z}_s)$$

on $l^\infty([1,\infty)\times\mathcal{F})$ *where* $\mathbb{Z}$ *is a Kiefer-Müller process on* $[1,\infty)\times\mathcal{F}$ *and* $\dot{\phi}(s^{-1}\mathbb{Z}_s)$ *is short-hand for* $\dot{\phi}$ *evaluated at the* $l^\infty(\mathcal{F})$*-map* $f \mapsto s^{-1}\mathbb{Z}_s(f)$*. The limit* $\dot{\phi}(s^{-1}\mathbb{Z}_s)$ *is a Gaussian process on* $l^\infty([1,\infty)\times\mathcal{E})$*.*

Proof. Recall that we assume that

$$\phi : \mathbb{D}_\phi \subseteq \mathbb{D} = l^\infty(\mathcal{F}) \mapsto l^\infty(\mathcal{E}) = \mathbb{E}$$

is H-differentiable at P tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}_\phi$. That is, there exists is a continuous linear map $\dot{\phi}_\theta : \mathbb{D}_0 \mapsto \mathbb{E}$, such that

$$\lim_{n\to\infty} \left\| \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} - \dot{\phi}_\theta(h) \right\|_{\mathcal{E}} = 0$$

for all converging sequences $t_n \to 0$ and $h_n \to h$ such that $h \in \mathbb{D}_0$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for every n. Define $h_s : \mathbb{D} \mapsto \mathbb{E}$ as the restriction map $h_s(f) = h(s_0, f)\big|_{s_0=s}$ for $h \in l^\infty([1,\infty)\times\mathcal{F})$ and let

$$\begin{aligned}
\mathbb{P}_\phi &= \left\{h \in l^\infty([1,\infty)\times\mathcal{F}) : \text{for all } s \geq 1,\ h_s \in \mathbb{D}_\phi\right\},\\
\mathbb{P}_0 &= \left\{h \in l^\infty([1,\infty)\times\mathcal{F}) : \text{for all } s \geq 1,\ h_s \in \mathbb{D}_0,\ \lim_{s\to\infty} h_s = 0\right\},\\
\mathbb{P}_n &= \left\{h \in l^\infty([1,\infty)\times\mathcal{F}) : \text{for all } s \geq 1,\ h_s \in \mathbb{D}_n,\ \lim_{s\to\infty} h_s = 0\right\}
\end{aligned}$$

where

$$\mathbb{D}_n = \left\{h \in l^\infty(\mathbb{F}) : P + \frac{1}{\sqrt{n}}\, h \in \mathbb{D}_\phi\right\}.$$

Define

$$\Phi : \mathbb{P}_\phi \mapsto l^\infty([1,\infty)\times\mathcal{E}), \dot{\Phi}_P : \mathbb{P}_0 \mapsto l^\infty([1,\infty)\times\mathcal{E})$$

by

$$\Phi(h)(s,e) = \phi(h_s)(e), \qquad \dot{\Phi}_P(h)(s,e) = \dot{\phi}(h_s)(e),$$

Define $g_n : \mathbb{P}_m \mapsto l^\infty([1,\infty)\times\mathcal{E})$ and $c_n : \mathbb{P}_m \mapsto l^\infty(\mathcal{E})$ by

$$g_n(h) = \sqrt{n}\left[\Phi\left(P + \frac{1}{\sqrt{n}}h\right) - \Phi(P)\right], \qquad c_n(h) = \sqrt{n}\left[\phi\left(P + \frac{1}{\sqrt{n}}h\right) - \phi(P)\right].$$

Although we know that H-differentiability of ϕ implies the validity of the extended continuous mapping theorem (Theorem 1.11.1 van der Vaart & Wellner, 1996) on c_n for the spaces $\mathbb{D}_n$ and $\mathbb{D}_0$, we wish to use the mapping theorem on g_n with the spaces $\mathbb{P}_n$ and $\mathbb{P}_0$. To do this, we suppose that $h_n \to h$ with $h_n \in \mathbb{P}_n$ and $h \in \mathbb{P}_0$ and must show that also $g_n(h_n) \to \dot{\Phi}(h)$. As $P + \frac{1}{\sqrt{n}}\, h_{n,s} \in \mathbb{D}_\phi$ for each s, H-differentiability of ϕ at P tangentially to $\mathbb{D}_0$ implies that

$$\sup_{e\in\mathcal{E}} |g_n(h_n)(s,e) - \dot{\phi}(h)(e)| \to 0$$

for each s, which is seemingly weaker than the required

$$\sup_{s\in[1,\infty),e\in\mathcal{E}} |g_n(h_n)(s,e) - \dot\phi(h)(e)| = \sup_{e\in\mathcal{E}} \sup_{s\in[1,\infty)} |g_n(h_n)(s,e) - \dot\Phi(h)(s,e)| \to 0.$$

However, the inner supremum must be achieved by an $s \in [1,\infty)$. Indeed, as $h_{n,s}$ is vanishing when $s \to \infty$, we have that

$$\lim_{s\to\infty} g_n(h_n)(s,e) = g_n(0) = \sqrt{n}\,[\Phi(P) - \Phi(P)] = 0$$

by the continuity of ϕ at P and

$$\lim_{s\to\infty} \dot\Phi(h)(s,e) = \dot\Phi(0) = 0$$

by the linearity of $\dot\phi$. Let $s(e)$ be the attained maximum of $\sup_{s\in[1,\infty)} |g_n(h_n)(s,e) - \dot\Phi(h)(s,e)|$ and pick, say, the smallest one if the point of maximum is not unique. We have that

$$\begin{aligned}\sup_{e\in\mathcal{E}} \sup_{s\in[1,\infty)} |g_n(h_n)(s,e) - \dot\Phi(h)(s,e)| &= \sup_{e\in\mathcal{E}} |g_n(h_n)(s,e) - \dot\Phi(h)(s,e)| \\ &= \sup_{e\in\mathcal{E}} |c_n(h_{s(e),n})(e) - \dot\Phi(h_{s(e)})(e)|.\end{aligned}$$

However, as $h_{n,s} \in \mathbb{D}_n$ and $h_s \in \mathbb{D}_0$ for any $s \geq 1$, we have that $\tilde h_n = h_{s(e),n}$ is just a sequence in $\mathbb{D}_n$ converging to $\tilde h = h_{s(e)}$, an element of $\mathbb{D}_0$. Indeed, let $e \in \mathcal{E}$ be given. Then

$$\|h_{s(e),n} - h_{s(e)}\|_{\mathcal{F}} \leq \sup_{s\geq 1} \|h_{n,s} - h_s\|_{\mathcal{F}} = \|h_n - h\|_{[1,\infty)\times\mathcal{F}} \to 0$$

where the convergence follows as we know that $h_n \to h$ in $l^\infty([1,\infty),\mathcal{F})$. We can conclude with $g_n(h_n) \to \dot\phi(h)$, proving the validity of the extended continuous mapping theorem.

As $\mathbb{X}_n = \sqrt{n}[P_{[sn]} - P]$ converges weakly to a separable limit on $l^\infty([1,\infty)\times\mathcal{F})$, we are left with showing that $\mathbb{X}_n$ is concentrated on $\mathbb{P}_n$. There are two defining properties of $\mathbb{P}_n$. The first is trivially fulfilled by $\mathbb{X}_n$ for each n. Notice that if ϕ is to be used as a statistical functional, clearly

$$P_n = P + \frac{1}{\sqrt{n}}\sqrt{n}[P_n - P] \in \mathbb{D}_\phi,$$

and hence

$$\sqrt{n}[P_n - P] \in \mathbb{D}_n = \left\{ q \in l^\infty(\mathcal{F}) : P + \frac{1}{\sqrt{n}}\, q \in \mathbb{D}_\phi \right\}.$$

for each n. As

$$P + \frac{1}{\sqrt{n}}\mathbb{X}_n = P + \frac{1}{\sqrt{n}}\,\sqrt{n}[P_{[sn]} - P] = P_{[sn]},$$

this means that also

$$P + \frac{1}{\sqrt{n}}\,\mathbb{X}_n(s,f) \in \mathbb{D}_n$$

for every $s \geq 1$.

However, the second defining property is only fulfilled with probability one. Indeed, Talagrand (1987) (see also Theorem 6.6.A of Dudley, 1999) shows that as $\mathcal{F}$ is Glivenko–Cantelli and made up of measurable and integrable functions, we have that

$$P\left(\lim_{n\to\infty} \|P_n - P\|_{\mathcal{F}} = 0\right) = 1,$$

even though $\|P_n - P\|_{\mathcal{F}}$ might not itself be measurable. As

$$\{\lim_{s\to\infty} \mathbb{X}_n(s,e) = 0\} = \{\lim_{n\to\infty} \|P_n - P\|_{\mathcal{F}} = 0\} =: \tilde{\Omega},$$

the process $\mathbb{X}_n$ is included in $\mathbb{P}_n$ with probability one, which suffices to allow the application of the extended continuous mapping theorem, as the exclusion of a *measurable* set with probability zero does not change the (outer) probability structure of the problem. This is seen as follows. Given a $B \subseteq \Omega$, we have that

$$P^*(B \cap \tilde{\Omega}) = P\left(\left(B \cap \tilde{\Omega}\right)^*\right) = P(B^* \cap \tilde{\Omega}) = P(B^*) = P^*(B),$$

where the second equality comes from the measurability of $\tilde{\Omega}^C$ and exercise 1.2.15 in van der Vaart & Wellner (1996). Hence, we may conclude with

$$\sqrt{m}\left[\phi(P_{[sn]}) - \phi(P)\right] = g_n(t, \mathbb{X}_n) \xrightarrow[n\to\infty]{\mathscr{W}_*} \dot{\Phi}_P(\mathbb{X}_s) = \dot{\phi}(s^{-1}\mathbb{Z}_s)$$

on $[1,\infty) \times \mathcal{E}$ for a Kiefer-Müller process $\mathbb{Z}$ on $[1,\infty) \times \mathcal{F}$ from the extended continuous mapping theorem. Finally, the Gaussianity of the limit process follows either from the functional definition of Gaussian processes in Banach spaces or Lemma 3.9.8 of van der Vaart & Wellner (1996). □

Theorem 1. *Let $\mathbb{Z}_s(f) = \mathbb{Z}(s,f)$ be a Kiefer-Müller process indexed by $[0,1)\times\mathcal{F}$ and $\dot{\phi}\mathbb{Z}_s$ is $\dot{\phi}$ evaluated at the map $f \mapsto \mathbb{Z}_s(f)$. Given assumptions 1-3, the following is true.*

(1) *For $N_\varepsilon = \sup\{n : \|\phi(P_n) - \phi(P)\|_{\mathcal{F}}\}$, we have that*

$$\varepsilon^2 N_\varepsilon \xrightarrow[n\to\infty]{\mathscr{W}_*} \|\dot{\phi}\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}}. \tag{11}$$

(2) *Given an estimator $\hat{\theta}_n \xrightarrow[n\to\infty]{a.s.*} \theta$, let $N_\varepsilon = \sup\{n : \|\hat{\theta}_n - \theta\|_{\mathcal{E}} > \varepsilon\}$. Assume $\hat{\theta}_n$ is close to being H-differentiable in the sense that $\hat{\theta}_n = \phi(P_n) + R_n$ where $\sqrt{m}\sup_{n\geq m}\|R_n\|_{\mathcal{E}}$ is $o_{P^*}(1)$. We then have*

$$\varepsilon^2 N_\varepsilon \xrightarrow[n\to\infty]{\mathscr{W}_*} \|\dot{\phi}\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}}. \tag{12}$$

In both cases, $\dot{\phi}\mathbb{Z}_s$ is a zero mean Gaussian process. If $\mathbb{D}_0$ is a linear space, then $\dot{\phi}\mathbb{Z}_s$ has a covariance function with the product structure

$$\rho\left((s_1,e_1),(s_2,e_2)\right) := \mathbb{E}\dot{\phi}\mathbb{Z}_{s_1}(e_1)\dot{\phi}\mathbb{Z}_{s_2}(e_2) = (s_1 \wedge s_2)\mathbb{E}\dot{\phi}W^\circ(e_1)\dot{\phi}W^\circ(e_2). \tag{13}$$

where W° is a P-Brownian bridge process on $\mathcal{F}$.

Proof. For the first part, we note that in light of eq. (3), it suffices to identify the weak limit of $\sup_{n\geq m}\sqrt{m}\|\phi(P_n)-\phi(P)\|_{\mathcal{E}}$. Thanks to the Lemma, this is easy, as

$$\begin{aligned}\sup_{n\geq m}\sqrt{m}\|\phi(P_n)-\phi(P)\|_{\mathcal{E}} &= \sup_{s\geq 1}\|\phi(P_{[sn]})-\phi(P)\|_{\mathcal{E}} = \sqrt{m}[\Phi(\mathbb{X}_m)-\phi(P)]\|_{\mathcal{E}} \\ &= \|\sqrt{m}[\Phi(\mathbb{X}_m)-\phi(P)]\|_{[1,\infty)\times\mathcal{E}} \xrightarrow[n\to\infty]{\mathscr{W}_*} \|\dot{\phi}s^{-1}\tilde{\mathbb{Z}}_s\|_{[1,\infty)\times\mathcal{E}}\end{aligned}$$

by the continuous mapping theorem. Finally, we know that $\mathbb{Z}_s(f) = s^{-1}\tilde{\mathbb{Z}}_{1/s}(f)$ is a Kiefer-Müller process on $(0,1]\times\mathcal{F}$. This proves the first claim, and we can readily extend this case to the second claim. Note that

$$P^*(\varepsilon^2 N_\varepsilon > y) = P^*\left(\sup_{s\geq 1}\sqrt{m}\|\hat{\theta}_{[ms]}-\theta\|_{\mathcal{E}} > \sqrt{y_0}\right).$$

Thanks to Lemma 1.10.2 (i) of van der Vaart & Wellner (1996), the stated convergence follows if

$$\left|\sup_{s\geq 1}\sqrt{m}\|\hat{\theta}_{[ms]}-\theta\| - \sup_{s\geq 1}\sqrt{m}\|\phi(P_{[ms]})-\theta\|_{\mathcal{E}}\right| \xrightarrow[n\to\infty]{\mathscr{P}_*} 0.$$

However, $\sup_{s\geq 1}\|\cdot\|_{\mathcal{E}} = \|\cdot\|_{[1,\infty)\times\mathcal{E}}$ respects the triangle inequality, so that the above difference is bounded by $\sqrt{m}\sup_{n\geq m}\|R_n\|_{\mathcal{E}}$ which converge to zero in probability by assumption.

