Direct and approximately valid probabilistic inference on a class of statistical functionals

Leonardo Cella* and Ryan Martin*

June 13, 2022

Abstract

Existing frameworks for probabilistic inference assume the quantity of interest is the parameter of a posited statistical model. In machine learning applications, however, often there is no statistical model/parameter; the quantity of interest is a statistical functional, a feature of the underlying distribution. Model-based methods can only handle such problems indirectly, via marginalization from a model parameter to the real quantity of interest. Here we develop a generalized inferential model (IM) framework for direct probabilistic uncertainty quantification on the quantity of interest. In particular, we construct a data-dependent, bootstrap-based possibility measure for uncertainty quantification and inference. We then prove that this new approach provides approximately valid inference in the sense that the plausibility values assigned to hypotheses about the unknowns are asymptotically well-calibrated in a frequentist sense. Among other things, this implies that confidence regions for the underlying functional derived from our proposed IM are approximately valid. The method is shown to perform well in key examples, including quantile regression, and in a personalized medicine application.

Keywords and phrases: bootstrap; empirical risk minimizer; estimating equation; M-estimator; nonparametric; plausibility function; Z-estimator.

1 Introduction

In applications, the quantities of interest—or *inferential targets*—are often "real" in the sense that they are features of the population under investigation, known to exist and have meaning. For example, moments and quantiles are real in the sense that all distributions have, say, a 0.85-quantile. On the other hand, shape, concentration, tail-index, etc. are parameters whose meaning relies on the context provided by a suitable statistical model. Consequently, any inferences drawn about, say, a shape parameter, would be meaningless if there is no "true" shape parameter associated with the population in question. It is important to realize that these issues cannot be remedied simply by "picking a better model." Indeed, modern machine learning applications often require inference on unknowns that are defined as, say, minimizers of expected loss functions. These are

*Department of Statistics, North Carolina State University; lolivei@ncsu.edu, rgmarti3@ncsu.edu

real in the sense above and the problems are often too complex to expect that they could be treated as parameters of an interpretable statistical model. So taking a traditional statistical approach to this machine learning problem amounts to introducing a statistical model and treating the real quantity of interest indirectly through the parameters of the posited statistical model. However, as the mantra goes, *All models are wrong*, so any inference about the real quantity of interest is immediately at risk of being rendered useless by model misspecification bias. This is precisely the reason why machine learners often prefer to attack the real problem directly without considering a statistical model or using the associated statistical tools developed for model-based inference. To bridge this gap, it is important that the statistical community address the problem of direct and reliable (probabilistic) uncertainty quantification about inferential targets that are not parameters of a posited statistical model. This is the goal of the present paper.

We do not consider ourselves "anti-statistical model." There are many good reasons to work within a model-based framework, including interpretability, computational and statistical efficiency, and convenience. Indeed, within the context of a statistical model, one has a likelihood function which can be used to make likelihood-based inference. This includes maximum likelihood estimation, likelihood ratio testing, and the asymptotic efficiency properties that these methods enjoy. In terms of probabilistic inference, both the Bayesian and generalized fiducial (Fisher 1935; Hannig et al. 2016) frameworks rely heavily on the likelihood function.[1] Consequently, inference about, say, expected loss minimizers under these frameworks is necessarily *indirect* and would put users at risk of model misspecification bias. The latter criticism can be at least partially remedied by making the model "nonparametric" in the sense of, e.g., Wasserman (2006), where the model parameter is the distribution itself (or some other infinite-dimensional object). Even this can be handled in a probabilistic way using Bayesian nonparametrics (Ghosal and van der Vaart 2017; Ghosh and Ramamoorthi 2003; Hjort et al. 2010) or generalized fiducial (Cui and Hannig 2019). While this might address the issue of model misspecification bias, it does so by making any inference about a "real" inferential target even less direct—inference about an infinite-dimensional object must be made first and then an extreme marginalization to the often low-dimensional inferential target carried out. This adds complexity and negatively affects the interpretability and computational/statistical efficiency that originally motivated the model-based approach.

There is another interpretation of "nonparametric" (e.g, Conover 1971) that is more in line with our perspective here. These methods allow for inference about certain inferential targets with no/minimal model assumptions—classical examples include the sign and signed-rank tests. The advantage of these methods is that they are *direct* in the sense that they do not attempt anything more than to answer specific questions about the inferential target. Unfortunately, these classical nonparametric statistical methods are tailored to very specific problems and, to our knowledge, do not readily extend to allow for probabilistic uncertainty quantification in any general or systematic way. In this paper—an extended version of Cella and Martin (2021a)—we aim to develop a framework which is general/flexible enough to handle modern machine learning applications while simultaneously providing (imprecise) probabilistic uncertainty quantification about the inferential target with reliability guarantees (at least approximately).

[1]There are variations on the Bayesian framework that can allow for inference on real quantities of interest (see Martin and Syring 2022, and the references therein) but we will not discuss this here.

More specifically, our goal in the present paper is to develop a (generalized) inferential model, or IM, a framework for direct—and valid—probabilistic inference on statistical functionals that are not parameters of a posited statistical model. In general, an IM is mapping that takes as input the observed data, along with any other relevant information about the data-generating process, and returns as output a lower and upper probability pair to be used for quantifying uncertainty about the unknowns; see Martin and Liu (2013, 2015) for the first considerations, and Martin (2019, 2021) for a modern perspective. The need for a lower and upper probability pair, instead of just a single probability like in the Bayesian and fiducial frameworks, is to ensure that inference drawn from the IM are reliable, or *valid*. Further background on this is presented in Section 2 below. Despite the benefits of having provably valid probabilistic uncertainty quantification, the original Martin–Liu construction has a shortcoming: like the Bayesian and other fiducial-like frameworks, it relies on a statistical model to define the quantity of interest and characterize its relationship to the observable data. Therefore, direct inference on quantiles or other statistical functionals would appear to be out of reach.

To close this gap, we draw inspiration from the work presented in Martin (2015, 2018) and, more recently, in Cella and Martin (2021b, 2022), towards relaxing the requirement that a connection between observable data and quantities of interest be described via a data-generating process. The first applications of this idea focused on streamlining the IM construction, but these still rely on specification of a statistical model. Our key observation is that, by eliminating the requirement that the user start by writing out the data-generating process, we create an opportunity to construct valid probabilistic uncertainty quantification without the specification of a statistical model.

After some background on IMs and generalized IMs in Section 2, we turn to the problem of inference on statistical functional defined either as the minimizer of an expected loss or as the solution of an estimating equation (e.g., Godambe 1991; Huber 1981, 1964); see Section 3.1 for the relevant definitions. The simple quantile example mentioned above fits in this framework, as do many modern machine learning problems. Having made this connection, it is relatively simple, at least in principle, to construct a generalized IM that is exactly valid in the sense described below. The output of this generalized IM, defined in Section 3.2, takes the form of a data-dependent consonant belief/plausibility function or, equivalently, a necessity/possibility measure pair and, as consequence of the validity property, confidence regions derived from this output achieve the nominal coverage probability exactly, for all sample sizes. Unfortunately, this simple construction is impractical because it depends on aspects of the problem that would be unknown in every real-world application. To overcome this, in Section 3.3, we leverage the powerful bootstrap machinery (e.g., Efron 1979) to construct a principled approximation to the aforementioned generalized IM. With the introduction of bootstrap, exact validity cannot be achieved, but we prove, in Section 3.4, that the bootstrap-based generalized IM is approximately valid in the large-sample limit. To our knowledge, this is the first general implementation of (asymptotically) valid, prior- and model-free probabilistic uncertainty quantification. Illustrations are presented in Section 4, including, quantile regression, a classical "model-free" application. We also consider, in Section 5, an application of the proposed generalized IM approach to a relevant problem in personalized medicine, namely, *dynamic treatment regimes* (e.g., Tsiatis et al. 2020). We conclude in Section 6 with a brief summary and discussion of some open problems.