We are left with proving that $\dot{\phi}\mathbb{Z}$ has the stated covariance structure of eq. (13). Construct a sequence $W_1^\circ, W_2^\circ, \ldots$ of independent P-Brownian Bridges, and define

$$\mathbb{Z}_n(s,f) := \frac{1}{\sqrt{n}}\sum_{i=1}^{[ns]} W_i^\circ(f)$$

which is a Gaussian mean zero process with covariance function given by

$$\operatorname{Cov}\left[\mathbb{Z}_n(s_1,f_1),\mathbb{Z}_n(s_2,f_2))\right] = \frac{[ns_1]\wedge[ns_2]}{n}\operatorname{Cov}\left[\mathbb{Z}_n(1,f_1),\mathbb{Z}_n(1,f_2)\right].$$

This covariance function converges to the covariance function of a Kiefer-Müller process on $(0,1]\times\mathcal{F}$, so that the finite dimensional distributions of $\mathbb{Z}_n$ converge weakly to those of $\mathbb{Z}$. We now prove that $\mathbb{Z}_n$ is tight so that $\mathbb{Z}_n \xrightarrow[n\to\infty]{\mathscr{W}_*} \mathbb{Z}$. Let $\varrho_P(f) = (P(f-Pf)^2)^{1/2}$ be the variance seminorm. Following the proof of Theorem 2.12.1 of van der Vaart & Wellner (1996), we need to show that for any $\varepsilon,\eta>0$, there exists a $\delta>0$ so that

$$\limsup_{n\to\infty} P^*\left(\sup_{|s-t|+\varrho(f,g)<\delta}|\mathbb{Z}_n(s,f)-\mathbb{Z}(t,g)|>\varepsilon\right) < \eta.$$

By the triangle inequality, the supremum in the above display is bounded by

$$\sup_{|s-t|<\delta}\|\mathbb{Z}_n(s,f)-\mathbb{Z}_n(t,f)\|_{\mathcal{F}} + \sup_{0\leq t\leq 1}\|\mathbb{Z}_n(t,f)\|_{\mathcal{F}_\delta} \tag{14}$$

where $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \varrho(f-g) < \delta\}$. We can hence bound the probability of each of these terms being larger than ε separately. By the generalized Lévy inequality (see e.g. De la Pena & Gine, 1999, Theorem 1.1.5), we have that

$$\begin{aligned} P\left(\sup_{0\leq t\leq 1} \|\mathbb{Z}_n(t,f)\|_{\mathcal{F}_\delta} > \varepsilon\right) &= P\left(\max_{k\leq n} \|\frac{1}{\sqrt{n}}\sum_{i=1}^{k} W_i^\circ(f)\|_{\mathcal{F}_\delta} > \varepsilon\right) \\ &\leq 9P\left(\|\mathbb{Z}_n(1,f)\|_{\mathcal{F}_\delta} > \varepsilon/30\right). \end{aligned}$$

An inspection of the covariance of $\mathbb{Z}_n(1,f)$ reveals that it is a P-Brownian Bridge for each n. As $\mathcal{F}$ is Donsker, a P-Brownian Bridge is continuous with respect to ϱ_P, so that $\|\mathbb{Z}_n(1,f)\|_{\mathcal{F}_\delta}$ converges to zero in probability as $\delta \to 0^+$. To bound the probability that the first term of eq. (14) is larger than ε, the arguments contained in the proof of Theorem 2.12.1 in van der Vaart & Wellner (1996) imply that

$$\begin{aligned} P\left(\sup_{|s-t|<\delta} \|\mathbb{Z}_n(s,f) - \mathbb{Z}_n(t,f)\|_{\mathcal{F}} > \varepsilon\right) &\leq \left\lceil \frac{1}{\delta} \right\rceil P\left(\max_{k\leq n\delta} \|\frac{1}{\sqrt{n}}\sum_{i=1}^{k} W_i^\circ(f)\|_{\mathcal{F}} > \varepsilon\right) \\ &= \left\lceil \frac{1}{\delta} \right\rceil P\left(\max_{k\leq n\delta} \|\frac{1}{\sqrt{\delta n}}\sum_{i=1}^{k} W_i^\circ(f)\|_{\mathcal{F}} > \frac{\varepsilon}{\delta}\right). \end{aligned}$$

Note again that $\mathbb{Z}_{n\delta}$ is a P-Brownian Bridge W° for each n. By the generalized Lévy inequality, the above display is bounded by

$$9\left\lceil \frac{1}{\delta} \right\rceil P\left(\|\mathbb{Z}_{n\delta}(1,f)\|_{\mathcal{F}} > \frac{\varepsilon}{30\delta}\right) = 9\left\lceil \frac{1}{\delta} \right\rceil P\left(\|W^\circ\|_{\mathcal{F}} > \frac{\varepsilon}{30\delta}\right).$$

the finite second moment of $\|W^\circ\|_{\mathcal{F}}$ (van der Vaart & Wellner, 1996, Lemma 2.3.9) enables us to envoke the Borell inequality (van der Vaart & Wellner, 1996, Proposition A.2.1) which imples that $\|W^\circ\|_{\mathcal{F}}$ has exponentially decreasing tails. Hence, the above display converges to zero. We assumed that $\mathbb{D}_0$ is a linear space, so that we can apply $\dot{\phi}$ to $\mathbb{Z}_n$, which converges weakly to $\dot{\phi}\mathbb{Z}$ by the continuous mapping theorem. The linearity of $\dot{\phi}$ also shows that

$$\dot{\phi}\mathbb{Z}_n(s,e) = \frac{1}{\sqrt{n}}\sum_{i=1}^{[ns]} \dot{\phi}W_i^\circ(e),$$

which has covariance function

$$\begin{aligned} \rho_n\left((s_1,e_1),(s_2,e_2)\right) &= \mathrm{Cov}\left[\dot{\phi}(\mathbb{Z}_n(s_1,f))(e_1), \dot{\phi}(\mathbb{Z}_n(s_2,f))(e_2)\right] \\ &= \frac{[ns_1]\wedge[ns_2]}{n}\,\mathrm{Cov}\left[\dot{\phi}(\mathbb{Z}_n(1,f))(e_1), \dot{\phi}(\mathbb{Z}_n(1,f))(e_2)\right]. \end{aligned}$$

As $\dot{\phi}\mathbb{Z}_n$ is Gaussian and converges weakly to $\dot{\phi}\mathbb{Z}$ and as $\dot{\phi}\mathbb{Z}_1 = \dot{\phi}W^\circ$ for a P-Brownian Bridge W°, we have that $\rho_n \to \rho$, where ρ is defined in eq (13). □

Several remarks are in order.

Remark 1. When $\phi(P_n)$ is a random variable, so that $\mathcal{E} = \{e\}$ is a singleton, the covariance structure of eq. (13) shows that $\dot\phi\mathbb{Z}_s = \sqrt{\operatorname{Var} \mathrm{IF}_\phi(X)}\mathbb{B}_s$ for a Brownian Motion $\mathbb{B}_s$ and where IF_ϕ is the influence function of ϕ. Thus Theorem 1 is a proper generalization of the basic result in Hjort & Fenstad (1992).

Remark 2. We note that the proofs of Lemma 1 and the first two parts of Theorem 1 does not use the assumed linearity of $\dot\phi$, and is still true when the definition of H-differentiability is weakened to only assume eq. (7). The chain-rule still applies, and several new maps can be shown to be H-differentiable in this weaker sense. See Römisch (2005) for a survey of such results. Our proof also applies in the case of set-valued functionals when an appropriate metric for comparing sets is assumed, such as the Attouch-Wets topology.

Remark 3. The limit of $\varepsilon^2 N_\varepsilon$ depends only on three things. Firstly, the Kiefer-Müller process is a mean zero Gaussian process, with covariance structure defined through P. Secondly, both N_ε and the limit variable is defined in terms of the uniform topology on $\mathcal{E}$. Thirdly, while N_ε is defined in terms of the full ϕ, the limit only depends on the much simpler $\dot\phi$. This is interesting from a statistical perspective and motivates the definition of

$$\sigma^2 := \frac{\operatorname{Median}\|\dot\phi\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}}}{\operatorname{Median}\|\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{F}}} \tag{15}$$

as a measure of variance for $\phi(P_n)$. There are two main reasons for scaling the median of the limit variable of $\varepsilon^2 N_\varepsilon$ with $\operatorname{Median}\|\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{F}}$. Firstly, all stochasticity of $\theta_n = \phi(P_n)$ originates from P_n, making it natural to separate the variability of P_n and the variability inherent in the structure of ϕ itself. Secondly, notice that if $\hat\theta = \bar X_n$ is the empirical mean of iid random variables $X_1, X_2, \ldots, X_n$, then $\dot\phi\mathbb{Z}_s = \sigma B_s$ for a Brownian Motion process B_s. Hence,

$$\operatorname{Median}\|\dot\phi\mathbb{Z}_s\|^2 = \sigma^2 \operatorname{Median} \sup_{0\le s\le 1} |B_s|^2.$$

so that the σ^2 of eq. (15) coincides with the standard definition of variance.

Remark 4. The structure of the class of H-differentiable functionals depends on the topology of both $\mathbb{D}$ and $\mathbb{E}$. For a collection $\mathcal{C} \subseteq \mathbb{D}$ we call ϕ a $\mathcal{C}$-differentiable functional at θ if

$$\lim_{t\to 0} \sup_{h\in\mathcal{C},\, \theta+th\in\mathbb{D}_\phi} \left\| \frac{\phi(\theta+th)}{t} - \dot\phi_\theta(h) \right\| = 0.$$

H-differentiability is equivalent to $\mathcal{C}$-differentiability when $\mathcal{C}$ is the class of all compact sets. If other topologies on $\mathbb{D}$ or $\mathbb{E}$ are used, this changes the class of H-differentiable functionals in non-trivial ways. We note that the investigation of Dudley (1992) works with Fréchet differentiability functionals with p-variation norms on the $\mathbb{D}$-space. Fréchet differentiability is $\mathcal{C}$-differentiability when $\mathcal{C}$ is the class of all

bounded sets of $\mathbb{D}$, which is strictly stronger than H-differentiability – when the same topology is used. However, the classes of H-differentiable and Fréchet differentiable functionals are incommensurable when different topologies are used. See Section 5.2 of Shao (2003) for examples of this incommensurability, and exercise 5.27 of Shao (2003) for a class of functionals of the classical empirical distribution which are Fréchet differentiable with respect to the L_1-norm, but not H-differentiable with respect to the uniform norm. We have followed van der Vaart & Wellner (1996) in working with the uniform topology on both $\mathbb{D}$ and $\mathbb{E}$.

Remark 5. When working with estimators of the form $\hat{\theta}_n = \phi(P_n) + R_n$, we can no longer guarantee the measurability of $\{N_\varepsilon < \infty \text{ for each } \varepsilon > 0\}$ as eq. (10) need not hold. If $R_n \not\equiv 0$ but $R_n \xrightarrow[n\to\infty]{\text{a.s.*}} 0$, this only provides a the existence of a version of the measurable cover of $\|\hat{\theta}_n - \phi(P)\|$, which we denote by $\|\hat{\theta}_n - \phi(P)\|^\star$, that converges to zero almost surely. Although the convergence of eq. (12) is valid without measurability, we can only guarantee the measurability of $\{N_\varepsilon^\star < \infty\}$ for $\varepsilon > 0$ where $N_\varepsilon^\star := \sup\{n : \|\hat{\theta}_n - \theta\|_{\mathcal{E}}^\star > \varepsilon\}$.

3. Sequential confidence sets

As in Hjort & Fenstad (1992) and Stute (1983), our results about the limiting distribution of $\varepsilon^2 N_\varepsilon$ can be used to construct sequential fixed-volume confidence regions. As our limit result encompasses all H-differentiable functionals, this leads to new confidence sets for many estimators, the Nelson–Aalen estimator being one of them. In this connection we remark that Bandyopadhyay et al. (2003) find fixed-value confidence intervals for the H-differentiable functional

$$\phi(F_{X,Y}) = \int F_X \, \mathrm{d}F_Y = P(X \le Y). \tag{16}$$

The basis for their construction of a fix-volume confidence set for $P(X \le Y)$ is a direct application of a special case of Theorem 1.

The connection between the limit of N_ε and the construction of fixed-width confidence sets is as follows. Calculate or approximate the upper α quantile of the limit variable of the theorem and denote this quantile by λ_α. Fix the radius of the confidence set as ε_0 and compute $m = [\lambda_\alpha/\varepsilon_0^2]$. By the distributional convergence, we get that

$$\begin{aligned} P(\varepsilon^2 N_\varepsilon < \lambda_\alpha) &= P(\|\phi(P_n) - \phi(P)\|_{\mathcal{E}} \le \varepsilon_0 \text{ for all } n \ge m) \\ &= P(\phi(P) \in B\left(\varepsilon_0, \phi(P_n)\right) \text{ for all } n \ge m) \end{aligned} \tag{17}$$

is close to $1 - \alpha$ where

$$B(\varepsilon, y) = \{x : \|x - y\|_{\mathcal{E}} \le \varepsilon\}$$

is an ε-ball in $l^\infty(\mathcal{E})$. This has intuitive appeal. Whereas confidence sets are usually of the form

$$P(\phi(P) \in C_n) \geq 1 - \alpha, \qquad \text{for all } n \geq m$$

and thus only give a probability statement for one $n \geq m$ at the time, a fixed-volume confidence set gives a simultaneous answer for all $n \geq m$. This is intuitively pleasing, and Hjort & Fenstad (1992) humorously mentioned that even Serfling's physician (Serfling, 1980, page 49) is interested in sequential fixed-volume confidence regions.

The difficult step in constructing the fixed width confidence set of eq. (17) is to calculate λ_α. In some special cases, as in the case of eq. (16), the limit distribution of $\varepsilon^2 N_\varepsilon$ can be found in a closed form expression. This seems out of reach for a completely general H-differentiable ϕ. However, in some cases we can find useful approximations for tail-probabilities of $\|\dot\phi \mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}}$. Although this quantile can in theory be simulated directly from the Donsker Theorem, this is often very time consuming, if even possible.

When the limit variable $\dot\phi\mathbb{Z}_s$ is Gaussian, we have the well-developed theory of Gaussian tail bounds at our disposal. Under typical conditions, $\dot\phi\mathbb{Z}_s$ has zero mean – see Section 3.9.2 of van der Vaart & Wellner (1996). In this case we can use Proposition A.2.1 of van der Vaart & Wellner (1996) that gives the Borell inequality in the form

$$P(\|\dot\phi\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}} \geq \lambda) = P(\|\dot\phi\mathbb{Z}_s\|_{(0,1]\times\mathcal{E}} \geq \sqrt{\lambda}) < 2\exp\left(-\frac{\lambda}{8\mathbb{E}\|\dot\phi\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}}}\right) \tag{18}$$

for all $\lambda > 0$. The following Lemma shows that the above inequalities are non-trivial under our assumptions.

Lemma 2. *Let $\mathbb{Z}_s(f) = \mathbb{Z}(s,f)$ be a Kiefer-Müller process indexed by $[0,1)\times\mathcal{F}$ and $\dot\phi\mathbb{Z}_s$ is $\dot\phi$ evaluated at the map $f \mapsto \mathbb{Z}_s(f)$. Given assumptions 1-3, $\|\dot\phi\mathbb{Z}_s\|_{(0,1]\times\mathcal{E}}$ has finite second moment.*

Proof. By Proposition 1 below, we have

$$\mathbb{E}\|\dot\phi\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}} = \int_0^\infty P(\|\dot\phi\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}} > x)\,\mathrm{d}x \leq 2\int_0^\infty P(\|\dot\phi\mathbb{Z}_s\|^2_{\mathcal{E}} > x)\,\mathrm{d}x = 2\mathbb{E}\|\dot\phi\mathbb{Z}\|^2_{\mathcal{E}}$$

As $\dot\phi\mathbb{Z}$ is the weak limit of $\sqrt{n}[\phi(P_n) - \phi(P)]$ as $n \to \infty$, Lemma 2.3.9 of van der Vaart & Wellner (1996) shows that $\mathbb{E}\|\dot\phi\mathbb{Z}\|^2_{\mathcal{E}}$ is finite. □

The expectation of inequality 18 is simpler to approximate than the full distribution of $\|\dot\phi\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{E}}$ and provides a general bound for λ_α. However, $\mathbb{E}\|\dot\phi\mathbb{Z}\|^2_{\mathcal{E}}$ is often difficult to compute and the constants involved can be improved in special cases. The following subsections gives explicit bounds for some classes of special cases.

Remark 6. The confidence sets presented in this section rely on the approximation $P(\varepsilon^2 N_\varepsilon < \lambda_\alpha) \approx 1 - \alpha$ through Theorem 1. An alternative construction

of approximate sequential confidence sets for a fixed $\varepsilon > 0$ can be based on the following observation. Let

$$(s, e) \mapsto R_{ms}(e) = \left[\phi(P_{\lfloor ms \rfloor})(e) - \phi(P)(e)\right] - \left[\dot{\phi}(P_{\lfloor ms \rfloor} - P)\right]$$

and suppose a bound of the type

$$P\left(\sup_{s \geq 1, e \in \mathcal{E}} |R_{ms}(e)| > y\right) \leq r(y) \tag{19}$$

is known. Following the notation of Section 1, the triangle inequality shows that

$$P(\varepsilon^2 N_\varepsilon > y) \leq P\left(\sqrt{m} \sup_{s \geq 1, e \in \mathcal{E}} |\dot{\phi}(P_{\lfloor ms \rfloor} - P)(e)| > \sqrt{y_0}/2\right) + r\left(\sqrt{y_0}/2\right). \tag{20}$$

By the linearity of $\dot{\phi}$, the first term is the supremum of a sequential empirical process, for which non-asymptotic bounds exist. The inequality of Talagrand (1996) applies to sequential empirical processes as well, as it is proved through estimating the Laplace transform, and the exponentiated partial sum is a submartingale, so that Doob's inequality can be applied. However, although good constants for the Talagrand inequality are given in Massart (2000) for the non-sequential empirical process, we are unaware of analogous results for the sequential case. Supposing such constants known, one could bound any quantile from eq. (20). However, it may be difficult to find useful r-functions for eq. (19). Analogously to the unspecified precision underlying $P(\varepsilon^2 N_\varepsilon < \lambda_\alpha) \approx 1 - \alpha$, one could also give conditions securing $\sup_{s \geq 1, e \in \mathcal{E}} |R_{ms}(e)| = o_p(1)$ and ignore the second term of eq. (20) when solving for y in eq. (20).

3.1. **A reduction to the Kolmogorov–Smirnov limit.** The weak limit of $\varepsilon^2 N_\varepsilon$ is almost the limit of the Kolmogorov–Smirnov Goodness-of-fit functional for the estimator $\phi(P_n)$. Approximating such goodness-of-fit limits is a well-known problem and have been studied in many settings. The following result relates the $\varepsilon^2 N_\varepsilon$ limit to that of the Kolmogorov–Smirnov functional.

Proposition 1. *Let $\mathbb{Z}_s(f) = \mathbb{Z}(s, f)$ be a Kiefer-Müller process indexed by $[0, 1) \times \mathcal{F}$ and $\dot{\phi}\mathbb{Z}_s$ is $\dot{\phi}$ evaluated at the map $f \mapsto \mathbb{Z}_s(f)$. Given assumptions 1-3, we have*

$$P(\|\dot{\phi}\mathbb{Z}_s\|_{(0,1] \times \mathcal{E}} > \lambda) \leq 2P(\|\dot{\phi}\mathbb{Z}\|_{\mathcal{E}} > \lambda).$$

where $\mathbb{Z}$ is an $\mathcal{F}$-Brownian Bridge.

Proof. Fix an integer $k > 0$ and let $m = 2^k$. For $k = 1, 2, \ldots, m$ and $t \in [0, 1]^d$, let

$$U_k(e) = \dot{\phi}\mathbb{Z}_{j/m}(e) - \dot{\phi}\mathbb{Z}_{(j-1)/m}(e)$$

which is a symmetric stochastic process, and where $U_1, U_2, \ldots, U_k$ are independent of each other. As $\dot{\phi}\mathbb{Z}_{j/m}(e) = \sum_{i=1}^{j} U_i(e)$, the general Lévy's inequality given e.g.

in Proposition A.1.2 in van der Vaart & Wellner (1996), shows that

$$P\left(\sup_{1\leq j\leq m}\|\dot{\phi}\mathbb{Z}_{j/m}\|_{\mathcal{E}}>\lambda\right)=P\left(\sup_{1\leq j\leq m}\left\|\sum_{i=1}^{j}U_i\right\|_{\mathcal{E}}>\lambda\right)\leq 2P\left(\left\|\sum_{i=1}^{m}U_i\right\|_{\mathcal{E}}>\lambda\right),$$

which equals $2P(\|\dot{\phi}\mathbb{Z}_1\|_{\mathcal{E}}>\lambda)$. As $\mathbb{Z}_1$ is an $\mathcal{F}$-Brownian Bridge, the claimed upper bound follows from monotone convergence as $k\to\infty$. □

The above result leads e.g. to explicit bounds for the limit distribution of $\varepsilon^2 N_\varepsilon$ for the two-dimensional empirical distribution function through the results of Adler & Brown (1986). Let $\mathbb{W}$ be a two-dimensional real valued F-Brownian-Bridge on $\mathbb{R}^2$ and $\mathbb{K}$ an F-Kiefer-process on $(0,1]\times\mathbb{R}^2$. The above lemma, symmetry of zero mean Gaussian processes and Theorem 3.1 of Adler & Brown (1986) shows that for any F, we have

$$\begin{aligned}P\left(\sup_{(s,t)\in(0,1]\times\mathbb{R}^2}|\mathbb{Z}_s(t)|>\sqrt{\lambda}\right)&\leq 2P\left(\sup_{t\in\mathbb{R}^2}|\mathbb{W}(t)|>\sqrt{\lambda}\right)\\&\leq 4P\left(\sup_{t\in\mathbb{R}^2}\mathbb{W}(t)>\sqrt{\lambda}\right)\leq 4\sum_{k=1}^{\infty}(8k^2\lambda-2)e^{-2k^2\lambda}.\end{aligned}$$

3.2. **Gaussian Local Martingales.** If $\dot{\phi}W^\circ$ is a univariate local martingale indexed by $[0,\tau)$ the limit variable of N_ε has a particularly simple structure.

Theorem 2. *Assume that $\mathbb{D}_0$ is linear, that $\mathcal{E}$ is $[0,\tau)$ for some $0<\tau<\infty$, and that for each s, the process $\dot{\phi}(\mathbb{Z}_s)(t)$ is a square integrable continuous local martingale in t starting at zero. Let $\left\langle\dot{\phi}W^\circ,\dot{\phi}W^\circ\right\rangle_s$ be the covariation process of $\dot{\phi}W^\circ$ and define $\sigma^2(t)=\inf\left\{s:\left\langle\dot{\phi}W^\circ,\dot{\phi}W^\circ\right\rangle_s>t\right\}$. Then the limit variable of Theorem 1 has the same distribution as $\sigma^2\|\mathbb{S}\|^2_{[0,1]^2}$ where $\mathbb{S}$ is a Brownian Sheet on $[0,1]^2$ and $\sigma^2=\sigma^2(\tau)$ is non-stochastic.*

Proof. The Dambis Dubuins-Schwarz Theorem (Revuz & Yor, 1999, Theorem V.1.6) shows that there exists a version W of Brownian Motion so that $W(\sigma^2(t))=\dot{\phi}W^\circ(t)$. As $\dot{\phi}W^\circ$ is a continuous mean zero Gaussian process with a product covariance structure given by eq. (13), its quadratic variation process is non-stochastic (see exercise V.1.14 Revuz & Yor, 1999). Hence,

$$\mathbb{E}\dot{\phi}W^\circ(t)\dot{\phi}W^\circ(s)=\mathbb{E}W(\sigma^2(t))W(\sigma^2(s))=\sigma^2(t)\wedge\sigma^2(s).$$

Theorem 1 shows that $\dot{\phi}\mathbb{Z}$ is a continuous mean zero Gaussian process with a product covariance structure given by eq. (13). As the distribution of a mean zero Gaussian process is determined by its covariance structure, this shows that defining $\mathbb{S}$ by $\dot{\phi}\mathbb{Z}=\mathbb{S}(s,\sigma^2(t))$ makes $\mathbb{S}(s,t)$ a Brownian Sheet on $[0,1]\times[0,\sigma^2(\tau)]$. Let N be the

limit variable of Theorem 1. As $\dot{\phi}W^\circ$ is continuous, its quadratic variation is also continuous, which makes its inverse $\sigma^2(t)$ continuous as well. Hence,

$$N = \left(\sup_{0\le s\le 1} \sup_{0\le t\le \tau} \left| \mathbb{S}(s, \sigma^2(t)) \right| \right)^2 = \left(\sup_{0\le s\le 1} \sup_{0\le t\le 1} \left| \mathbb{S}(s, t\sigma^2(\tau)) \right| \right)^2 .$$

The time scaling property of the Brownian Sheet then shows that

$$N = \sigma^2(\tau) \left(\sup_{0\le s\le 1} \sup_{0\le t\le 1} \left| \tilde{\mathbb{S}}(s,t) \right| \right)^2 = \sigma^2 \|\tilde{\mathbb{S}}\|^2_{[0,1]^2}$$

where $\tilde{\mathbb{S}}$ is a Brownian Sheet on $[0,1]^2$. □

This leads directly to the following result concerning the Nelson–Aalen estimator. Its proof follows as a direct consequence of Theorem 2 from the well-known fact that the Nelson–Aalen estimator is composed of H-differentiable maps (van der Vaart & Wellner, 1996, Example 3.9.19) and has a Gaussian Martingale limit. We also note that a completely analogous corollary is also valid for the Kaplan–Meier estimator (see example 3.9.31 of van der Vaart & Wellner (1996) and Theorem IV.3.2 of Andersen et al. (1992)).

Corollary 1. *Let N_ε be the last time the Nelson–Aalen estimator $\hat{\Lambda}_n$ is more than ε away from Λ with respect to supremum distance and let*

$$\sigma^2(t) = \int_{[0,t]} \frac{1 - \Delta\Lambda(z)}{P\{Z \ge z\}} d\Lambda(z).$$

Then

$$\varepsilon^2 N_\varepsilon \xrightarrow[\varepsilon\to 0^+]{\mathcal{W}} \sigma^2 \left(\sup_{0\le s\le 1} \sup_{0\le t\le 1} |\mathbb{S}(s,t)| \right)^2 \tag{21}$$

for a Brownian Sheet $\mathbb{S}$ on $[0,1]^2$ and where $\sigma^2 = \sigma^2(\tau)$.

This can also be seen independently when working directly with the heuristics leading to Theorem 1 through

$$\mathbb{Y}_m(s,t) = \sqrt{m}(\hat{\Lambda}_{[ms]}(t) - \Lambda(t))$$

using martingale calculus. Using theory presented in Andersen et al. (1992), convergence of $\mathbb{Y}_m(s,t)$ to the Brownian Sheet $W(s, \sigma^2(t))$ as $m \to \infty$ can be proven. However, such a proof would use the fine structure of ϕ. In contrast, the above corollary is a trivial consequence of Theorem 2, and only rests on the well-known martingale structure of $\dot{\phi}\mathbb{Z}_s$.

In the setting of Theorem 2, we can reach tight and general bounds for the m of eq. (17). Let $b = \sqrt{\lambda_\alpha}/\sigma$ where λ_α is the upper α quantile of $\sigma^2\|S\|_{[0,1]^2}$. We have that

$$P(\|B_s\|_{[0,1]} > b) \le P(\|\mathbb{S}(s,t)\|_{[0,1]^2} > b) = \alpha \le 2P(\|B_s\|_{[0,1]} > b), \tag{22}$$

where B is Brownian motion on $[0,1]$ and where the upper bound is analogous to Proposition 1. Hence,

$$A^{-1}(\sqrt{\alpha}) \leq b \leq A^{-1}(\sqrt{\alpha}/2)$$

where

$$A(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \left[\Phi((2k+1)\lambda) - \Phi((2k-1)\lambda)\right]$$

is the cumulative distribution function of $\|B_s\|_{[0,1]}$ given in Section 2.7 of Sen (1981). As $m = [\lambda_\alpha/\varepsilon^2]$, we get that

$$\sigma^2 A^{-1}(\sqrt{\alpha})^2/\varepsilon_0^2 \leq m \leq \sigma^2 A^{-1}(\sqrt{\alpha}/2)^2/\varepsilon_0^2 + 1.$$

One may improve on this bound by approximating the distribution of $\|\mathbb{S}(s,t)\|_{[0,1]^2}$ directly instead of using eq. (22).

3.3. **An application to risk averse stochastic problems.** As discussed in Shapiro & Ruszczynski (2008), there is a rich class of applications in operations research where one encounters problems of the form

$$\min_{x \in X} g(x) \tag{23}$$

where $g(x) = \mathbb{E}G(x,\xi)$ is the expected loss of a loss-function G defined in terms on a random vector ξ which has a known distribution and is supported on a set $\Xi \subseteq \mathbb{R}^d$. Often $g(x)$ is difficult to compute, but $G(x,\xi)$ is simpler to compute, while ξ is possible to simulate. As numerical optimization of eq. (23) requires many evaluations of $g(x)$ at different values of x, a well-motivated procedure is to approximate $g(x)$ by

$$\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n} G(x,\xi_i)$$

where $\xi_1, \xi_2, \ldots, \xi_n$ are *iid* realizations of ξ. The so-called sample average approximation to the stochastic problem of eq. (23) is then

$$\min_{x \in X} \hat{g}(x). \tag{24}$$

Shapiro (2008) derives limit theorems for the sample average approximation for certain minimax stochastic problems by showing that under certain assumptions that are natural in many operation research problems, the estimator of eq. (24) is a H-differentiable functional of the empirical distribution. Under uniqueness assumptions on the optimization problem, the functional delta method then shows that $\sqrt{n}(v_n - v)$ is asymptotically normal, where $v_n = \min_{x\in X} \hat{g}(x)$ and $v = \min_{x \in X} g(x)$. For concreteness, let us work with the following risk averse stochastic problem, given by

$$\min_{x \in X} \rho_\lambda \left[G(x,\xi)\right]$$

where $G : \mathbb{R}^m \times \Xi$ and $\rho_\lambda(Z) := \mathbb{E}Z + \lambda\mathbb{E}[Z - \mathbb{E}Z]_+$ is the so-called absolute semideviation risk measure with $\lambda \in [0, 1]$. A most fundamental problem for using sample average approximations is how to choose n. First of all, one needs to guarantee that approximating $g(x)$ by $\hat{g}(x)$ does not distort the minimum value too much. Secondly, one needs to make sure that the size of n that guarantees such a sufficient precision level is not so large as to exceed the computational burden of working work directly with $g(x)$. Through assuming an exponential bound of the moment generating function of ξ, Shapiro (2008) provides a formula for $n(\alpha, \varepsilon)$ such that for a given $\alpha > 0$,

$$P(|\hat{v}_{n(\alpha,\varepsilon)} - v| < \varepsilon) \geq 1 - \alpha \tag{25}$$

where

$$n(\alpha, \varepsilon) = \frac{C_1}{\varepsilon^2}\left(\log\frac{C_2}{\varepsilon} + \log\alpha^{-1}\right) \tag{26}$$

for constants C_1, C_2 depending on G, X and the distribution of ξ only. Without assuming exponential bounds for the moment generating function of ξ, Theorem 1 identifies the limit distribution of $\varepsilon^2 N_\varepsilon = \varepsilon^2 \sup\{n : |v_n - v| > \varepsilon\}$. Assuming the uniqueness conditions stated in Shapiro (2008), v_n is asymptotically Gaussian, so that Remark 1 and the computations of Section 3.2 shows that

$$n \geq N(\alpha, \varepsilon) := \sigma^2 A^{-1}(\sqrt{\alpha}/2)^2/\varepsilon^2 \tag{27}$$

implies that

$$P(|\hat{v}_m - v| < \varepsilon \text{ for all } m \geq n) \tag{28}$$

is close to $1 - \alpha$ for sufficiently small ε. Here σ^2 is the asymptotic variance of $\sqrt{n}(v_n - v)$ which is given in Equation 3.11 of Shapiro (2008) as

$$\sigma^2 = \mathrm{Var}\left\{G(x^*, \xi) + \lambda\alpha^*\left[G(x^*, \xi, -\mathbb{E}G(x^*, \xi)\right]_+ + \lambda(1 - \alpha^*)\left[-G(x^*, \xi) - \mathbb{E}G(x^*, \xi)\right]_+\right\}$$

defined in terms of

$$x^* = \operatorname*{argmin}_{x \in X} \rho_\lambda\left[G(x, \xi)\right], \qquad \alpha^* = P(G(x^*, \xi) \leq \mathbb{E}G(x^*, \xi)).$$

This result is valid under much less stringent assumptions than that of Shapiro (2008), but is asymptotic in contrast to the finite sample bound of $n(\alpha, \varepsilon)$ in eq. (26). It is interesting to note that $n(\alpha, \varepsilon)$ is larger than $N(\alpha, \varepsilon)$ by a factor of $\log\varepsilon^{-1}$. This seems to originate from the coarseness of the exponential inequalities used in Shapiro (2008).