2 Background on IMs

Let Z_i, for $i = 1, \ldots, n$, denote data points taking values in a space $\mathbb{Z}$, and let $Z^n = (Z_1, \ldots, Z_n) \in \mathbb{Z}^n$. Here the space $\mathbb{Z}$ is very general, so, for example, this covers the case where $Z_i = (X_i, Y_i)$ is a predictor and response variable pair, where X_i could be high-dimensional. In this background section, we introduce a statistical model, which is a collection of probability distributions, P_ω^n, for Z^n, indexed by a parameter $\omega \in \Omega$. The key point is that the parameter ω determines *everything* about the distribution of Z^n. Consequently, if the real quantity of interest is some feature, $\theta \in \Theta$, of the Z^n distribution, then θ would be expressed as a function of ω, i.e., $\theta = \theta(\omega)$. For example, if the model is Gaussian, so that $\omega = (\mu, \sigma)$ is the mean and standard deviation pair, and if the inferential target θ is the 0.75-quantile, then $\theta = \mu + \sigma z_{0.75}$, where $z_{0.75}$ is the corresponding quantile of the standard normal distribution. Regardless of what form the mapping $\omega \to \theta$ takes, inferences about θ would be obtained by applying this mapping to inferences about ω. For example, if a confidence region for ω were available, then its image under the mapping $\omega \to \theta$ would be a corresponding confidence region for θ.

In this paper, inferences are based on data-dependent, probabilistic quantifications of uncertainty—or what Martin (2019, 2021) refers to as an *inferential model* (IM). An IM is a mapping that takes the observed data $Z^n = z^n$ and the information encoded in the statistical model to a sub-additive capacity (Choquet 1954) defined on a collection of subsets of Ω, say, the Borel σ-algebra. Specifically, a capacity γ is a set function that satisfies $\gamma(\varnothing) = 0$, $\gamma(\Omega) = 1$, and is monotone: $A \subseteq B$ implies $\gamma(A) \leq \gamma(B)$; sub-additivity requires that $\gamma(A \cup B) \leq \gamma(A) + \gamma(B)$ whenever $A \cap B = \varnothing$. Of course, probability measures are capacities, so the familiar frameworks like Bayesian, fiducial (Fisher 1935), generalized fiducial (Hannig et al. 2016), structural (Fraser 1968), and confidence distributions (Schweder and Hjort 2016; Xie and Singh 2013) are IMs in this sense. However, capacities are more general than ordinary probabilities, so an IM's output could also take the form of, say, a plausibility function (Dempster 1968, 2014; Denœux 2014; Shafer 1976), a possibility measure (Dubois and Prade 1988), or something else more complicated. For the observed data $Z^n = z^n$, relative to the posited statistical model, denote the IM's capacity by $\overline{\Pi}_{z^n}$. Define its dual/conjugate as

$$\underline{\Pi}_{z^n}(B) = 1 - \overline{\Pi}_{z^n}(B^c), \quad B \subseteq \Omega,$$

and note that sub-additivity implies $\underline{\Pi}_{z^n}(B) \leq \overline{\Pi}_{z^n}(B)$. For this reason, the IM's output can be referred to as a pair $(\underline{\Pi}_{z^n}, \overline{\Pi}_{z^n})$ of lower and upper probabilities. If the IM output is additive, not just sub-additive, then $\underline{\Pi}_{z^n}$ and $\overline{\Pi}_{z^n}$ are equal and we are back to the more familiar Bayesian or fiducial case. The motivation for non-additivity will be explained below. Observe that, through the mapping $\omega \to \theta$, assertions about θ correspond to assertions about ω, so we can quantify uncertainty about θ via marginalization, e.g.,

$$\overline{\Pi}_{z^n}(\{\omega : \theta(\omega) \in A\}), \quad A \subseteq \Theta.$$

This formalizes our above description of how inferences about ω are mapped to θ.

The interpretation of the IM output is as follows. The sets $B \subseteq \Omega$ are assertions or hypotheses about ω and $\underline{\Pi}_{z^n}(B)$ and $\overline{\Pi}_{z^n}(B)$ are lower and upper probabilities for the claim "$\omega \in B$" based on the given data $Z^n = z^n$ and the posited statistical model. If

$\underline{\Pi}_{z^n}(B)$ were large, then the data z^n strongly supports the claim; alternatively, if $\overline{\Pi}_{z^n}(B)$ is small, then the data z^n strongly contradicts the claim. For situations in between, in which $\underline{\Pi}_{z^n}(B)$ and $\overline{\Pi}_{z^n}(B)$ are relatively small and large, respectively, the data is not sufficiently informative to support or contradict the claim. In such situations, the data analyst ought to consider a "don't know" conclusion (e.g., Dempster 2008) and either collect more informative data or shift focus to a less complex assertion.

The quality of an IM is determined by the reliability of inferences drawn from it, so we are concerned with the statistical properties of the IM's output, i.e., on the properties of $\overline{\Pi}_{Z^n}$ as a function of data $Z^n \sim P_\omega^n$. We focus here on the upper probability just for brevity; all of what follows could also be described in terms of the lower probability, and we show both in our presentation of the new developments in Section 3. The basic idea is as follows. Based on the interpretation of the IM output described above, erroneous inference could be made if, for example, to an assertion B that happened to be true, the IM assigned small $\overline{\Pi}_{Z^n}(B)$. An IM would be unreliable if such erroneous inferences were not controllably rare, so the validity property is designed specifically to provide the control necessary to make its inferences reliable. This is completely in line with Fisher's logic behind his tests of significance (Fisher 1973, p. 42). More formally, an IM with output $\overline{\Pi}_{Z^n}$ is said to be *valid* if

$$\sup_{\omega \in B} P_\omega^n\{\overline{\Pi}_{Z^n}(B) \leq \alpha\} \leq \alpha, \quad \text{for all } \alpha \in [0,1] \text{ and all } B \subseteq \Omega. \tag{1}$$

In the above expression, the true ω is contained in B, so B is a "true" assertion. Then the event $\{\overline{\Pi}_{Z^n}(B) \leq \alpha\}$ is potentially problematic, especially when α is small, as it corresponds to a case where the inferences drawn could be wrong. However, the right-most inequality in (1) ensures that this potentially problematic event has an explicit and relatively small probability under the posited model. This calibration makes it possible for the IM to avoid the "unacceptable" and "systematically misleading conclusions" that Reid and Cox (2015) warn us about. The validity condition (1) can also be compared to the (slightly weaker) *fundamental frequentist principle* in Walley (2002). Of course, the validity property as stated above is relative to the posited statistical model and, therefore, if that model happens to be wrong in some sense, then property (1) is meaningless; it is precisely for this reason that we look to extend the IM construction and corresponding validity property beyond those idealized cases where there is a statistical model and it is assumed to be correctly specified.

There are a number of desirable consequences of the validity property. First, since (1) includes the "for all $B \subseteq \Omega$" clause, the validity property carries over immediately to marginal inferences on the quantity of interest θ. Second, one can readily derive statistical procedures—hypothesis tests and confidence regions—from the IM's output, and the validity property guarantees that these will control the frequentist error rates of those procedures. More details about this will be presented in Section 3.

This brings us to the motivation for considering non-additive IMs. It turns out that *IMs whose output is additive cannot be valid*. This is the so-called *false confidence theorem* of Balch et al. (2019); see, also, Martin (2019, 2022). Demonstrations of the challenges with additive IMs, especially for marginal inference about a function $\theta = \theta(\omega)$ of the full model parameter ω, can be found in Fraser (2011), Martin (2021), Cunen et al. (2020), and Martin et al. (2021), so we will not reproduce the details here. The point is, in order

to ensure validity and to enjoy its desirable consequences, it is necessary to consider genuinely non-additive IMs.

This begs the question: how to construct a valid IM? The first constructions were presented in Martin and Liu (2013, 2015), and Martin (2019) gives a detailed overview. The original formulation started by expressing the statistical model in terms of a functional relationship, $Z^n = a(\omega, U^n)$, between data Z^n, unknown parameter ω, and an unobservable auxiliary variable, say, U^n. This is effectively the same starting point as fiducial, but Martin and Liu's approach differs in how this auxiliary variable is handled. At the observed value z^n of Z^n, the expression becomes $z^n = a(\omega, u^n)$, where u^n is the unobserved value of U^n, so the question turns to how we can quantify uncertainty about the fixed, unobserved value u^n. On the one hand, a fiducial approach quantifies uncertainty about u^n using the *a priori* distribution of U^n, which leads to a z^n-dependent probability distribution for ω that does not satisfied the validity property in (1). On the other hand, Martin and Liu argue that the epistemological status of a fixed, unobserved value of a random variable is very different from a random variable and, therefore, uncertainty about u^n should be quantified with something different/more conservative than a probability distribution. Their original proposal used random sets (e.g., Molchanov 2005; Nguyen 2006) to quantify uncertainty about u^n but, more recently, it was recognized that quantifying uncertainty using possibility measures was a more direct route to a valid IM (Liu and Martin 2021). Moreover, the latter construction ensures that the IM's output is also a possibility measure, and Martin (2021) argued that valid IMs of this form are the most efficient. Therefore, we will focus here on the case where the IM output takes the form of a possibility measure. To avoid repetition, we save the details of this construction for Section 3—the novelty in the main results section of this paper is in dealing with the statistical model-free context, not the basic steps of the IM construction.