4. Further applications

This section surveys other statistically motivated applications of Theorem 1.

4.1. **The multivariate case.** Although we have suppressed it from our notation, Theorem 1 is valid also in the multivariate case. Given a norm $\|\cdot\|_{\mathbb{R}^d}$ on $\mathbb{R}^d$, such as the Euclidean or the maximum norm, we can work with

$$l^\infty(\mathcal{E}) = \left\{ f \in M(\mathcal{E} \mapsto \mathbb{R}^d) : \sup_{e\in\mathcal{E}} \|f(e)\|_{\mathbb{R}^d} < \infty \right\}$$

where $M(\mathcal{E} \mapsto \mathbb{R}^d)$ is the space of all functions from $\mathcal{E}$ to $\mathbb{R}^d$. Suppose that $\hat\theta_{1,n} \xrightarrow[n\to\infty]{\text{a.s.*}} \theta_1$ and $\hat\theta_{2,n} \xrightarrow[n\to\infty]{\text{a.s.*}} \theta_2$ are two sequences of estimators pertaining to the regularity conditions of Theorem 1 and let

$$\begin{aligned} N_\varepsilon := \sup\left\{ n : \left\|\hat\theta_{1,n} - \theta_1\right\| > \varepsilon \text{ and } \left\|\hat\theta_{2,n} - \theta_2\right\| > \varepsilon \right\} \\ = \sup\left\{ n : \max\left\{ \left\|\hat\theta_{1,n} - \theta_1\right\|, \left\|\hat\theta_{1,n} - \theta_1\right\| \right\} > \varepsilon \right\} \end{aligned}$$

be the last time an error larger than ε is committed both for $\hat\theta_{1,n}$ and $\hat\theta_{2,n}$. As the map $F \mapsto (F, F)$ is linear and hence trivially H-differentiable, the chain-rule of H-differentiability and Theorem 1 show that

$$\varepsilon^2 N_\varepsilon \xrightarrow[\varepsilon\to 0^+]{\mathscr{W}*} \sup_{(i,s,e)\in\{1,2\}\times(0,1]\times\mathcal{E}} |\mathbb{Z}_{i,s}(e)|^2 = \|\mathbb{Z}_s(e)\|^2_{(0,1]\times\mathcal{E}}$$

for a vector-valued Kiefer-Müller process $\mathbb{Z}_s = (\mathbb{Z}_{1,s}, \mathbb{Z}_{2,s})$. Note that $\mathbb{Z}_{1,s}$ and $\mathbb{Z}_{2,s}$ are independent if $\sqrt{n}(\hat\theta_{1,n} - \theta_1)$ is asymptotically independent of $\sqrt{n}(\hat\theta_{2,n} - \theta_2)$.

4.2. **The number of ε-misses and two new variables.** So far we have only worked with the variable N_ε. However, weak convergence of several statistically interpretable variables also follow from Lemma 1.

Corollary 2. *Let*

$$Q_\varepsilon = \sum_{n=1}^\infty I\{\|\phi(P_n) - \phi(P)\| \geq \varepsilon\}$$

be the number of errors larger than ε. Further let

$$R_\varepsilon(a,b) = \frac{\sum_{n=1}^\infty I\{a\varepsilon \leq \|\phi(P_n) - \phi(P)\| \leq b\varepsilon\}}{\sum_{n=1}^\infty I\{\|\phi(P_n) - \phi(P)\| \geq \varepsilon\}}$$

be the ratio of errors of sizes contained in $[a\varepsilon, b\varepsilon]$ relative to all errors larger than ε and

$$M_\varepsilon = \frac{\sum_{n=1}^\infty \|\phi(P_n) - \phi(P)\| I\{\|\phi(P_n) - \phi(P)\| \geq \varepsilon\}}{\sum_{n=1}^\infty I\{\|\phi(P_n) - \phi(P)\|_\mathcal{E} \geq \varepsilon\}},$$

the mean size of errors larger than ε. We then have that

$$\varepsilon^2 Q_\varepsilon \xrightarrow[\varepsilon\to 0^+]{\mathscr{W}} \int_0^\infty I\left\{ \|\dot\phi \mathbb{Z}_s\|_\mathcal{E} \geq 1 \right\} \mathrm{d}s.$$

Denoting the limit variable of $\varepsilon^2 Q_\varepsilon$ by Q, we further have

$$R_\varepsilon(a,b) \xrightarrow[\varepsilon\to 0^+]{\mathscr{W}} Q^{-1} \int_0^\infty I\left\{ a \leq \|\dot\phi \mathbb{Z}_s\|_\mathcal{E} \leq b \right\} \mathrm{d}s,$$

which we will call $R(a, b)$. Finally, we also have

$$\varepsilon^{-1} M_\varepsilon \xrightarrow[\varepsilon \to 0^+]{\mathscr{W}} Q^{-1} \int_0^\infty \|\dot{\phi}\mathbb{Z}_s\|_{\mathcal{E}} I\left\{\|\dot{\phi}\mathbb{Z}_s\|_{\mathcal{E}} \geq 1\right\} \mathrm{d}s.$$

Proof. We will only consider Q_ε, as the other cases follow similarly. Let us first show that for

$$Q_\varepsilon(l) = \sum_{n=[l/\varepsilon^2]}^{\infty} I\{\|\phi(P_n) - \phi(P)\| \geq \varepsilon\}$$

we have

$$\varepsilon^2 Q_\varepsilon(l) \xrightarrow[\varepsilon \to 0^+]{\mathscr{W}} \int_l^\infty I\left\{\|\dot{\phi}\mathbb{Z}_s\|_{\mathcal{E}} \geq 1\right\} \mathrm{d}s$$

each $l > 0$ and we afterwards let $l \to 0^+$. Indeed, as

$$\sum_{n=[l/\varepsilon^2]}^{\infty} I\{\|\phi(P_n) - \phi(P)\| \geq \varepsilon\} = \int_{[l/\varepsilon^2]}^{\infty} I\{\|\phi(P_{[s]n}) - \phi(P)\| \geq \varepsilon\} \,\mathrm{d}s$$

a change of variables gives

$$\varepsilon^2 Q_\varepsilon(l) = \int_l^\infty I\{\sqrt{m}\|\phi(P_{[ms]}) - \phi(P)\| \geq 1\} \,\mathrm{d}s + o_{P^*}(1) = Q_l(\mathbb{X}_n) + o_{P^*}(1),$$

where Q_l is the mapping

$$D \mapsto \int_l^\infty I\{\sup_{f \in \mathcal{F}} |D_s(f)| \geq 1\} \,\mathrm{d}s.$$

As Q_l is a continuous mapping in $l^\infty([l, \infty) \times \mathcal{E})$, the claimed limit follows from the continuous mapping Theorem and a trivial extension of Lemma 1 to prove convergence on $l^\infty([l, \infty) \times \mathcal{E})$ (when $l > 0$) instead of $l^\infty([1, \infty) \times \mathcal{E})$. The full convergence follows if we show that for each $\delta > 0$ we have

$$\lim_{c \to \infty} \limsup_{n \to \infty} P^*\left(\sup_{l \leq 1/c} |D_l(\mathbb{X}_n) - D_0(\mathbb{X}_n)| \geq \delta\right) = 0.$$

The linearity of the integral and subadditivity of outer measures implies that

$$\begin{aligned} P^*\left(\sup_{l \leq 1/c} |Q_l(\mathbb{X}_n) - Q_0(\mathbb{X}_n)| \geq \delta\right) &= P^*\left(\int_0^{1/c} I\{\sqrt{n}\|\phi(P_{[ns]}) - \phi(P)\| \geq 1\} \,\mathrm{d}s \geq \delta\right) \\ &\leq P^*\left(c^{-1} I\{\sup_{0 < s \leq 1/c} \sqrt{n}\|\phi(P_{[ns]}) - \phi(P)\| \geq 1\} \geq \delta\right) \\ &= P^*\left(I\{\sup_{0 < s \leq 1/c} \sqrt{n}\|\phi(P_{[ns]}) - \phi(P)\| \geq 1\} \geq c\delta\right) \end{aligned}$$

which is zero for $c\delta > 1$. □

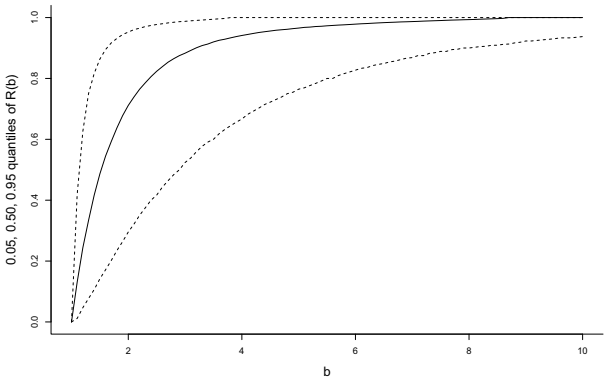


FIGURE 1. Median value and lower and upper 0.05 quantiles of the variable $R(1,b)$ (the limit of $R_\varepsilon(1,b)$) for a range of b values for the simple average.

While Hjort & Fenstad (1992) worked with Q_ε, both M_ε and R_ε are new. Note that R_ε does not require a normalization with respect to ε to gain a weak limit, and as such has a very direct interpretation. For an illustration of the R_ε result, Figure 1 displays the median value and the lower and upper 0.05 quantiles of the variable $R(1,b)$, the limit of $R_\varepsilon(1,b)$, for a range of b values (these calculations relate to the case of a one-dimensional simple average). We learn e.g. that about half of all errors ever committed above ε are below $1.53\,\varepsilon$, the rest above $1.53\,\varepsilon$. Amazingly, this fact is established even though we may never observe or even simulate the underlying $R_\varepsilon(1,b)$ variables.

4.3. **Measures of asymptotic relative efficiency.** Suppose that $\phi_1(P_n)$ and $\phi_2(P_n)$ are H-differentiable statistical functionals both estimating $\phi(P)$. A concrete example is the median versus the mean when the density of P is symmetric. Let $N_{i,\varepsilon}$ be the last time $\phi_i(P_n)$ is further than ε away from $\phi(P)$. A natural measure for the asymptotic relative efficiency of $\phi_1(P_n)$ compared to $\phi_2(P_n)$ is then

$$\text{ARE} := M_1/M_2$$

where M_i is the median of N_i, the limit variable of $\varepsilon^2 N_{i,\varepsilon}$ as $\varepsilon \to 0^+$. Recall that $\phi_1(P_n)$ and $\phi_2(P_n)$ is implicitly dependent on which space P_n is defined. Indeed, suppose ϕ_1 and ϕ_2 are functionals of $l^\infty(\mathcal{F}_1)$ and $l^\infty(\mathcal{F}_2)$. If $\mathcal{F}_1 \neq \mathcal{F}_2$, a more natural extension of the measure of variance proposed in Remark 3 is

$$\text{ARE} := \left(\frac{M_1}{\text{Median}\|\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{F}_1}}\right) / \left(\frac{M_2}{\text{Median}\|\mathbb{Z}_s\|^2_{(0,1]\times\mathcal{F}_2}}\right) \tag{29}$$

If $\mathcal{F}_1 = \mathcal{F}_2$, the two measures agree.

These asymptotic relative efficiency measures do not distinguish between estimators with the same H-differential. To distinguish between such cases, a second order perspective is required. The $\varepsilon^2 Q_\varepsilon$-limit result of Corollary 2 may be the starting-point for providing a.r.e measures when $\varepsilon^2 N_{1,\varepsilon}$ and $\varepsilon^2 N_{1,\varepsilon}$ have the same limit. Indeed, let $Q_{i,\varepsilon}$ be the number of errors committed by $\phi_i(P_n)$ for $i = 1, 2$. As done in Hjort & Fenstad (1995) and Hjort & Khasminskii (1993) for estimators connected with averages, one can work with the asymptotic relative deficiency measure

$$\text{ARD} = \lim_{\varepsilon \to 0^+} \mathbb{E}\{Q_{1,\varepsilon} - Q_{2,\varepsilon}\},$$

which in such cases provides more detail than the a.r.e measure of eq. (29).

Acknowledgements

We are grateful to Jon A. Wellner for helpful comments which lead to the correction of an inequality in Section 3 that lead to Proposition 1 and to Alex Koning for hospitality and for discussions on Gaussian tail bounds and the paper Koning & Protasov (2003) while the first author visited the Econometric institute of Erasmus University. We would also like to thank an anonymous referee that suggested the approach of Remark 6, and comments that led to improvements of the paper.

References

Adler, R. & Brown, L. (1986). Tail behaviour for suprema of empirical processes. *The Annals of Probability* **14**, 1–30.

Andersen, P., Borgan, Ø., Gill, R. & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer Verlag.

Bahadur, R. (1967). Rates of convergence of estimates and test statistics. *The Annals of Mathematical Statistics* **38**, 303–324.

Bandyopadhyay, U., Das, R. & Biswas, A. (2003). Fixed width confidence interval of $P(X < Y)$ in partial sequential sampling scheme. *Sequential Analysis* **22**, 75–94.

Bickel, P., Klaassen, A., Ritov, Y. & Wellner, J. (1993). *Efficient and adaptive inference in semi-parametric models*. Johns Hopkins University Press, Baltimore.

De la Pena, V. & Gine, E. (1999). *Decoupling: from dependence to independence*. Springer Verlag.

Dudley, R. (1992). Frechet differentiability, p-variation and uniform Donsker classes. *The Annals of Probability* **20**, 1968–1982.

Dudley, R. (1999). *Uniform central limit theorems*. Cambridge university press.

Hjort, N. L. & Fenstad, G. (1992). On the last time and the number of times an estimator is more than ϵ from its target value. *The Annals of Statistics* **20**, 469–489.

Hjort, N. L. & Fenstad, G. (1995). Second-order asymptotics for the number of times an estimator is more than ϵ from its target value. *Journal of Statistical Planning and Inference* **48**, 261–275.

Hjort, N. L. & Khasminskii, R. Z. (1993). On the time a diffusion process spends along a line. *Stochastic Process. Appl.* **47**, 229–247.

Kao, C. (1978). On the time and the excess of linear boundary crossings of sample sums. *The Annals of Statistics* **6**, 191–199.

Koning, A. & Protasov, V. (2003). Tail behaviour of Gaussian processes with applications to the Brownian pillow. *Journal of Multivariate Analysis* **87**, 370–397.

Massart, P. (2000). About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability* , 863–884.

Revuz, D. & Yor, M. (1999). *Continuous Martingales and Brownian Motion*. Comprehensive Studies in Mathematics. Springer, 3rd ed.

Robbins, H., Siegmund, D. & Wendel, J. (1968). The limiting distribution of the last time $S_n \geq n\varepsilon$. *Proc. Nat. Acad. Sci. USA* **61**, 1228–1230.

Römisch, W. (2005). Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*. Wiley, New York, 2nd ed.

Sen, P. (1981). *Sequential nonparametrics*. Wiley New York.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

Shao, J. (2003). *Mathematical Statistics*. Springer Texts in Statistics. Springer.

Shapiro, A. (2008). Asymptotics of minimax stochastic programs. *Statistics and Probability Letters* **78**, 150–157.

Shapiro, A. & Ruszczynski, A. (2008). Lectures on stochastic programming. *Preprint, Georgia Tech* .

Shorack, G. & Wellner, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley.

Stute, W. (1983). Last passage times of M-estimators. *Scandinavian Journal of Statistics* **10**, 301–305.

Talagrand, M. (1987). The Glivenko-Cantelli problem. *The Annals of Probability* **15**, 837–870.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae* **126**, 505–563.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer.

Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway

E-mail address: steffeng@math.uio.no

Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway

E-mail address: nils@math.uio.no

II

THE COPULA INFORMATION CRITERION AND ITS IMPLICATIONS FOR THE MAXIMUM PSEUDO LIKELIHOOD ESTIMATOR

STEFFEN GRØNNEBERG

Abstract. This chapter surveys the asymptotic theory of estimation of a copula from a frequentistic perspective and presents the problems involved in frequentistic model selection among several candidate copulae when using the Maximum Pseudo Likelihood Estimator (MPLE). Frequentistic copula model selection has recently been addressed through the development of the Copula Information Criterion (CIC) – a model selection formula which extends the Maximum Likelihood based Akaike Information criterion (AIC) to the MPLE. We present the developments leading to the CIC with a focus on its implications, while deferring proofs of underlying limit theorems to the original CIC paper.

The CIC is in fact two different formulae, one for mis-specified copula models and another for correctly specified copula models, parallelling the Takeuchi Information Criterion and the Akaike Information Criterion respectively.

These formulae show that there does not exist (in a certain technical sense) an AIC formula for MPL estimation when the parametric copula has extreme behavior near the edge of the unit cube. This means that one cannot estimate first-order bias-correction terms of a desired part of the attained Kullback–Leibler divergence between the MPL estimated copula and the data generating copula in a class of copulae which has received much attention in for example econometrics. This provides a demarcation for which types of copulae it is sensible to estimate with the MPLE. Interestingly, the main motivating factor for using the MPLE is also the source of the non-existence of a general MPLE based AIC formula.

Contents

1. Introduction

Suppose n-dimensional stochastic vectors $X_1, X_2, \ldots, X_N$ are observed, which are independent of each other, and all coming from the same, unknown data generating distribution

$$F^\circ(x) = C^\circ\left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right). \tag{1}$$

We assume that F° is continuous, and we wish to model the copula C° through one, or perhaps several parametric classes. In the praxis of parametric copula modelling, there are four basic problems which are naturally met in any investigation. First, if our model is

$$f_\theta(x) = c_\theta\left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right) \prod_{i=1}^n f_i^\circ(x_i),$$

where the marginals F_i° are completely unknown, how should θ be estimated? Second, how should the parametric form of c_θ be chosen? Third, how should one select among several candidate models on the basis of observed data? And fourth, is the final model (or models) adequate?

The first problem has various solutions, where the Maximum Pseudo Likelihood Estimator (MPLE) discussed in Genest et al. (1995) is the most popular. The second problem is implicit in all multivariate model building, and much of this book is devoted solely to provide flexible solutions to this problem. The fourth problem is usually dealt with through goodness-of-fit tests which are based on the MPLE, and there exists several investigations in this area (see Genest et al. (2006)).

The development of the CIC started from noticing that the third issue has been ignored, or dealt with in an incorrect manner. Several published papers, and many practitioners, have incorrectly used the "AIC formula"

$$\mathrm{AIC}^\bullet = 2\ell_{N,\max} - 2\,\mathrm{length}(\theta) \tag{2}$$

as a model selection criterion, with $\ell_{n,\max} = \ell_n(\hat\theta)$ being the maximum pseudo likelihood, from the traditional Akaike information criterion

$$\mathrm{AIC} = 2\ell^\#_{N,\max} - 2\,\mathrm{length}(\theta),$$

where $\ell^\#_{N,\max}$ is the usual maximum likelihood for a fully parametric model. One computes this $\mathrm{AIC}^\bullet$ score for each candidate model and in the end chooses the model with highest score.

This ignores the fact that the pseudo likelihood is not a proper likelihood, and unfortunately it does not lead to a correct formula. Grønneberg & Hjort (2008) derive a proper generalization of the AIC for the MPLE and name it the Copula Information Criterion (CIC). The formula is given by

$$\boxed{\text{CIC} = 2\ell_{N,\max} - 2(\hat{p}^* + \hat{q}^* + \hat{r}^*)} \tag{3}$$

with expressions for $\hat{p}^* + \hat{q}^* + \hat{r}^*$ different from (and more complicated than) merely length(θ). These quantities even vary non-trivially with the model parameter – in clear contrast with length(θ) which is invariant to the actual value of θ.

But the story does not end here, as the CIC formula of Grønneberg & Hjort (2008) does not exist for a large class of copula families such as copulae with extreme tail dependence. This lack of existence is, however, not a deficiency of the arguments used in Grønneberg & Hjort (2008), but is an inherent limitation for the asymptotic behaviour of the MPLE. This makes model selection with the MPLE a more complex problem than the fully parametric case, and the CIC formula can only attack model selection problems concerning copulae which are sufficiently well-behaved along the edges of the unit cube. The implications of this is discussed in the conclusion of the chapter.

To understand these developments and the difficulties involved in the model selection problem for copula estimation with the MPLE, one needs to understand some fundamental issues concerning the MLE, the AIC and the MPLE. The present chapter is, in addition to the introduction and concluding remarks, divided into three parts. The first part is Section 2, which presents the MLE, the AIC and the MPLE from a perspective which naturally leads to the CIC formula. The second part of our story is Section 3, which derives the two CIC formulae. Finally, we include a brief simulation example in Section 4. Although we will omit the technical asymptotic developments needed to make the arguments rigorous, we will discuss the needed mathematical structures to such a degree that the above mentioned exploding bias correction terms can be presented without simplification.

Let us first introduce some general notation that we use throughout the chapter. Let $F_1^\circ, F_2^\circ, \ldots, F_n^\circ$ be the marginal distributions of F°, and let

$$F_\perp^\circ(x) := (F_1^\circ(x_1), F_2^\circ(x_2), \ldots, F_n^\circ(x_n))$$

be the vector of marginal distributions. We will denote all sizes related to the true data generating distribution F° by circle superscripts, and all empirical estimates through replacing the circle with a hat, so that for example $\hat{F}_N$ can be seen right away to estimate F°. The assumed continuity of F° implies the existence of a unique copula C° defined implicitly through

$$F^\circ(x) = C^\circ(F_\perp^\circ(x)) \tag{4}$$

or equivalently through the more explicit

$$C^\circ(v) = F^\circ(F_\perp^{\circ\,-1}(u)) \tag{5}$$

where

$$F_\perp^{\circ\,-1}(u) = \left(F_1^{\circ\,-1}(u_1), F_2^{\circ\,-1}(u_2), \ldots, F_n^{\circ\,-1}(u_n)\right)$$

is the vector of inverse marginal distributions.

2. The developments leading to the CIC

The MPLE and the AIC both generalize the MLE, but in completely different ways. The AIC generalizes the MLE to multimodel estimation, while the MPLE generalizes the MLE to situations where the marginals are unknown. The CIC generalize both the MPLE and the AIC in that it implements the AIC-generalization of the MLE to the MPL estimator. In order to present this generalization we thus need to present the fundaments of the MLE, the AIC and the MPLE.

The MPLE sets out to estimate a copula parameter θ in a parametric model

$$f_\theta(x) = c_\theta\left(F_\perp^\circ(x)\right) \prod_{i=1}^n f_i^\circ(x_i)$$

where the marginal distributions $F_\perp^\circ$ are completely unspecified. Its precise form is defined through the following two considerations.

(1) It asymptotically minimizes the Kullback–Leibler divergence between the true data generating copula c° and a parametric copula c_θ. This generalizes the standard MLE.
(2) The estimation of the θ that minimizes Kullback–Leibler divergence between c° and c_θ is invariant to a large class of symmetries. An empirical estimate $\hat{\theta}$ should be invariant to the same symmetries.

Although the motivation for using the ML estimator to estimate a parametric model which is correctly specified is well known, its connection to the minimization of Kullback–Leibler divergence in the general case is not. This perspective naturally leads to the model selection strategy of Akaike, and Sections 2.1 and 2.2 treat these two themes. The above mentioned invariance considerations are even less well-known (it seems not to have been made explicit in any previous expositions), and we use Section 2.3 for its discussion, where we also define the MPLE precisely. Finally, Section 2.4 discusses the fact that the MPLE is not semiparametrically efficient, and argues that the concept of semiparametrically efficiency is a very different way of constructing estimators, and is often in a natural opposition to symmetry considerations. The central argument is that the MPLE is not a semiparametric estimator per se, but focuses on estimating the copula parameter θ° which is least false with respect to Kullback–Leibler divergence while respecting the related symmetry considerations. In doing so, it does provide nonparametric estimates of the vector of

marginal distributions $F_\perp^\circ$, but this infinite-dimensional part of the MPLE is merely a by-product of symmetry considerations.

2.1. **The fully parametric MLE.** Let us quickly review how the MLE is justified when we refuse to make the assumption of having the true data generating distribution f° contained in the parametric model to be fitted. For more details, with a model selection perspective in mind, see Claeskens & Hjort (2008). Suppose (for the moment) that we wish to fit a *fully parametric* density

$$f_{\theta,\gamma}(x) = c_\theta(F_{1,\gamma(1)}(x_1), \ldots, F_{n,\gamma(d)}(x_n) \prod_{i=1}^n f_{i,\gamma(i)}^\circ(x_i)$$

to observed data $X_1, \ldots, X_N \sim F^\circ$. The MLE paradigm tries to estimate

$$(\theta_{\mathrm{ML}}^\circ, \gamma_{\mathrm{ML}}^\circ) = \operatorname*{argmax}_{\theta,\gamma} \int \log f_{\theta,\gamma} \, \mathrm{d}F^\circ \tag{6}$$

from empirical data through replacing the unknown F° with the known multivariate empirical distribution $\hat{F}_n$ defined by

$$\hat{F}_N(x) := \frac{1}{N} \sum_{i=1}^N \prod_{j=1}^n I\{X_{j,i} \le x_j\} = \frac{1}{N} \sum_{i=1}^N I\{X_i \le x\}.$$

Recall that $\int \log f_{\theta,\gamma} \, \mathrm{d}F^\circ$ is a so-called multivariate Lebesgue–Stieltjes integral, and is just another way of writing $\mathbb{E} \log f_{\theta,\gamma}(X)$. We will use this notation throughout the chapter, as it leads to a very simple and rather general principle that often gives consistent empirical estimators for many quantities of interest through replacing "the circle with a hat" in F° and $\hat{F}_N$. The Lebesgue–Stieltjes integral has certain continuity properties, so that under quite general conditions "uniform (strong) consistency" of $\hat{F}_N$, meaning that

$$\lim_{N\to\infty} \sup_{x \in \mathbb{R}^n} |\hat{F}_N(x) - F^\circ(x)| = 0 \text{ almost surely},$$

implies that for each θ we have

$$\lim_{N\to\infty} \int \log f_{\theta,\gamma} \, \mathrm{d}\hat{F}_N = \int \log f_{\theta,\gamma} \, \mathrm{d}F^\circ \text{ almost surely}. \tag{7}$$

This is close to showing that the plug-in step of "putting a hat on" F° works in the sense that $(\hat{\theta}, \hat{\gamma}) \xrightarrow[n\to\infty]{\text{a.s.}} (\theta^\circ, \gamma^\circ)$. For $\hat{F}_N$, we have

$$\int \log f_{\theta,\gamma} \, \mathrm{d}\hat{F}_N = \frac{1}{N} \sum_{i=1}^N \log f_{\theta,\gamma}(X_i),$$

so eq. (7) is just a another way of stating the strong law of large numbers. But this perspective will give us a simple way of making the consistency of the MPLE plausible. For the standard MLE, the "plug-in" step takes us from

$$(\theta_{\mathrm{ML}}^\circ, \gamma_{\mathrm{ML}}^\circ) = \operatorname*{argmax}_{\theta,\gamma} \int \log f_{\theta,\gamma} \, \mathrm{d}F^\circ$$

to the empirical estimate

$$(\hat{\theta}_{\mathrm{ML}}, \hat{\gamma}_{\mathrm{ML}}) = \operatorname*{argmax}_{\theta,\gamma} \int \log f_{\theta,\gamma} \, \mathrm{d}\hat{F}_N,$$

which is also the standard definition of the MLE.

The ML-estimator was originally motivated by assuming that $f^\circ = f_{\theta^\circ_{\mathrm{ML}},\gamma^\circ_{\mathrm{ML}}}$ and then proceeding to find the estimator which asymptotically has the least variance for the true parameter. In spite of this motivation, the MLE can be calculated even when f° is not assumed to be expressible through $f_{\theta,\gamma}$ and the above consistency result is valid no matter what the true density f° is. Hence, the maximum likelihood estimator will consistently maximize $\int \log f_{\theta,\gamma} \, \mathrm{d}F^\circ$. We now show that the parameter configuration which maximize $\int \log f_{\theta,\gamma} \, \mathrm{d}F^\circ$ is a "least false" parameter in the following sense.

The relative entropy ("Kullback–Leibler divergence") between f° and $f_{\theta,\gamma}$ is

$$\mathrm{KL}(f^\circ, f_{\theta,\gamma}) = \int f^\circ \log \frac{f^\circ}{f_{\theta,\gamma}} \, \mathrm{d}x = \int f^\circ \log f^\circ \, \mathrm{d}x - \int f^\circ \log f_{\theta,\gamma} \, \mathrm{d}x,$$

where the second term is recognized from eq. (6). As the first term in the above display does not vary with (θ, γ), we have

$$\operatorname*{argmin}_{\theta,\gamma} \mathrm{KL}(f^\circ, f_{\theta,\gamma}) = \operatorname*{argmax}_{\theta,\gamma} \int \log f_{\theta,\gamma} \, \mathrm{d}F^\circ = (\theta^\circ_{\mathrm{ML}}, \gamma^\circ_{\mathrm{ML}}),$$

so that finding the maximum likelihood estimate will asymptotically reach the parameter $(\theta^\circ, \gamma^\circ)$ which minimize the Kullback–Leibler divergence between f° and $f_{\theta,\gamma}$. We call $(\theta^\circ, \gamma^\circ)$ the least false parameter (with respect to Kullback–Leibler divergence).

Kullback–Leibler divergence $\mathrm{KL}(f, g)$ is zero if and only if $f = g$ almost surely with respect to the Lebesgue measure, which means that we can use Kullback–Leibler divergence to distinguish between two densities. This property is the absolute minimal assumption needed to provide motivation to minimize $\mathrm{KL}(f^\circ, f_{\theta,\gamma})$ with respect to the parameter sets. There are also deeper motivations for using precisely Kullback–Leibler divergence, and not just any other function which is zero if and only if $f = g$ almost surely, as it is connected with the mathematical concept of information and entropy. See Claeskens & Hjort (2008) for a general discussion.