3 Direct model-free probabilistic inference

3.1 Setup

The discussion in the previous section focused on the case where a statistical model was specified, i.e., that the data $Z^n = (Z_1, \ldots, Z_n)$ had a distribution P_ω^n indexed by a parameter $\omega \in \Omega$. Suppose, however, that the model parameter, ω, is not directly of interest. Instead, the goal is inference on some feature θ of the underlying distribution. Under the posited model, $\theta = \theta(\omega)$ is a function of ω, and marginal inference can be carried out more or less as usual. The main obstacle is that forcing θ to be a function of the model parameter, ω, is potentially restrictive, since the model could be misspecified. As a somewhat extreme example, suppose the quantity of interest, θ, is the variance of the distribution of Z_1. If we model this with a Poisson distribution having rate parameter $\omega > 0$, then the usual estimator of θ would be the sample mean, which would be a poor estimate of the variance if the distribution is not Poisson.

To avoid the risk of model misspecification bias, we opt to proceed without specifying a model. That is, we assume $Z^n = (Z_1, \ldots, Z_n)$ consists of independent and identically distributed (iid) components with $Z_i \sim P$; the joint distribution of Z^n is denoted by P^n. Note that P is free to be any distribution, no constraints due to dependence on a parameter ω. In this more general case, the quantity of interest $\theta = \theta(P)$ is a functional

of the underlying distribution. And since there is no restriction on P, there is similarly no restriction on θ, hence no risk of model misspecification bias.

So far, we have said very little about what specifically the quantity of interest, θ, is. This will be important in what follows so, to end this problem-setup section, we give some further details about the origins of θ.

- Start with a loss function $\ell_\vartheta(z)$ that takes pairs $(\vartheta, z) \in \Theta \times \mathbb{Z}$ to real numbers. This loss function is a measure of the compatibility of a data point z with a generic value ϑ, with large values of $\ell_\vartheta(z)$ indicating less compatibility. Then the inferential target is defined as the minimizer of the expected loss, i.e.,

$$\theta = \arg\min_{\vartheta \in \Theta} R(\vartheta), \quad \text{where} \quad R(\vartheta) = \int \ell_\vartheta(z) \, P(dz).$$

  To estimate θ based on data z^n from P, one defines an empirical version of the risk and take $\hat{\theta}_{z^n}$ to be the corresponding minimizer, i.e.,

$$\hat{\theta}_{z^n} = \arg\min_{\vartheta \in \Theta} R_{z^n}(\vartheta), \quad \text{where} \quad R_{z^n}(\vartheta) = \frac{1}{n} \sum_{i=1}^n \ell_\vartheta(z_i).$$

  This framework is called *M-estimation*.

- Alternatively, start with a (vector-valued) function $\psi_\vartheta(z)$ and define the inferential target as the root of the expectation, i.e., θ is a solution to the (vector) equation

$$\Psi(\vartheta) = 0, \quad \text{where} \quad \Psi(\vartheta) = \int \psi_\vartheta(z) \, P(dz).$$

  As above, to estimate θ based on data z^n from P, define a corresponding empirical version of the expectation and take $\hat{\theta}_{z^n}$ to be a solution to the equation

$$\Psi_{z^n}(\vartheta) = 0, \quad \text{where} \quad \Psi_{z^n}(\vartheta) = \frac{1}{n} \sum_{i=1}^n \psi_\vartheta(z_i).$$

  The above equation is sometimes referred to as an *estimating equation*, and this general framework is called *Z-estimation*. Some authors, including Boos and Stefanski (2013, Chap. 7), do not distinguish between M- and Z-estimation.

The familiar maximum likelihood framework is a special case of M- and sometimes Z-estimation. Suppose the statistical model P_θ is indexed by θ, and that p_θ is the density function. Then $\ell_\vartheta(z) = -\log p_\vartheta(z)$ would be a loss function and the corresponding M-estimator is the maximum likelihood estimator. Similarly, if interchange of derivatives and integrals is allowed, then $\psi_\vartheta(z) = -\frac{\partial}{\partial \vartheta} \log p_\vartheta(z)$ can be used to express the maximum likelihood estimator as a Z-estimator. But there are many other M- and Z-estimators in the literature, this is just one very familiar case.

In the following two subsections, we describe this paper's proposal to provide (approximately) valid and distribution-free probabilistic inference through a so-called *generalized IM*. We present this in two steps, starting in Section 3.2 with the main idea in order to develop intuition. The real proposal and its justification comes in Section 3.3.

3.2 Generalized IM: basic idea

As discussed in Section 2 above, based on the original formulation at least, to construct a valid IM for θ we would require a functional relationship, à la Dawid and Stone (1982), that describes how to simulate data Z^n in terms of θ. In our present context, however, this is not possible because θ is not a "model parameter" that determines the distribution of Z^n, so a different approach is needed. Fortunately, the *generalized IM* construction developed in Martin (2015, 2018), which was designed to address an altogether different challenge, can be modified to suit our present needs.

First, we will need a function that will rank generic values ϑ of θ in terms of how well they align with the data $Z^n = z^n$. Denote this measure by $T_{z^n}(\vartheta)$. Throughout we will assume that ϑ having smaller values of $T_{z^n}(\vartheta)$ are higher ranked in terms of how well they align with the data z^n. Naturally, the definition of the T function would take into consideration how the functional, θ, is defined. For example, if there was a statistical model, P_θ^n, indexed by θ, then a natural choice of T would be

$$T_{z^n}(\vartheta) = -\log\bigl\{p_\vartheta^n(z^n) \,/\, p_{\hat{\theta}}^n(z^n)\bigr\},$$

where p_ϑ^n is the model's density function and $\hat{\theta}$ is the maximum likelihood estimator. A similar thing could be done using profile likelihoods if θ were a function $\theta = \theta(\omega)$ of the model parameter ω. Our focus here, however, is on the situation in which there is no statistical model, and the functional θ is defined as described at the end of the previous subsection. For the case where θ is a risk minimizer, a natural choice of T is

$$T_{z^n}(\vartheta) = R_{z^n}(\vartheta) - R_{z^n}(\hat{\theta}_{z^n}), \quad \vartheta \in \Theta.$$

Similarly, for cases where θ is a solution to an estimating equation, a natural choice is

$$T_{z^n}(\vartheta) = n\Psi_{z^n}(\vartheta)^\top \, S_{z^n}(\vartheta)^{-1} \, \Psi_{z^n}(\vartheta), \quad \vartheta \in \Theta,$$

where $S_{z^n}(\vartheta) = n^{-1}\sum_{i=1}^n \psi_\vartheta(Z_i)\,\psi_\vartheta(Z_i)^\top$ is the empirical estimate of the covariance matrix of the random vector $\Psi_{Z^n}(\vartheta)$. The intuition behind both of these choices is that, under certain regularity conditions, the distribution of $T_{Z^n}(\theta)$—as a function of $Z^n \sim P^n$ with $\theta = \theta(P)$—would, at least approximately, be free of any unknowns. This is not unlike what Wilks's theorem provides in the classical setting of likelihood ratio tests. Our proposal here does not rely on these asymptotic properties and, hence, does not require regularity conditions; see Lemma 1. However, the approximate validity result in Theorem 1 does require certain regularity, which we discuss below.