2.2. **Kullback–Leibler divergence and model selection.** Maximizing the likelihood function asymptotically reaches the parameter configuration that minimizes the Kullback–Leibler divergence between f° and $f_{\theta,\gamma}$. In the presence of several competing parametric models

$$f_{1,\alpha(1)}, \ldots, f_{K,\alpha(K)},$$

it is natural to define the best model as the model which minimizes Kullback–Leibler divergence to the truth. Let

$$\alpha(k)^\circ = \operatorname*{argmin}_{\alpha(k)} \mathrm{KL}(f^\circ, f_{k,\alpha(k)})$$

denote the least false parameter configuration when constrained to the k'th parametric class, so that the parametric model with the index

$$k^\circ = \operatorname*{argmin}_{1\leq k\leq K} \mathrm{KL}(f^\circ, f_{k,\alpha(k)^\circ})$$

is the best (in the Kullback–Leibler sense) model *among the ones we are presently considering* – i.e., the global minimizer of Kullback–Leibler divergence in the space of all parameter configurations possible among all considered models. As k° only depends on the data generating distribution F° through a multivariate Lebesgue–Stieltjes integral, the plug-in principle suggests estimating k° with

$$\tilde{k}_N = \operatorname*{argmax}_{1\leq k\leq K} \int \log f_{\hat{\alpha}(k)} \,\mathrm{d}\hat{F}_N$$

where

$$\hat{\alpha}_N(k) = \operatorname*{argmax}_{\alpha(k)} \int \log f_{k,\alpha(k)} \,\mathrm{d}\hat{F}_N.$$

This is the *main conceptual step in developing the Akaike Information Criterion*, and the precise AIC formula is simply refinements of this observation. Although $\tilde{k}_N$ is a consistent estimator, it has non-negligible bias (in a sense to be made precise) for small[1] N. The above definition of $\tilde{k}_N$ simply defines the estimated best model as the one with the highest log-likelihood at the maximum likelihood estimate, and the standard AIC formula derives first order bias-corrections *in a rather specific way*. A Taylor expansion together with well-known asymptotic likelihood theory show that

$$\int \log f_{k,\hat{\alpha}(k)} \,\mathrm{d}\hat{F}_N - \int \log f_{k,\hat{\alpha}(k)} \,\mathrm{d}F^\circ = \bar{Z}_N + \frac{1}{N} p_N(k) + o_p(N^{-1})$$

in which $\mathbb{E}\bar{Z}_N = 0$ while $p_N(k)$ converges in distribution to a $p(k)$ with expectation $p^*(k)$. Asymptotic likelihood theory provides an expression for the expectation of $p(k)$, and so we can estimate its expectation. This leads to a first order bias correction term of

$$\int \log f_{k,\hat{\alpha}(k)} \,\mathrm{d}\hat{F}_N,$$

in which it is crucial to notice that this expression is defined in terms of $\hat{\alpha}(k)$, the empirical estimate which is potentially being used, and not $\alpha^\circ(k)$, the least false parameter configuration which is unknown. If we work under the assumption that f°

[1]First order bias correction terms are insignificant for large N, and so if N is sufficiently large, the estimator $\tilde{k}_N$ yields a sensible model selection strategy.

is in the parametric class under consideration, we get the rather amazing conclusion that $p^*(k) = \text{length}(\alpha(k))$, giving the famous AIC strategy

$$\hat{k}_N^{\text{AIC}} = \underset{1\leq k\leq K}{\text{argmax}} \left[\int \log f_{k,\hat{\alpha}(k)} \, \mathrm{d}\hat{F}_N - \frac{1}{N} \text{length}(\alpha(k)) \right]$$

requiring *no empirical estimation of the bias-correction term*. For this strategy to be conceptually and formally consistent, we need to assume nested models. If this assumption cannot be justified, one can use the Takeuchi Information Criterion, which uses plug-in estimators of $p^*(k)$, and hence is of higher variability. See Claeskens & Hjort (2008) for a more detailed discussion. We will define the development of first order bias correction terms as *the AIC-programme*, and it is this we will carry out to conclude with the Copula Information Criterion. We stress the importance of the $o_p(N^{-1})$ term, and note that it is the N^{-1} which defines to what resolution we need to provide bias corrections if we are to implement the above "AIC programme".

A feature of the AIC formula is that it works with the expectation of $p(k)$, the weak limit of $p_N(k)$. This is perhaps first and foremost motivated through mathematical convenience as there is no general expression for $\mathbb{E}p_N(k)$. However, a more subtle point is that $\mathbb{E}p_N(k)$ can be infinite for even simple models such as the binomial model (Chapter 2 of Claeskens & Hjort (2008)). The AIC formula solves this potential explosion (that is, the non-existence of expectations) through going to the limit, and there everything works out nicely. For the CIC case, which transfers the above derivations to parameter estimates based on the MPLE and not the MLE, we get an additional bias correction term r_n which has the unfortunate feature that $\mathbb{E}r_N$ is finite only if the expectation of the limit variable of r_n has finite expectation. Thus, going to the limit does not help. Several common copulae models have an exploding $\mathbb{E}r_N$, leading to non-existing bias-correction terms *with respect to the above defined AIC programme*.

2.3. **The MPLE, the empirical copula and invariance considerations.** We would like to fit a parametric copula c_θ without specifying the marginal distributions. So we work under the assumption that observed data have a parametric distribution given by

$$f_\theta(x) = c_\theta\left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right) \prod_{j=1}^n f_j^\circ(x_j).$$

If the parametric form of the copula includes the correct copula c°, we wish to find the true parameter value. Otherwise, we wish to find the θ which minimizes Kullback–Leibler divergence between f_θ and the true density

$$f^\circ(x) = c^\circ\left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right) \prod_{j=1}^n f_k^\circ(x_j).$$

That is, the loss function we wish to minimize is $d(\theta) = \mathrm{KL}(f^\circ, f_\theta)$, where the minimum will be zero if and only if the model is correctly specified. Notice that we do not focus on estimating the marginals f_i°, but only on finding the least false copula inside the parametric class under consideration.

In many cases, the nonspecification of the marginals comes from lack of a priori knowledge of parametric forms for the marginals. If this is the case, the above posed estimation problem has important symmetry properties, which motivates the use of the MPLE from equivariance considerations of classical point estimation theory, as described e.g. in Lehmann & Casella (1998). First, the copula of any stochastic vector is left invariant to any (not necessarily linear) change in scale for the data. More precisely, assume that a stochastic vector X has distribution function $C^\circ(F_\perp^\circ)$. The copula C° of X is then invariant to the whole class of functions

$$\mathcal{S} := \big\{ H : \mathbb{R}^n \mapsto \mathbb{R}^n : H(x_1, \ldots, x_n) = (H_1(x_1), H_2(x_2), \ldots, H_n(x_n)), \\ \text{and each } H_i \text{ is monotonously increasing} \big\}$$

in the sense that for an $H \in \mathcal{S}$, the random vector $H(X)$ also has the copula C°. To see this, notice that the marginal distributions of $H(X)$ are given by $F_{H_i(X_i)}(v) = P\{H_i(X_i) \leq v\} = P\{X_i \leq H_i^{-1}(v)\}$, and so $F_{H(X),\perp}(x) = F_\perp(H^{-1}(v))$. Thus,

$$F_{H(X),\perp}(x)(H(X)) = F_\perp \circ H^{-1} \circ H(X) = F_\perp(X) \sim C^\circ,$$

which demonstrates the invariance. As the copula C° is completely unaffected under $\mathcal{S}$-transformations, this invariance will be shared by any parametric copula family c_θ. This should also be intuitively clear, as the copula represents the dependency structure of X, and each H in $\mathcal{S}$ merely changes the scale of each coordinate. This change in scale does not transform the (intuitive notion of) dependency among the elements of X.

The loss function $d(\theta) = \mathrm{KL}(f^\circ, f_\theta)$ is also invariant to the class $\mathcal{S}$, as it in fact does not depend on the marginals $F^\circ_\perp$. To see this, notice that

$$\begin{aligned}
\mathrm{KL}(f^\circ, f_\theta) &= \int \log \frac{f^\circ}{f_\theta} \,\mathrm{d}F^\circ \\
&= \int \log c^\circ \left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right) \,\mathrm{d}F^\circ + \sum_{j=1}^n \int \log f_j^\circ(x_k) \,\mathrm{d}F^\circ \\
&\quad - \int \log c_\theta \left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right) \,\mathrm{d}F^\circ - \sum_{j=1}^n \int \log f_j^\circ(x_j) \,\mathrm{d}F^\circ \\
&= \int \log \frac{c^\circ \left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right)}{c_\theta \left(F_1^\circ(x_1), \ldots, F_n^\circ(x_n)\right)} \,\mathrm{d}F^\circ \qquad (8) \\
&= \int \log \frac{c^\circ \left(v_1, \ldots, v_n\right)}{c_\theta \left(v_1, \ldots, v_n\right)} \,\mathrm{d}C^\circ(v) \qquad (9) \\
&= \mathrm{KL}(c^\circ, c_\theta),
\end{aligned}$$

where the transition from eq. (8) to (9) applies the change of variables formula for multivariate Lebesgue–Stieltjes integrals.

This validates the principle of equivariance (see Lehmann & Casella (1998)), meaning that any estimator of $\hat{\theta}$ should be invariant to transformations of $\mathcal{S}$. It is well-known from the problem of testing independence that multivariate rank statistics are "maximally invariant" (see Lehmann & Romano (2005) for precise definitions) with respect to the transformations in $\mathcal{S}$, and so our estimator needs to be a functional of multivariate rank statistics.

Univariate ranks are equivalently represented through the marginal empirical distribution function. Analogously, multivariate ranks are equivalently represented through the empirical copula

$$\hat{C}_N(v) = \frac{1}{N} \sum_{i=1}^N \prod_{j=1}^n I\{\hat{F}_{N,j}(X_{i,j}) \leq v_j\}$$

so that any functional of the multivariate ranks is a functional of the empirical copula. Here $\hat{F}_{N,\perp}$ is the vector of marginal empirical distributions multiplied by $N/(N+1)$ to keep the observations away from the edge of the unit cube. That is,

$$\hat{F}_{N,\perp}(x) = \left(\hat{F}_{N,1}(x_1), \hat{F}_{N,2}(x_2), \ldots, \hat{F}_{N,n}(x_n)\right), \qquad (10)$$

where

$$\hat{F}_{N,j}(x_j) = \frac{1}{N+1} \sum_{i=1}^N I\{X_{i,j} \leq x_j\}.$$

When observing that the least false copula parameter θ° can be written as

$$\theta^\circ = \operatorname*{argmin}_\theta \mathrm{KL}(f^\circ, f_\theta) = \operatorname*{argmax}_\theta \int \log c_\theta \,\mathrm{d}C^\circ,$$

and when one knows that the empirical copula is a uniformly strongly consistent estimator of the data generating copula in the sense that

$$\sup_v |\hat{C}_N(v) - C^\circ(v)| \xrightarrow[n\to\infty]{\text{a.s.}} 0, \tag{11}$$

a very natural estimator of θ° is the MPLE given by

$$\hat{\theta} = \operatorname*{argmax}_\theta \int \log c_\theta \, \mathrm{d}\hat{C}_N = \operatorname*{argmax}_\theta \frac{1}{N} \sum_{i=1}^{N} \log c_\theta \left(\hat{F}_{N,\perp}(X_i) \right).$$

2.4. **What about semiparametric efficiency?** It is well known that the MPLE is not universally semiparametrically efficient in the sense of e.g. Bickel et al. (1993). In the context of model selection of semiparametric copula models, it can be argued that this lack of semiparametric efficiency is not a serious deficiency. The semiparametric efficiency concept is defined for models that include the true data generating distribution, which is certainly not the case in any investigation where non-nested model selection is needed.

Although there does exist a semiparametric copula estimation routine which is universally semiparametrically efficient (given in Chen et al. (2006)), it does not respect the symmetry considerations leading to the MPLE. While the Chen et al. (2006) method is well-motivated only when the parametric copula model includes the data generating copula, the symmetry considerations motivating the MPLE are valid no matter what copula is the data-generating one. Although it would be desirable that the MPLE is semiparametrically efficient, this is not the problem the MPLE sets out to solve. There should be no surprise if estimators derived from equivariance considerations, and that happen to be interpretable also as semiparametric estimators, are not semiparametrically efficient, as these two concepts most often represent opposing interests.

2.5. **Large-sample theory for the MPLE.** In Section 2.2, we saw that the large-sample theory of the MLE was needed to derive bias corrections that motivated the AIC formula. This section will state the large-sample results which form a basis for the CIC. The results are justified in Genest et al. (1995); Tsukahara (2005); Chen & Fan (2005), and we state them without further justification.

Recall the definition of $\hat{F}_{N,\perp}$ in eq. (10) and define

$$\ell_N(\theta) = \sum_{i=1}^{N} \log c_\theta \left(\hat{F}_{N,\perp}(X_i) \right)$$

as the "pseudo likelihood" function. Let

$$\hat{A}_N(\theta) = \frac{1}{N} \ell_N(\theta) = \int \log c_\theta \, \mathrm{d}\hat{C}_N$$

be the normalized pseudo likelihood function so that

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \ell_N(\theta) = \operatorname*{argmax}_{\theta} \hat{A}_N(\theta).$$

And while $\ell_N(\theta) \to \infty$, we have normalized $\hat{A}_N$ so that

$$\hat{A}_N(\theta) \xrightarrow[n\to\infty]{\text{a.s.}} \int \log c_\theta \, \mathrm{d}C^\circ =: A(\theta).$$

Classical Taylor expansion-based proofs of normality for M-estimators (estimators which optimize a criterion function) require the asymptotic distribution of the score function

$$U_N := \frac{\partial \hat{A}_N(\theta_0)}{\partial \theta}.$$

As $U_N = \int \phi(v, \theta_0) \, \mathrm{d}\hat{C}_N$, where $\phi(\cdot, \theta) = \partial/\partial\theta \log c(\cdot, \theta)$, the score function is a multivariate rank statistic, whose asymptotic behaviour is derived in Ruymgaart et al. (1972); Ruymgaart (1974). We get

$$\sqrt{N}\, U_N \xrightarrow[n\to\infty]{\mathscr{W}} U \sim N_p(0, \Sigma)$$

where Σ is somewhat inflated compared to the standard Maximum Likelihood setting.

We have

$$\Sigma = \mathcal{I} + \operatorname{Cov}\left\{ \sum_{j=1}^{n} \int_{[0,1]^n} \frac{\partial \phi(v, \theta_0)}{\partial v_j} (I\{\xi_j \le v_j\} - v_j) \, \mathrm{d}C^\circ(v) \right\}$$

in which $\mathcal{I}$ is the Information matrix $\mathcal{I} = \mathbb{E}\phi(\xi, \theta_0)\phi(\xi, \theta_0)^t$ and $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ is a random vector distributed according to C°. Note that the above covariance is taken with respect to ξ.

Regularity conditions then secure

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow[n\to\infty]{\mathscr{W}} J^{-1}U \sim N_p(0, J^{-1}\Sigma J^{-1}), \tag{12}$$

where

$$J = -A''(\theta_0) = -\int_{[0,1]^n} \frac{\partial^2 \log c_{\theta_0}(v)}{\partial\theta\partial\theta^t} \, \mathrm{d}C^\circ.$$

If $c^\circ = c_{\theta_0}$, the well known information matrix equality $J = \mathcal{I}$ is valid. This means that the limit covariance of eq. (12) is simplified to

$$J^{-1} + J^{-1} \operatorname{Cov}\left\{ \sum_{j=1}^{n} \int_{[0,1]^n} \frac{\partial \phi(v, \theta_0)}{\partial v_j} (I\{\xi_j \le v_j\} - v_j) \, \mathrm{d}C^\circ(v) \right\} J^{-1}.$$

3. Model selection with the MPLE

We are now ready to implement the AIC-programme for the MPLE parallelling the developments of Section 2.2. All proofs and technical subtleties are omitted, for which the reader can refer to Grønneberg & Hjort (2008).