Then the basic idea behind the original generalized IM construction was to forget about establishing a functional relationship between the full data, the quantity of interest, and an unobservable auxiliary variable. Instead, just create a link between an appropriate summary, $T_{Z^n}(\theta)$, and an unobservable auxiliary variable, say, U. For problems involving a statistical model, this can be done with T the log relative likelihood summary above (Cahoon and Martin 2020, 2021); for prediction problems, this can be done with T depending on the non-conformity score used in conformal prediction (Cella and Martin 2021b, 2022). Here we propose the following association

$$T_{Z^n}(\theta) = G^{-1}(U), \quad U \sim Q = \mathsf{Unif}(0,1), \tag{2}$$

where $Z^n \sim P^n$, $\theta = \theta(P)$, and $G = G_P$ is the distribution function of $T_{Z^n}(\theta)$, which depends on P; also, "$Q = \mathsf{Unif}(0,1)$" means that Q is the probability measure corresponding to the uniform distribution on $[0,1]$. Throughout, we will assume that T is such that $T_{Z^n}(\theta)$ has an absolutely continuous distribution, so that G is strictly increasing and the inverse is well-defined. Next, if we set Z^n equal to the observed z^n, then the above equation becomes

$$T_{z^n}(\theta) = G^{-1}(u^\star), \tag{3}$$

where $u^\star$ is a fixed value, unknown to us because P and, hence, θ and G are unknown. To quantify uncertainty about the unobserved $u^\star$, we introduce a possibility measure $\overline{\Pi}$ defined on $[0,1]$. What is unique about a possibility measure is that the upper probabilities are determined by an ordinary point function, $\pi : [0,1] \to [0,1]$, where the maximum value 1 is attained, according to the formula

$$\overline{\Pi}(K) = \sup_{u \in K} \pi(u), \quad \text{for all } K \subseteq [0,1].$$

The function π is called the *possibility contour*. To achieve validity, we cannot choose just any possibility measure—it must be *consistent* with $Q = \mathsf{Unif}(0,1)$ in the sense that

$$Q(K) \leq \overline{\Pi}(K), \quad \text{for all measurable } K \subseteq [0,1], \tag{4}$$

as described in, e.g., Definition 9 of Hose and Hanss (2021). This choice ensures that Q is in the credal set corresponding to $\overline{\Pi}$. This properties differs from stochastic dominance (e.g., Denœux 2009) because the inequality holds for all events K, not just one-sided intervals. Since Q is one of the simplest probabilities distributions, (4) is relatively easy to arrange. There are several different ways this can be done, but here we insist on defining $\overline{\Pi}$ based on the contour

$$\pi(u) = 1 - u, \quad u \in [0,1].$$

That this yields a possibility measure compatible with $Q = \mathsf{Unif}(0,1)$ is easy to see:

$$\overline{\Pi}(K) = \sup_{u \in K}(1-u) = 1 - \inf K \geq \int_K du = Q(K).$$

The rationale behind this choice of π is that, since "good" values of ϑ are those that make $T_{z^n}(\vartheta)$ small, and small values of $T_{z^n}(\vartheta)$ correspond to small values of u, we want $\pi(u)$ to be large for small u values. It turns out that $\pi(u) = 1 - u$ as above determines the *maximally specific* (e.g., Dubois and Prade 1986) possibility measure $\overline{\Pi}$ that is consistent with $Q = \mathsf{Unif}(0,1)$ in the sense above.

Following the fiducial-style logic, which is often referred to as the *extension principle* in this non-additive probabilistic framework (e.g., Zadeh 1975), if the possibility measure $\overline{\Pi}$ with contour π is a quantification of uncertainty about $u^\star$, then we push this through (3) to get the data-dependent possibility measure $\overline{\Pi}_{z^n}$ on Θ with contour

$$\pi_{z^n}(\vartheta) = \pi\big(G(T_{z^n}(\vartheta))\big) = 1 - G(T_{z^n}(\vartheta)), \quad \theta \in \Theta.$$

That is, a generalized IM for θ under this new distribution-free framework assigns upper probabilities to assertions A according to the formula

$$\overline{\Pi}_{z^n}(A) = \sup_{\vartheta \in A}\big\{1 - G(T_{z^n}(\vartheta))\big\}, \quad A \subseteq \Theta. \tag{5}$$

The corresponding lower probability is defined by conjugacy, i.e., $\underline{\Pi}_{z^n}(A) = 1 - \overline{\Pi}_{z^n}(A^c)$.

In certain applications (e.g., Section 5), there may be select features of the inferential target θ that are also of interest. These can be represented as $\phi = \phi(\theta)$, where the notation ϕ is used to represent both the unknown feature and the function mapping the original inferential target to that feature. In such cases, the extension principle can be applied again to construct a marginal IM for ϕ from that for θ. That is, define the possibility contour

$$\pi_{z^n}^{\phi}(\varphi) = \sup_{\vartheta:\phi(\vartheta)=\varphi} \pi_{z^n}(\vartheta), \quad \varphi \in \phi(\Theta).$$

Then lower and upper probabilities for ϕ can be obtained as we did before for θ:

$$\overline{\Pi}_{z^n}^{\phi}(C) = \sup_{\varphi \in C} \pi_{z^n}^{\phi}(\varphi) \quad \text{and} \quad \underline{\Pi}_{z^n}^{\phi}(C) = 1 - \overline{\Pi}_{z^n}^{\phi}(C^c), \quad C \subseteq \phi(\Theta).$$

Validity of the above-defined generalized IM, in the sense below, is a consequence of the consistency between $Q = \mathsf{Unif}(0,1)$ and the possibility measure $\overline{\Pi}$. But it is straightforward to check this property directly, as we do next.

Lemma 1. *The generalized IM above, with output determined by the possibility measure $\overline{\Pi}_{Z^n}$ defined in (5) is valid in the sense that, for all $\alpha \in [0,1]$ and all $A \subseteq \Theta$, the two equivalent conditions hold:*

$$\sup_{P:\theta(P)\in A} P^n\big\{\overline{\Pi}_{Z^n}(A) \leq \alpha\big\} \leq \alpha \quad \textit{and} \quad \sup_{P:\theta(P)\notin A} P^n\{\underline{\Pi}_{Z^n}(A) > 1-\alpha\} \leq \alpha.$$

In particular, $\pi_{Z^n}(\theta(P)) \sim \mathsf{Unif}(0,1)$ as a function of $Z^n \sim P^n$.

Proof. We give the proof for the claim in terms of the upper probability; the lower probability claim follows from this and conjugacy. Fix P and let $\theta = \theta(P)$. For any A that contains θ, monotonicity implies that

$$\overline{\Pi}_{Z^n}(A) \geq \overline{\Pi}_{Z^n}(\{\theta\}),$$

and, for the possibility measure version above, the right-hand side is simply $\pi_{Z^n}(\theta)$, i.e., the possibility contour evaluated at θ. But since $\pi_{Z^n}(\theta) = 1 - G(T_{Z^n}(\theta))$ and G is the distribution function of $T_{Z^n}(\theta)$, it follows immediately that $\pi_{Z^n}(\theta) \sim \mathsf{Unif}(0,1)$, as a function of $Z^n \sim P^n$. Therefore,

$$P^n\{\overline{\Pi}_{Z^n}(A) \leq \alpha\} \leq P^n\{\pi_{Z^n}(\theta) \leq \alpha\} = \alpha,$$

and, since this holds for all A and for all P such that $\theta(P) \in A$, the claim follows. □

We discussed above, in Section 2, the practical interpretation of the validity property. There is another interpretation that readers familiar with the imprecise probability literature might be more comfortable with. If one interprets the IM output as a credal set of probability distributions consistent with $\overline{\Pi}_{z^n}$ in the sense of (4), then the upper probability version of the validity property states that the event "*all the probabilities in the credal set* assign mass $\leq \alpha$ to the true assertion A" is controllably rare. This interpretation also helps to explain why, despite the use of possibility measures, etc., we can still claim that IMs offer "probabilistic" uncertainty quantification.

The following corollary gives an important consequence of the generalized IM's validity property. That is, denote the α level sets of the possibility contour as

$$\mathcal{P}_\alpha(z^n) = \{\vartheta \in \Theta : \pi_{z^n}(\vartheta) > \alpha\}, \quad \alpha \in [0,1]. \tag{6}$$

We refer to these as $100(1-\alpha)\%$ plausibility regions for θ, i.e., these are collections of "sufficiently plausible" values of θ based on data z^n. Validity implies that these are also nominal confidence regions.

Corollary 1. *The generalized IM's plausibility regions in* (6) *are nominal confidence regions in the sense that*

$$\sup_P P^n\big\{\mathcal{P}_\alpha(Z^n) \not\ni \theta(P)\big\} \leq \alpha, \quad \alpha \in [0,1].$$

Proof. Fix P and let $\theta = \theta(P)$. Then it is easy to see that $\mathcal{P}_\alpha(Z^n) \not\ni \theta$ if and only if $\pi_{Z^n}(\theta) \leq \alpha$. Then the claim follows immediately from Lemma 1. □

Remark. For an IM whose output is a possibility measure, a stronger notion of validity can be established. Indeed, virtually the same proof as that above shows

$$\sup_P P^n\{\overline{\Pi}_{Z^n}(A) \leq \alpha \text{ for some } A \ni \theta(P)\} \leq \alpha, \quad \alpha \in [0,1]. \tag{7}$$

To be clear, "for some $A \ni \theta(P)$" corresponds to a union[2] over all such A. Therefore, the event above is much larger event than that for a fixed A that contains $\theta(P)$. So the α upper bound on the probability of a larger event makes for a stronger validity conclusion. In practice, this stronger notion of validity ensures that erroneous conclusions are controllably rare not just in ideal cases where assertions A are specified in advance, but also in the more challenging scenarios where data are used to determine which assertions are to be evaluated. This extension is possible due to the uniformity in (7) being inside the event, which means that it is a rare event that the data analyst can even find a (possibly data-dependent) true assertion to which the IM would assign small upper probability. Since assigning small upper probability to a true assertion corresponds to a case where inference might be erroneous, the uniformity baked in to (7) provides the data analyst some additional comfort and security.

While the formulation just described is simple and achieves the desirable validity property exactly, there is one major problem: its implementation requires knowledge of the distribution function G. Even if a parametric model for P was known to be true, it would be completely unrealistic to expect the distribution of $T_{Z^n}(\theta)$ to be known, so this would never be the case in our present situation where no statistical model is assumed. Next, we put forth a practical version of the generalized IM approach described above.

[2]In general, this is an uncountable union, so its measurability would not be automatic. But it can be readily seen from the argument in the proof of Lemma 1 that the uncountable union equals "$\pi_{Z^n}(\theta) \leq \alpha$," so there is in fact no measurability issue.

3.3 Generalized IM: practical construction

The above formulation is deceptively simple. The obstacle hidden in that presentation is the fact that the distribution function G—based on the distribution of the complicated function $T_{Z^n}(\theta)$, for $Z^n \sim P^n$, with $\theta = \theta(P)$—is unavailable. Fortunately, we can overcome this obstacle by making use the powerful *bootstrap* procedure developed in the seminal work by Efron (1979). The basic idea behind the bootstrap is that iid samples from the empirical distribution of the observed data z^n should closely resemble iid samples from P. Our proposal, therefore, is to approximate the unknown distribution G using this bootstrap strategy. The details of this boostrap-based generalized IM proposal, and its (approximate) validity, are presented in this and the following subsections.

The bootstrap requires an extra level of randomization and proceeds as follows. Our presentation below, which is based on that in Kosorok (2008), may look a bit different from the "sample with replacement from the observed data" common in the literature, but rest assured that it is the same. Let $\xi = (\xi_1, \ldots, \xi_n)$ denote a random n-vector, independent of the data Z^n, with a multinomial distribution, $P_{\text{boot}} = \mathsf{Mult}_n(n^{-1}1_n)$. Then we define a corresponding bootstrap version of the quantity $T_{Z^n}(\theta)$, the form of which depends on whether it is based on minimizing a risk or solving an estimating equation. Start with bootstrap versions of the driving functions R_{z^n} and Ψ_{z^n}:

$$R_{z^n}^\xi(\vartheta) = \frac{1}{n}\sum_{i=1}^n \xi_i \, \ell_\vartheta(z_i) \quad \text{and} \quad \Psi_{z^n}^\xi(\vartheta) = \frac{1}{n}\sum_{i=1}^n \xi_i \, \psi_\vartheta(z_i).$$

The corresponding minimizer/root will be denoted by $\hat\theta_{z^n}^\xi$. The intuition is that the n-vector ξ represents the number of occurrences of each original observation in the bootstrap replicate. The above expressions depend on the random vector ξ and their distribution as a function of ξ will be relevant to us here. This distribution will be approximated in a Monte Carlo way by sampling many copies of ξ from its (multinomial) distribution.

Then the corresponding bootstrap version of $T_{z^n}(\theta)$, in the M- and Z-estimation case, respectively, is given by

$$\begin{aligned} T_{z^n}^\xi(\hat\theta_{z^n}) &= R_{z^n}^\xi(\hat\theta_{z^n}) - R_{z^n}^\xi(\hat\theta_{z^n}^\xi) \\ T_{z^n}^\xi(\hat\theta_{z^n}) &= n \Psi_{z^n}^\xi(\hat\theta_{z^n})^\top \, S_{z^n}^\xi(\hat\theta_{z^n})^{-1} \, \Psi_{z^n}^\xi(\hat\theta_{z^n}), \end{aligned} \tag{8}$$

where the matrix squeezed in the middle is $S_{z^n}^\xi(\vartheta) = n^{-1}\sum_{i=1}^n \xi_i \, \psi_\vartheta(Z_i) \, \psi_\vartheta(Z_i)^\top$. Note the parallels between the random variables $T_{Z^n}(\theta)$ as a function of $Z^n \sim P^n$ with $\theta = \theta(P)$ fixed and $T_{z^n}^\xi(\hat\theta_{z^n})$ as a function of ξ with z^n fixed. That is, in the M-estimation case, say, θ minimizes the expected loss with respect to P whereas $\hat\theta_{z^n}$ minimizes the expected loss with respect to $\xi \sim P_{\text{boot}}$. If we denote by G_{boot} the distribution function of $T_{z^n}^\xi(\hat\theta_{z^n})$ as a function of $\xi \sim P_{\text{boot}}$ for fixed z^n, then the same argument presented in the previous subsection can be used to justify the following formula for a possibility contour:

$$\begin{aligned} \pi_{z^n}^{\text{boot}}(\vartheta) &= 1 - G_{\text{boot}}(T_{z^n}(\vartheta)) \\ &= P_{\text{boot}}\{T_{z^n}^\xi(\hat\theta_{z^n}) > T_{z^n}(\vartheta)\}, \quad \vartheta \in \Theta. \end{aligned} \tag{9}$$

And since our goal is probabilistic inference about θ, the theory developed in the previous subsection suggests a generalized IM whose output is a possibility measure:

$$\overline{\Pi}_{z^n}^{\text{boot}}(A) = \sup_{\vartheta \in A} \pi_{z^n}^{\text{boot}}(\vartheta), \quad A \subseteq \Theta. \tag{10}$$

Algorithm 1: Direct, model-free generalized IM

initialize: data z^n, definition of $T_{z^n}(\cdot)$ in (8), and a grid of ϑ values;
for *b in* $1, \ldots, B$ **do**
  sample $\xi^b \sim P_{\text{boot}}$;
  evaluate $T_{z^n}^{\xi_b}(\hat{\theta}_{z^n})$ according to (8);
  for *each* ϑ *value on the grid* **do**
    evaluate $\hat{\pi}_{z^n}^{\text{boot}}(\vartheta)$ as in (11);
  end
end
return $\hat{\pi}_{z^n}^{\text{boot}}(\vartheta)$ for each ϑ on the grid and/or $\overline{\Pi}_{z^n}^{\text{boot}}(A) \approx \max_{\vartheta \in A \cap \text{grid}} \hat{\pi}_{z^n}^{\text{boot}}(\vartheta)$.

As before, the lower probability $\underline{\Pi}_{z^n}^{\text{boot}}$ is defined by conjugacy.

This version of the possibility contour still is not practical, since evaluating probabilities with respect to P_{boot} requires a sum over all n^n possible values of ξ. For a practical alternative, we suggest a Monte Carlo approximation based on taking samples of ξ from P_{boot}. That is, for a user-specified bootstrap sample size B, define

$$\hat{\pi}_{z^n}^{\text{boot}}(\vartheta) = \frac{1}{B} \sum_{b=1}^{B} 1\{T_{z^n}^{\xi_b}(\hat{\theta}_{z^n}) > T_{z^n}(\vartheta)\}, \quad \xi_b \sim P_{\text{boot}}, \quad b = 1, \ldots, B. \tag{11}$$

For low-dimensional θ, this function can be plotted to visualize our uncertainty quantification based on data Z^n; see Figures 1(a), 2(b), 3(a) and 5(a). Moreover, the possibility measure $\overline{\Pi}_{z^n}^{\text{boot}}(A)$ can be approximated by replacing the supremum with a maximum over a user-specified grid of ϑ values. These details are summarized in Algorithm 1.