Suppose we have K copulae models $c_{1,\theta(1)}, \ldots, c_{K,\theta(K)}$ and wish to choose which to use on the basis of empirical data. We assume that the MPLE is to be used in the estimation of the copula parameters. This means we define the best parameter configuration of each of the models to be the $\theta^\circ(k)$ which minimizes Kullback–Leibler divergence between c° and $c_{k,\theta(k)}$. In this perspective, there is only one natural way to extend the AIC principle to our current setting, and that is to define the best copula model to be the one with index

$$k^\circ := \operatorname*{argmin}_{1\le k\le K} \mathrm{KL}(c^\circ, c_{k,\theta^\circ(k)}).$$

As for the AIC case, we can naively use

$$\tilde{k}_N := \operatorname*{argmax}_{1\le k\le K} \int \log c_{\hat{\theta}(k)} \,\mathrm{d}\hat{C}_N, \tag{13}$$

which is consistent, but with poor small sample behaviour. We can make small-sample corrections to the estimate $\tilde{k}_N$ analogous to the AIC formula. The definition of k° as the best parametric copula model is the decisive step of the development to the CIC. The remaining steps are entirely analogous to Section 2.2, and although their validity requires some mathematical sophistication, the conceptual side of the CIC is now fully developed.

As in the development of the AIC formula, we can use a Taylor expansion together with the limit theorems of Section 2.5 to conclude that

$$\hat{A}_N(\hat{\theta}) - A(\hat{\theta}) = \bar{Z}_N + N^{-1}p_N + \hat{A}_N(\theta^\circ) - A(\theta^\circ) + o_P(N^{-1})$$

where $\mathbb{E}\bar{Z}_N = 0$ and p_N is of a known form and converges to a Gaussian distribution.

But in contrast to the developments of the standard AIC in section 2.2, this expansion is not sufficient to conclude with a model selection formula. To see this, notice that in the standard ML case with known marginals, the $\hat{A}_N(\theta^\circ) - A(\theta^\circ)$ would be included in the mean zero variable $\bar{Z}_N$, as we *would* have

$$\begin{aligned}\mathbb{E}\hat{A}_N(\theta^\circ) = \mathbb{E}\int \log c_{\theta^\circ}(v)\tilde{C}_N = \mathbb{E}\frac{1}{N}\sum_{i=1}^{N} \log c_{\theta^\circ}(F_\perp^\circ(X_i)) \\ = \int \log c_{\theta^\circ}(F_\perp^\circ(x))\,\mathrm{d}F^\circ = \int \log c_{\theta^\circ}(v)\,\mathrm{d}C^\circ = A(\theta^\circ)\end{aligned} \tag{14}$$

in which $\tilde{C}_N$ is the empirical distribution based on observations $F_\perp^\circ(X_1), \ldots, F_\perp^\circ(X_N)$. As we are interested in bias correction terms, and accordingly only focus on the mean value behaviour, we could in the classical ML case ignore both $\bar{Z}_N$ and

$\hat{A}_N(\theta^\circ) - A(\theta^\circ)$. We then only had to investigate the behaviour of p_N, and find an estimator $\hat{p}^*$ for $p^* = \mathbb{E}p$ where $p_N \xrightarrow[N\to\infty]{W} p$ to get the classical AIC formula.

In the MPLE case, we encounter the complication that

$$\mathbb{E}\hat{A}_N(\theta^\circ) = \mathbb{E}\int \log c_{\theta^\circ}(v)\tilde{C}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\log c_{\theta^\circ}(F_{N,\perp}(X_i)) \neq A(\theta^\circ),$$

in which we have the stochastic and far from trivial stochastic function $F_{N,\perp}(X_i)$ inside of c_{θ° – in contrast to the $F^\circ_\perp(X_i)$ we had in eq. (14). Remember that the AIC gives bias-corrections up to the $o_P(N^{-1})$ precision level. As we define this to be the AIC-programme, we have to take the behaviour of $\hat{F}_{N,\perp}$ into consideration to provide a genuine extension of the standard AIC. A two-term Taylor expansion of $\log c_{\theta^\circ}(\cdot)$ around $F^\circ_\perp(X_i) - \hat{F}_{N,\perp}(X_i)$ replaces the problematic $\hat{F}_{N,\perp}$ with $F^\circ_\perp$ – which we had in the standard ML case – and also quantifies the magnitude of error we are committing. This error is of the desired order $o_P(N^{-1})$. We get that

$$\begin{aligned}(15)\quad \hat{A}_N(\theta^\circ) = N^{-1}\sum_{i=1}^{N}\Big[&\log c(F^\circ_\perp(X_i),\theta^\circ) + \zeta'(F^\circ_\perp(X_i),\theta^\circ)^t(\hat{V}_i - F^\circ_\perp(X_i)) \\ &+ \frac{1}{2}(\hat{V}_i - F^\circ_\perp(X_i))^t\zeta''(F^\circ_\perp(X_i),\theta^\circ)(\hat{V}_i - F^\circ_\perp(X_i))\Big] + o_P(N^{-1})\end{aligned}$$

where

$$\zeta'(v,\theta) = \frac{\partial \log c(v,\theta)}{\partial v} \text{ and } \zeta''(v,\theta) = \frac{\partial^2 \log c(v,\theta)}{\partial v \partial v^t}$$

are the vector of derivatives and matrix of double derivatives of the log copula density respectively.

The first summation term of eq. (15) has expectation $A(\theta^\circ)$, as the ML case, but we also end up with two additional terms to deal with.

Through the use of empirical process theory, Grønneberg & Hjort (2008) concludes that

$$\hat{A}_N(\hat{\theta}) - A(\hat{\theta}) = \tilde{Z}_N + N^{-1}(p_N + q_N + r_N) + o_P(N^{-1})$$

in which $\mathbb{E}\tilde{Z}_N = 0$. Further,

$$\begin{aligned} q^*_N = \mathbb{E}q_N &\to \int_{[0,1]^n} \zeta'(v;\theta_0)^t\big(\mathbf{1} - v\big)\,\mathrm{d}C^\circ(v) \\ r^*_N = \mathbb{E}r_N &\to r^* = \mathbf{1}^t\Upsilon\mathbf{1}\end{aligned}$$

where $\Upsilon = (\Upsilon_{a,b})_{1\le a,b\le n}$ is the symmetric matrix with

$$\begin{aligned}\Upsilon_{a,a} &= \frac{1}{2}\int_{[0,1]^n} \zeta''_{a,a}(u;\theta_0)u_a(1-u_a)\,\mathrm{d}C^\circ, \\ \Upsilon_{a,b} &= \frac{1}{2}\int_{[0,1]^n} \zeta''_{a,b}(u;\theta_0)\left[C_{a,b}(u_a,u_b) - u_a v_b\right]\,\mathrm{d}C^\circ\end{aligned}$$

and $\mathbb{E}r_N$ is finite only if Υ is. Here $C_{a,b}$ is the cumulative copula of $(X_{1,a}, X_{1,b})$.

Empirical estimates of these correction terms are readily be made. We deal with correctly specified and mis-specified models separately. We construct an "AIC-like" CIC, valid under the assumptions of a correctly specified parametric copula model, and also a "TIC-like" CIC which estimates the bias-correction terms consistently even without the assumption of a correctly specified parametric copula model.

In the "AIC-like" CIC formula, simplifications can be made, and we get a formula which is visually very similar to the classical AIC formula. We get

$$\widehat{\mathrm{CIC}}_{\mathrm{AIC}} = 2\ell_{N,\max} - 2(\hat{p}^* + \hat{r}^*).$$

The estimator $\hat{p}^*$ is given by

$$\hat{p}^* = \mathrm{length}(\theta) + \mathrm{Tr}\left(\hat{\mathcal{I}}^{-1}\hat{W}\right),$$

where $\hat{\mathcal{I}}^{-1}$ and $\hat{W}$ is the empirical estimates formed through using $c_{\hat{\theta}}$ as plug-in estimates of c° in the defining formulae of $\mathcal{I}$ and W, where $\hat{\mathcal{I}}^{-1}$ is a generalized inverse of $\hat{\mathcal{I}}$. The estimator $\hat{r}^*$ is given by $\hat{r}^* = \mathbf{1}^t\hat{\Upsilon}\mathbf{1}$, defined in terms of the plug-in estimators

$$\begin{aligned}
\hat{\Upsilon}_{a,a} &= \frac{1}{2}\int_{[0,1]^n} c(v;\hat{\theta})\zeta''_{a,a}(v;\hat{\theta})v_a(1-v_a)\,\mathrm{d}v,\\
\hat{\Upsilon}_{a,b} &= \frac{1}{2}\int_{[0,1]^n} c(v;\hat{\theta})\zeta''_{a,b}(v;\hat{\theta})\left[C_{a,b}(v_a,v_b;\hat{\theta}) - v_a v_b\right]\,\mathrm{d}v
\end{aligned}$$

where $C_{a,b}(v_a, v_b;\theta)$ is the cumulative copula of (Y_a, Y_b) where $(Y_1, Y_2, \ldots, Y_d) \sim C_\theta$.

The formula for $\hat{p}^*$ is almost the same as $\hat{p}^* = \mathrm{length}(\theta)$ in the AIC formula, but with an extra term $\mathrm{Tr}\left(\hat{\mathcal{I}}^{-1}\hat{W}\right)$ which is *always positive*. However, $\hat{r}^*$ *can be both positive and negative* – depending on the estimated dependency structure of the parametric copula.

One of the main advantages of the original AIC formula compared to the TIC is that the bias-correction term is only $\mathrm{length}(\theta)$, which does not have to be estimated on the basis of observed data. The "AIC-like" CIC does not have this advantage and we need to estimate high-order cumulants to apply it. An interpretation of the terms in the "AIC-like" CIC formula is that $\mathrm{Tr}\left(\hat{\mathcal{I}}^{-1}\hat{W}\right)$ takes into consideration the inflated (compared to the standard ML) covariance matrix of the asymptotic limit of the score function, while $\hat{r}^*$ stabilize the effects of using nonparametric marginal estimates $\hat{F}_{N,\perp}$ instead of the correct $F^\circ_\perp$.

If we do not assume a correctly specified model, we get the more complicated and more general "TIC-like" CIC formula

$$\widehat{\mathrm{CIC}}_{\mathrm{TIC}} = 2\ell_{N,\max} - 2(\hat{p}^* + \hat{q}^* + \hat{r}^*),$$

which is always valid. We use

$$\hat{p}^* = \operatorname{Tr}\left(\hat{J}^- \hat{\Sigma}\right), \qquad \hat{q}^* = \int_{[0,1]^n} \zeta'(v;\hat{\theta})^t \big(\mathbf{1} - v\big)\, \mathrm{d}\hat{C}(v),$$
$$\text{and} \qquad \hat{r}^* = \mathbf{1}^t \hat{\Upsilon} \mathbf{1}$$

where now

$$\hat{\Upsilon}_{a,a} = \frac{1}{2} \int_{[0,1]^n} \zeta''_{a,a}(v;\hat{\theta}) v_a (1 - v_a)\, \mathrm{d}\hat{C}_N,$$
$$\hat{\Upsilon}_{a,b} = \frac{1}{2} \int_{[0,1]^n} \zeta''_{a,b}(v;\hat{\theta}) \left[\hat{C}_{N,a,b}(v_a, v_b) - v_a v_b\right] \mathrm{d}\hat{C}_N$$

where $C_{N,a,b}$ is the empirical copula based on $(X_{1,a}, X_{1,b}), (X_{2,a}, X_{2,b}), \ldots, (X_{N,a}, X_{N,b})$. We use the standard empirical estimates of $\hat{J}^-$ and $\hat{\Sigma}$ given in e.g. Chen & Fan (2005), where $\hat{J}^-$ is a generalized inverse of $\hat{J}$.

3.1. **Non-existence of bias correction terms and implications for the MPLE.** Many practitioners of copulae are mainly interested in the copulae which have extreme tail dependence (see Joe (1997)). However, the bias correction terms q^* and r^* is defined through the differentials of $\log c_\theta(v)$ with respect to v. These will continuously grow when extreme behaviour near the edge of the unit cube is introduced, until they explode and do not have a finite expectation. Let us agree to call parametric copula models with non-existent r^* (or q^*) "edge-extreme". The implication of these exploding terms is that empirical estimates of q^* and r^* do not exist, as it simply does not make sense to estimate anything non-existent. Hence, there cannot be any generally applicable model selection formula in the sense of providing a first order bias-correction to the model relevant part of the attained Kullback–Leibler divergence between the MPL estimated model and c°. This poses a limitation for the use of the MPLE, which is shared by all two-stage copula estimators which estimate the marginals non-parametrically, say with $\tilde{F}_{N,\perp}$, and the copula through minimizing a pseudo likelihood

$$\sum_{i=1}^{N} \log c_\theta\left(\tilde{F}_{N,\perp}(X_i)\right).$$

To see this, notice the following.

The q^* and r^* terms can be traced back to Section 3 when we observed that

$$\mathbb{E}\hat{A}_N(\theta^\circ) \neq A(\theta^\circ). \tag{16}$$

But this is actually the case for all two-stage estimators[2], such as the IFM discussed in Joe (1997). In the IFM case, we have parametric marginal estimates. Going through the same procedures as Section 3 shows that

$$\hat{A}_N(\theta^\circ) = \frac{1}{N}\sum_{i=1}^{N} \log c_\theta\left(F_{\hat{\gamma},\perp}(X_i)\right)$$

where $F_{\hat{\gamma},\perp}$ is the vector of estimated marginal cumulative distributions found through standard ML estimates. If $F_\perp^\circ = F_{\gamma^\circ,\perp}$, so that the parametric class of marginal models is correctly specified, a Taylor expansion of

$$\log c_\theta(v)\big|_{v=F_{\perp,\hat{\gamma}}(X_i)},$$

not in the full v, but for $\gamma \mapsto F_{\perp,\gamma}$ around $\hat{\gamma} - \gamma^\circ$ yields terms parallelling q^* and r^* of the CIC that always exist under classical regularity conditions for all copulae. So the problem does not come from eq. (16), rather it comes from the need to perform a Taylor-expansion around v in terms such as

$$\log c_\theta(v)\big|_{v=\hat{F}_{N,\perp}(X_i)}. \tag{17}$$

Unless empirical estimators of $F_\perp^\circ$ can be found such that $N \sup|\tilde{F}_{N,\perp} - F_\perp^\circ| = O_P(1)$, this cannot be avoided at the precision level we have defined as the "AIC-programme". And one would even then have to demand regularity conditions on the C° integrability of functions of ζ' and ζ''. This would still be confining with respect to which types of parametric copulae that could have been estimated while still having AIC-like model selection formulae.