If marginalization to some feature $\phi = \phi(\theta)$ of the inferential target were desired, then this can be carried out exactly as described above. Just apply the extension principle to the possibility measure defined by the contour $\hat{\pi}_{z^n}^{\text{boot}}$ to get a marginal IM for ϕ. All the properties enjoyed by the IM for θ, especially the validity property in Theorem 1, apply equally to this marginal IM for ϕ. In particular, marginal plausibility regions for ϕ would achieve the nominal frequentist coverage probability asymptotically.

Note that the proposed generalized IM requires very little input from the user: only the data and a specification of how the inferential target is defined, which effectively identifies T. With just this essential information about the problem at hand, in particular, no statistical model specification required, the proposed method produces a full, data-dependent, probabilistic quantification of uncertainty about θ, which can be used to make (approximately) valid inference; see Section 3.4. In Section 4, we illustrate our proposed method with several practically relevant examples.

3.4 Asymptotic validity

Here we present the (asymptotically approximate) validity property enjoyed by the proposed bootstrap-based generalized IM. Recall that, if the distribution function G were known, as in Section 3.2, then validity followed almost immediately, as shown in Lemma 1. So, if the bootstrap version, G_{boot}, is an accurate approximation of G, then a suitable approximate validity property for the more practical bootstrap-based generalized IM will

follow. More formally, we say that the bootstrap approximation described above of the distribution of $T_{Z^n}(\theta)$, under $Z^n \sim P^n$ with $\theta = \theta(P)$, is *consistent* if

$$\sup_t \big| G^{(n)}(t) - G^{(n)}_{\text{boot}}(t) \big| \to 0 \quad \text{in } P^n\text{-probability as } n \to \infty, \tag{12}$$

where, to highlight their dependence on the sample size, $G^{(n)}$ denotes the exact distribution function for $T_{Z^n}(\theta)$ and $G^{(n)}_{\text{boot}}$ is its bootstrap version. While the intuition behind bootstrap consistency is clear—when n is large, iid sampling from the empirical distribution of Z^n should be roughly the same as iid sampling from P—the precise technical details are complicated and non-trivial. Fortunately, there is a substantial body of literature on bootstrap consistency, starting with Bickel and Freedman (1981) and Singh (1981) and, since then, Wellner and Zhang (1996), Chatterjee and Bose (2005), and Cheng and Huang (2010) who deal with the general M- and Z-estimation cases; see, also, Hall (1992), Shao and Tu (1995), van der Vaart and Wellner (1996), and Kosorok (2008) for textbook-style introductions to the bootstrap consistency theory. The general rule of thumb is that, if the M- or Z-estimator itself is asymptotically normal, i.e., if $n^{1/2}(\hat{\theta}_{Z^n} - \theta)$ converges in distribution to a Gaussian limit, which is true in a wide range of applications, then the bootstrap would be consistent in the sense of (12).

The following theorem makes more precise our above claim that bootstrap consistency is enough to establish approximate validity of our proposed generalized IM.

Theorem 1. *Suppose that the inference problem is such that the bootstrap version of the distribution of $T_{Z^n}(\theta)$, as a function of $Z^n \sim P^n$, with target $\theta = \theta(P)$, is consistent in the sense of* (12)*. Then the bootstrap-based generalized IM for θ, whose output is determined by the possibility measure $\overline{\Pi}^{\text{boot}}_{Z^n}$ defined in* (10)*, is approximately valid in the sense that, for all $\alpha \in [0,1]$, all $A \subseteq \Theta$, the following two equivalent properties hold:*

$$\begin{aligned} \limsup_{n\to\infty} P^n\{\overline{\Pi}^{\text{boot}}_{Z^n}(A) \leq \alpha\} \leq \alpha, \quad &\text{for all } P \text{ with } \theta(P) \in A \\ \limsup_{n\to\infty} P^n\{\underline{\Pi}^{\text{boot}}_{Z^n}(A) > 1-\alpha\} \leq \alpha, \quad &\text{for all } P \text{ with } \theta(P) \notin A. \end{aligned}$$

In particular, $\pi^{\text{boot}}_{Z^n}(\theta(P)) \to \mathsf{Unif}(0,1)$ in distribution, as $n \to \infty$, under $Z^n \sim P^n$.

Proof. Since

$$\pi_{Z^n}(\theta) = 1 - G^{(n)}(T_{Z^n}(\theta)) \quad \text{and} \quad \pi^{\text{boot}}_{Z^n}(\theta) = 1 - G^{(n)}_{\text{boot}}(T_{Z^n}(\theta)),$$

we immediately get

$$\pi^{\text{boot}}_{Z^n}(\theta) = \pi_{Z^n}(\theta) + \Delta_n, \tag{13}$$

where

$$|\Delta_n| = \big| G^{(n)}(T_{Z^n}(\theta)) - G^{(n)}_{\text{boot}}(T_{Z^n}(\theta)) \big| \leq \sup_t \big| G^{(n)}(t) - G^{(n)}_{\text{boot}}(t) \big|.$$

It follows from (12) that $\Delta_n = o_P(1)$. Since $\pi_{Z^n}(\theta) \sim \mathsf{Unif}(0,1)$ for all n, from (13) and Slutsky's theorem we get that $\pi^{\text{boot}}_{Z^n}(\theta(P)) \to \mathsf{Unif}(0,1)$ in distribution as $n \to \infty$. The other two properties in the theorem statement are a consequence of this. □

An immediate and relevant consequence of Theorem 1 is that the bootstrap-based generalized IM's plausibility region

$$\mathcal{P}_\alpha(z^n) = \{\vartheta : \hat{\pi}_{z^n}^{\text{boot}}(\vartheta) > \alpha\}, \quad \alpha \in [0,1], \tag{14}$$

is also an approximate confidence region in the sense that its coverage probability is converging to the nominal level $1-\alpha$ as $n \to \infty$. Similarly, for any $A \subset \Theta$, an asymptotic size-α test of a hypothesis "$\theta(P) \in A$" rejects the hypothesis if and only if $\overline{\Pi}_{z^n}(A) \leq \alpha$. These claims also apply to the summaries—plausibility regions and tests—of the marginal IMs for features $\phi = \phi(\theta)$ derived from the IM for θ.

The theorem above is quite general, but it is not universal, i.e., there are cases when the bootstrap fails to be consistent. These instances of bootstrap failure are associated with certain non-regularities, so our assumption (12) implicitly imposes regularity conditions on the M- or Z-estimation problem. See Section 6 for more discussion about non-regular cases. There we also address the natural question about the accuracy of the approximate validity claims in Theorem 1.

4 Examples

The goal of the present section is to illustrate the generalized IM construction above in various practically relevant examples involving quantiles. Each of the examples follows, roughly, the same structure. We start by providing the appropriate loss function or (vector-valued) estimating equation that links the observed data to the desired quantile. We then explore the IM's basic output, the possibility contour, obtained through Algorithm 1, in several ways. In particular,

- we plot it to visualize the plausibility of different values of the quantile of interest based on a single data set;
- we derive confidence regions for that quantile from it through (14);
- we carry out simulation studies to confirm the IM's approximate validity, checking its behavior for a range of sample sizes.

4.1 Quantiles

The most intuitive example of a quantity of interest that is not most naturally defined as a model parameter is the τ-th quantile, the exact point $\theta = \theta_\tau$ such that $F(\theta) = \tau$, for $\tau \in (0,1)$, where F is the distribution function of a random variable Z. Every distribution has quantiles, but very rarely are they model parameters. Of course, one can make model-based inference on a quantile by specifying a parametric model P_ω, for $\omega \in \Omega$, and defining $\theta = \theta(\omega)$ as the corresponding quantile, but, as argued above, this creates a risk of bias due to model misspecification and/or model selection. Our approach here is model-free, so these risks/challenges are avoided.

Suppose Z has a generic distribution P. Then it is well-known that a general τ^{th} quantile of P can be defined as the minimizer of the risk function $R(\vartheta) = \int \ell_\vartheta(z)\, P(dz)$, where the loss function is given by

$$\ell_\vartheta(z) = \tfrac{1}{2}\big\{(|z-\theta| - z) + (1-2\tau)\theta\big\}.$$

τ	GIM	Conservative	Bootstrap
0.25	0.95 (1.12)	0.96 (1.23)	0.96 (1.20)
0.50	0.95 (0.62)	0.98 (0.69)	0.96 (0.64)
0.75	0.93 (1.12)	0.96 (1.25)	0.94 (1.15)

Table 1: Estimated coverage probabilities and mean length of 95% confidence intervals for the quartiles based on the three following methods: generalized IM (GIM) with $B = 500$; the conservative method based on the binomial distribution; the standard bootstrap method with $B = 500$. The sample size is $n = 100$, with data coming from a Cauchy distribution with location and scale parameters 2 and 1, respectively.

In the special case $\tau = 0.5$, corresponding to the median, this can be reduced to

$$\ell_\vartheta(z) = |z - \vartheta|.$$

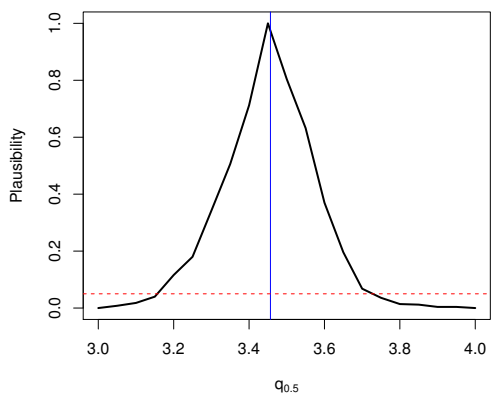
Consistency of the bootstrap, in the sense of (12), is established in this quantile inference problem by both Bickel and Freedman (1981) and Singh (1981).

As an illustration, suppose that P is a $\mathsf{Gamma}(4, 1)$. Interest here is in the median $\theta = \theta_{0.5}$ which, in this case, is roughly equal to 3.67. Figure 1(a) shows the plausibility contour in (11) with $B = 500$ and the loss function above for a single data set z^n with $n = 100$. The peak is at the M-estimator, i.e., the sample median, which is close to the true median, and the horizontal line determines the corresponding 95% plausibility interval, derived by (14). In order to check that approximate validity is attained, a simulation study was conducted where the above scenario is repeated 1000 times and, for each data set, $\hat{\pi}^{\text{boot}}_{z^n}(\theta)$ is evaluated at $\theta = 3.67$. Figure 1(b) shows that the distribution of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$ is close to $\mathsf{Unif}(0, 1)$, so approximate validity is verified. The same simulation is repeated for $\tau = 0.25$ and 0.75, showing that the approximate validity conclusion is not specific to the median. Finally, Figures 1(c) and (d) show that smaller sample sizes do not affect the good conclusions observed in Figure 1(b) too much.

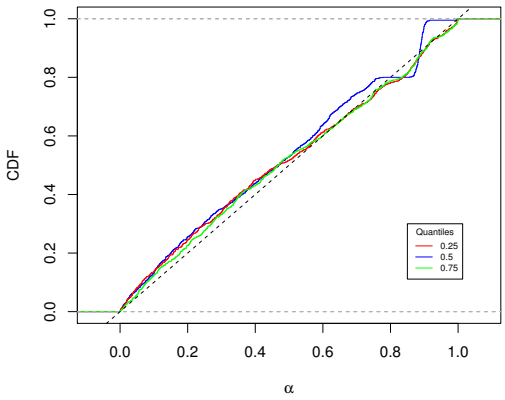
There are a number of different strategies available for constructing confidence intervals for population quantiles. For further illustration, we compare our method to two of them: an exact-but-conservative solution based on the binomial distribution and a basic bootstrap procedure, which resamples the data with replacement, computes the desired quantile and then reports, for a $(1 - \alpha)\%$ confidence interval, the $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ quantiles of this bootstrapped distribution. For our simulation, we consider P to be a Cauchy distribution with location and scale parameters equal to 2 and 1, respectively. We generated 1000 data sets of size $n = 100$ and, from each, 95% confidence intervals for θ_τ, with $\tau \in \{0.25, 0.50, 0.75\}$, based on the three methods are extracted. Table 1 reports the estimated coverage probabilities and mean length of these intervals. Note that approximately validity of the generalized IM solution is confirmed. Moreover, it is slightly more efficient than both the conservative and basic bootstrap methods.

4.2 Multivariate median

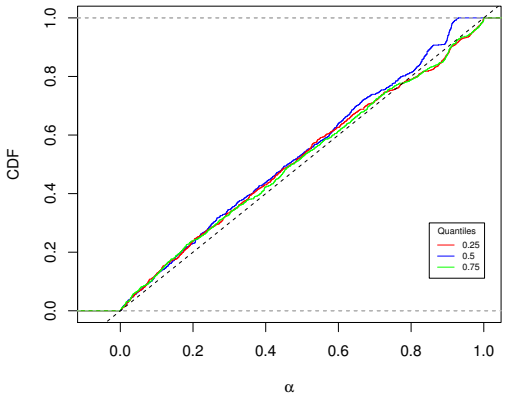
In univariate analysis, it is well known that the median is a more robust measure of the distribution's center than the mean. This is also the case in multivariate analysis. However, replacing the multivariate mean by a multivariate median is not so straightforward.



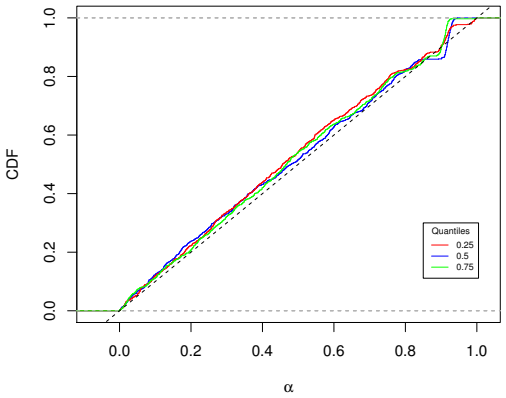
(a) Plausibility contour in (11)



(b) Distribution function of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$, $n = 50$



(c) Distribution function of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$, $n = 75$



(d) Distribution function of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$, $n = 100$

Figure 1: Details from the quantile example in Section 4.1. The results in Panel (b), (c) and (d) are based on 1000 data replications, and shown for $\tau \in \{0.25, 0.5, 0.75\}$.

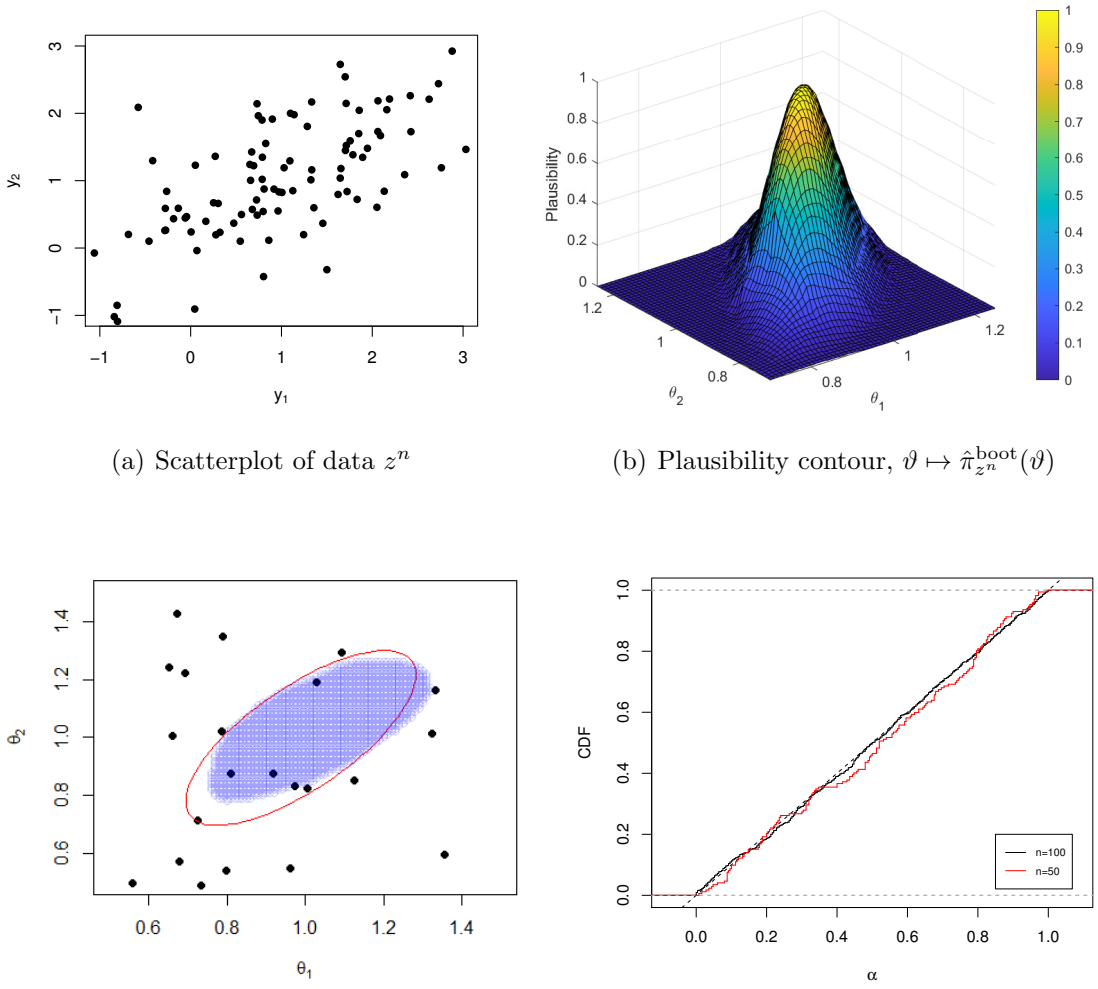
Indeed, since multivariate data do not have a natural ordering, there are various different ways of defining an order, each leading to a definition of the multivariate median or, more generally, multivariate quantiles (Becker et al. 2014).