Finally, we note that a solution which might seem promising is to utilize univariate Extreme Value Theory (EVT) to estimate the tails of the marginals. EVT gives general conditions for when the tails of univariate distributions can be approximated by Generalized Pareto distributions, and there is a well-developed machinery for finding empirical estimates for the parameters involved. As this would reduce the estimation of the functional form of the tails of the distributions to a low-dimensional problem, it would seem that a possible solution to the above problems would be to define $\hat{F}_{N,\perp}$ coordinate-wise as the standard univariate empirical distribution functions below thresholds, while using n estimated Generalized Pareto distributions above these thresholds. Such an approach for estimating the univariate distributions is discussed in McNeil & Saladin (1997), but the plug-in step of using such an $\hat{F}_{N,\perp}$ seems to be new. However, there are two problems concerning such an approach. Firstly, such EVT-estimates requires the specification of a point over threshold which is defined either algorithmically or manually. In practice this hinders a mathematical

[2]This seems to be a new observation, whose consequences have not been properly dealt with. The inequality (16) invalidates the AIC formula for all multi-stage estimation routines, and through following the derivation of the CIC it is not difficult to provide modifications of (or quantify consequences of using) the standard AIC formula in these settings.

theory of estimation based on asymptotics. Secondly, simulations show that standard *automated* routines for specifying the points over threshold and estimating the parameters of the Generalized Pareto distributions introduces so much new noise in the estimation process that the resulting copula parameter estimates are mostly inferior to the MPL estimates. These two issues show that such an EVT based solution does not seem to be fruitful.

3.2. **Philosophical implications of the CIC.** This very brief section discusses what implications the CIC formula has for the interpretation of the standard AIC formula.

The AIC formula is often seen heuristically as expressing a formalization of Occam's Razor. This interpretation is often presented as being some kind of general principle, intrinsic to the arguments underlying the AIC formula.

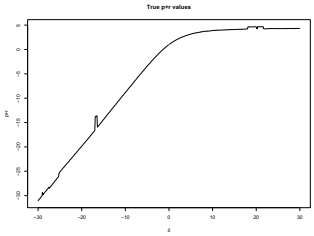
Although the $\hat{p}^*$ in the CIC formula retains the interpretation of being a "penalty for complexity", the full CIC formula has additional terms which can be *both positive and negative*, and the "penalization term" can all in all be negative. Examples of two such cases are found in Section 4. Hence the bias-correction term of the CIC no longer has the straight-forward interpretation of "penalizing for complexity", and can no longer be directly interpreted as a formalized Occam's Razor.

As the CIC is motivated through the same steps as the AIC, we see that the "penalization for complexity" interpretation of the AIC – although valid in the AIC case – is not a general principle which always follows from the underlying ideas of the AIC. The CIC seems to be the first information based model selection criterion that provides such a counterexample, hence the importance of this observation.

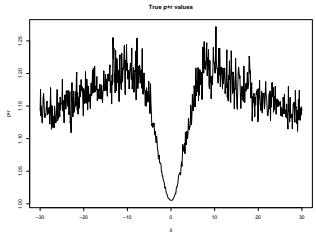
4. Illustrations

We include a brief illustration of the computational aspects of using the CIC, while confirming its validity numerically. Consider the Frank and the Plackett copulae (families B3 and B2 in Joe (1997) respectively) and denote their cumulative distribution functions by $C_{F,\delta}$ and $C_{P,\delta}$. Fig. 1 a-d shows the CIC values for the two models with varying δ. It is clear that the r^*-term dominates the CIC value, and that it reflects the degree of positive or negative dependence in the data. The random noise in the approximated p^* values is due to variation inherent in Monte-Carlo integration. Notice that for large degrees of negative dependence, both copulae give CIC formulae that are negative.

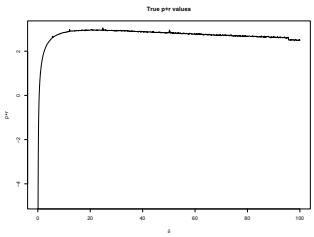
Assume $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ while the copula of (X,Y) is a copula mixture of the form $\lambda C_{F,\delta} + (1-\lambda)C_{P,\delta}$ with $\lambda = 80\%$. We want to use the known (near) unbiasedness of the AIC in the fully parametric case to illustrate that the CIC works as it should. We can do this by the following.



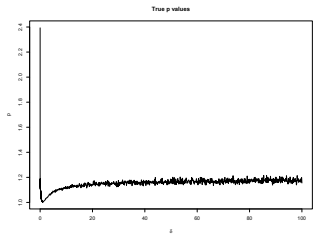
(a) Values of $p^* + r^*$ in the CIC formula for the Frank copula with varying δ.



(b) Values of p^* in the CIC formula for the Frank copula with varying δ.



(c) Values of $p^* + r^*$ in the CIC formula for the Planckett copula with varying δ.



(d) Values of p^* in the CIC formula for the Planckett copula with varying δ.

FIGURE 1. Plots of true CIC values under the assumption of a correctly specified parametric model for the Frank and Plackett copulae with varying dependence parameter.

If we restrict attention to parametric models with normal marginals and either a Frank or a Plackett copula, we have

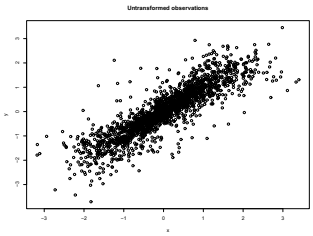
$$f_i(x, y; \delta) = c_i \left(\Phi^{-1}(x), \Phi^{-1}(y); \delta\right) \phi(x)\phi(y)$$

using the information that both marginals are known to be standard Normal and where $i \in \{F, P\}$. The true copula is known to be a mixture of the two. Denote this density by c°, and let f° be the full data-generating mechanism of (X, Y). We have

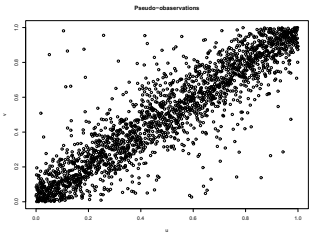
$$f^\circ(x, y) = c^\circ \left(\Phi^{-1}(x), \Phi^{-1}(y)\right) \phi(x)\phi(y).$$

This means that the Kullback–Leibler divergence between f° and $f_{i,\delta}$ is

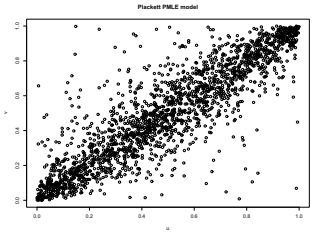
$$\mathrm{KL}(f^\circ, f_{i,\delta}) = \mathbb{E} \log \frac{f^\circ(X, Y)}{f_{i,\delta}(X, Y)} = \mathbb{E} \log \frac{c^\circ \left(\Phi^{-1}(X), \Phi^{-1}(Y)\right)}{c_i \left(\Phi^{-1}(X), \Phi^{-1}(Y); \delta\right)} = \mathrm{KL}(c^\circ, c_{i,\delta}).$$



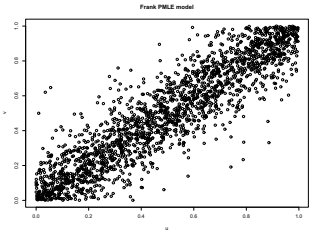
(a) Actual simulated data (X_i, Y_i)



(b) Marginal transformed observed data $(\hat{U}_i, \hat{V}_i)$.



(c) Simulated sample for the Plackett copula with dependence parameter as the MPLE of the original simulated data.



(d) Simulated sample for the Frank copula with dependence parameter as the MPLE of the original simulated data.

FIGURE 2. Plots of simulated data.

implying

$$\Delta\mathrm{KL}(f^\circ) := \mathrm{KL}(f^\circ, f_{F,\delta_F}) - \mathrm{KL}(f^\circ, f_{P,\delta_P}) = \mathrm{KL}(c^\circ, c_{F,\delta_F}) - \mathrm{KL}(c^\circ, c_{P,\delta_P}). \tag{18}$$

Consider the following three formulae.

1. The standard AIC formula $2\ell^{\#}_{N,\max} - 2\,\mathrm{length}(\delta)$ where $\ell^{\#}_{N,\max}$ is the observed maximum likelihood of the full likelihood of (X, Y) under the assumption that $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and with either a Frank or a Plackett copula specifying their simultaneous distribution. Denote the observed AIC-scores simply by AIC_F for the Frank-copula case and AIC_P for the Plackett-copula case and let $\Delta\mathrm{AIC} = \mathrm{AIC}_F - \mathrm{AIC}_P$.
2. The wrong, but typically applied AIC-like formula $2\ell_{N,\max} - 2\,\mathrm{length}(\theta)$, where $\ell_{N,\max}$ is the observed maximum pseudo-likelihood for the copula model. Denote the observed (but unjustified) AIC-scores by $\mathrm{AIC}^\bullet_F$ and $\mathrm{AIC}^\bullet_P$ and let $\Delta\mathrm{AIC}^\bullet = \mathrm{AIC}^\bullet_F - \mathrm{AIC}^\bullet_P$.

3. The CIC formula $2\ell_{N,\max} - 2(p^* + r^*)$ calculated under the assumption of a correctly specified model. Denote the observed CIC-scores by CIC_F and CIC_P and let $\Delta\mathrm{CIC} = \mathrm{CIC}_F - \mathrm{CIC}_P$.

Equation (18) shows that if the $\mathrm{AIC}^\bullet$ formula is correct, $\Delta\mathrm{AIC}^\bullet$ should be approximately equal to ΔAIC, but if the CIC-formula is correct, ΔCIC should be approximately equal to ΔAIC. A simulated sample of (X, Y) with the mixture copula is illustrated in Figure 2 e-g with $N = 2000$. It is not obvious which model is the best, as the fit of the MPLE models seems to be varying in different parts of the sample space. However, assume that we want to know which model has the least Kullback–Leibler divergence to the true model. Notice that we use the AIC-like formulae, and not the TIC-like formulae, which is an approximation typical in model selection practice, as the TIC-like formulae have a higher variability than the AIC-like formulae.

We ran 500 simulations as above – each with 2000 sample points, and for each simulation calculated the AIC, $\mathrm{AIC}^\bullet$ and CIC values. Table 1 shows that the CIC-formulae on average agrees with the fully parametric AIC value, while the mean of the incorrectly motivated $\mathrm{AIC}^\bullet$ misses the mean of AIC almost exactly by the average of $-2\Delta(p^* + r^*)$, the correction term which separates $\mathrm{AIC}^\bullet$ and CIC.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ΔAIC	−108.80	−26.73	−6.13	−5.28	16.87	84.95
ΔCIC	−122.90	−28.80	−4.65	−5.00	18.14	93.15
$\Delta\mathrm{AIC}^*$	−120.30	−26.23	−2.07	−2.43	20.72	95.72
$\Delta\mathrm{AIC} - \Delta\mathrm{CIC}$	−27.52	−7.42	−0.64	−0.28	6.51	39.26
$\Delta\mathrm{AIC} - \Delta\mathrm{AIC}^*$	−30.10	−9.99	−3.22	−2.85	3.94	36.69
MPLE δ_F	12.80	13.50	13.77	13.77	14.03	15.04
MPLE δ_P	43.06	47.05	48.74	48.71	50.12	56.13
$p_P^* + r_P^*$	2.78	2.83	2.84	2.84	2.85	2.96
$p_F^* + r_F^*$	4.00	4.04	4.06	4.06	4.08	4.13
$2\Delta(p^* + r^*)$	−2.65	−2.50	−2.44	−2.44	−2.39	−2.23

TABLE 1. Summary statistics for the simulation of 500 data-sets each consisting of 2000 samples

5. Concluding remarks

Standard semiparametric estimation theory, as summarized in Bickel et al. (1993), postulates that the true, data generating distribution is included in the space of all models spanned by the semiparametric model. The infinite-dimensional part of semiparametric models often spans such a large space that it is realistic to make this assumption. But for most practical uses of semiparametric copula models, this

is not realistic and motivates the investigation of semiparametric model selection techniques in the style of the AIC.

Standard semiparametric estimation theory is based on the assumption that the rationale for using a semiparametric model (in contrast with using a fully nonparametric model) is that the investigator possesses a priori knowledge of the correct finite dimensional part of the data generating distribution. This is often not the case in copula estimation.

The basis for the CIC investigation of Grønneberg & Hjort (2008) was to assess the consequences of using the "AIC-formula" of eq. (2). The main conclusions were

- The "penalization" for dimensionality of the copula model is only part of the story, and the correct sum of all bias correction terms can be negative.
- No proper generalization of the AIC formula exists for "edge extreme" copulae when parameters are estimated with the MPLE. The class of edge extreme copulae includes most copula models in common use.

Both of these points have practical implications for copula users. The first point has an obvious implication: Do not use the $\text{AIC}^\bullet$ formula of eq. (2) – its rationale is unjustified and its use can lead to systematic bias when selecting models. The second point has more subtle implications. It indicates that the estimation of parametric edge extreme copulae is fundamentally more complex without the knowledge of finite dimensional parametric marginals. Edge extreme copulae are often used to provide multivariate extreme value estimates such as Value At Risk calculations for the sum of dependent vectors for high quantiles. If this is the aim of the study at hand, the MPLE seems not to be the best choice.

A possible solution to the second point is to to ignore the bias-correction term which gets us in trouble, and work directly with $\tilde{k}_N$ of eq. (13). If N is sufficiently large, first order bias corrections are insignificant (see the footnote on p.7), making this a sensible model selection routine in some circumstances. This is implicitly done in Chen & Fan (2005) (although they did not notice that the "AIC formula" of eq. (2) is unjustified for the MPLE), and they provide statistical tests to assess the conclusion of the resulting model selection strategy.

Another way to address the second point is to look for alternative estimators of the copula parameter. It seems that the only well-known alternative to the MPLE is the sieve based estimator proposed in Chen et al. (2006), motivated through semiparametric efficiency considerations. But the concept of semiparametric efficiency is defined only when the model in question is correctly specified. This is clearly not the case for any investigation in which the (non-nested) model selection problem appears.

A third possible approach to the second point is to develop an analogue to the impressive machinery of Massart (2007) for the current situation. This seems currently out of reach, and would lead to a theory based on fundamentally different principles than the comparatively simple AIC formula.

If none of the candidate copula models are edge extreme, the CIC formula provides a general model selection strategy, but if at least one copula under consideration is edge extreme there are currently no fully satisfying solutions to the model selection problem. Finally, we note that model selection by cross-validation and boot-strap procedures are reasonable methods also for the MPLE. However, their theoretical properties are not yet well-understood.

Acknowledgements

This work is funded by Statistics for Innovation, (sfi)^2.

References

Bickel, P., Klaassen, A., Ritov, Y. & Wellner, J. (1993). *Efficient and adaptive inference in semi-parametric models*. Johns Hopkins University Press, Baltimore.

Chen, X. & Fan, Y. (2005). Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *The Canadian Journal of Statistics* **33**, 389–414.

Chen, X., Fan, Y. & Tsyrennikov, V. (2006). Efficient estimation of semiparametric copula models. *Journal of the American Statistical Association* **101**, 1228–1240.

Claeskens, G. & Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Genest, C., Ghoudi, K. & Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82**, 543–552.

Genest, C., Quessy, J.-F. & Rmillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transform. *Scandinavian Journal of Statistics* **33**, 337–366.

Grønneberg, S. & Hjort, N. (2008). The copula information criterion. Tech. Rep. 7, Department of Mathematics, University of Oslo.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall.

Lehmann, E. & Casella, G. (1998). *Theory of point estimation*. Springer.

Lehmann, E. & Romano, J. (2005). *Testing Statistical Hypotheses*. Springer.

Massart, P. (2007). *Concentration inequalities and model selection*. Citeseer.

McNeil, A. & Saladin, T. (1997). The peaks over thresholds method for estimating high quantiles of loss distributions. *Proceedings of 28th International ASTIN Colloquim* **28**.

Ruymgaart, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *The Annals of Statistics* **2**, 892–910.

Ruymgaart, F. H., Shorack, G. R. & van Zwet, W. R. (1972). Asymptotic normality of nonparametric tests for independence. *The Annals of Mathematical Statistics* **43**, 1122–1135.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *The Canadian Journal of Statistics* **33**, 357–375.

Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway

E-mail address: steffeng@math.uio.no

III

IV