The most common version of a multivariate median is the *spatial median*. This can be defined similar to the univariate median described above, as the minimizer of a risk function $R(\vartheta) = \int \ell_\vartheta(z)\, P(dz)$ where the loss is given by

$$\ell_\vartheta(z) = \|z - \vartheta\|_2 - \|z\|_2, \quad z, \vartheta \in \mathbb{R}^q, \quad q \geq 1,$$

where $\|\cdot\|_2$ is the usual ℓ_2-norm for vectors in $\mathbb{R}^q$. Alternatively, the spatial median can be defined as a Z-estimator, i.e., it satisfies the system of equations $\Psi(\vartheta) = \int \psi_\vartheta(z)\, P(dz) = 0$ where $\psi_\vartheta(z)$ is a q-vector with components

$$\psi_\vartheta(z)_j = \frac{z_j - \vartheta_j}{\|z - \vartheta\|_2}, \quad j = 1, \ldots, q.$$



(a) Scatterplot of data z^n

(b) Plausibility contour, $\vartheta \mapsto \hat{\pi}_{z^n}^{\text{boot}}(\vartheta)$

(c) Confidence regions for θ

(d) Distribution function of $\hat{\pi}_{Z^n}^{\text{boot}}(\theta)$

Figure 2: Results for the illustration in Section 4.2. Panel (c): classical 95% confidence ellipse based on asymptotic normality (red) and the 95% plausibility region (blue).

Consistency of the bootstrap for the multivariate median was established recently as part of Theorem 1 in Bhattacharya and Ghosal (2022).

For a quick illustration, Figure 2(a) shows the data $z_i \in \mathbb{R}^2$ for $i = 1, \ldots, n = 200$, which are samples from bivariate normal with mean $\theta = (1, 1)^\top$, unit variances, and correlation 0.7. In Figure 2(b), the plausibility contour in (11) is shown, based on the ψ_ϑ function defined above, with T the quadratic form in (8), and $B = 500$. The shaded area in Figure 2(c) represents the 95% plausibility region for θ derived by (14) and, in black, the classic 95% confidence ellipse based on the asymptotic normality. Note how the IM solution is more efficient. Figure 2(d) shows the resulting empirical distribution of the simulation study where the above scenario is repeated 1000 times and, for each data set, $\hat{\pi}_{Z^n}^{\text{boot}}(\theta)$ is evaluated. Approximate validity is once again verified. This is also true when a smaller sample of size $n = 50$ is considered.

4.3 Quantile regression

Let Y be a response variable coupled with a covariate $X \in \mathbb{R}^p$. The goal of *quantile regression* is to estimate the quantile for the conditional distribution of Y, given X. Fix a probability $\tau \in (0,1)$ and let $Q_x(\tau)$ denote the τ^{th} conditional quantile of Y, given $X = x$. Then the quantile regression model says

$$Q_x(\tau) = x^\top \theta,$$

where $\theta = \theta_\tau \in \mathbb{R}^p$ is the vector of regression coefficients of interest. This "model" describes the functional form of the quantile function, but does not determine the distribution of Y, given $X = x$. Towards making inference on θ, Koenker and Bassett (1978) show that θ is a risk-minimizer with respect to the loss function

$$\ell_\vartheta(x, y) = |y - x^\top \vartheta| - (2\tau - 1)x^\top \vartheta.$$

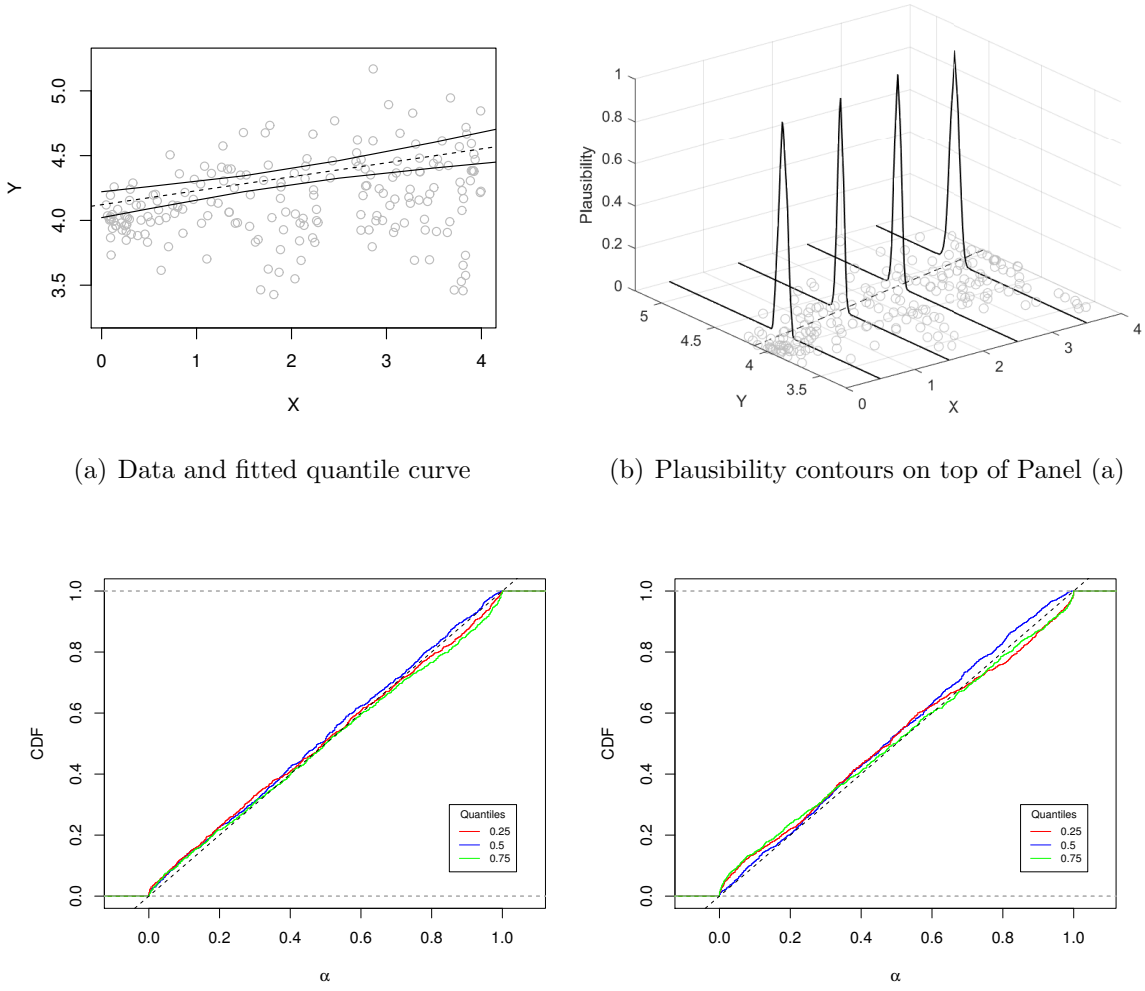
Properties of the quantile regression M-estimator, e.g., its asymptotic normality, are investigated in Koenker (2005) and, in particular, bootstrap consistency is established in Hahn (1995). Here we present an illustration of the proposed generalized IM solution to the quantile regression problem.

Let $X_i \overset{\text{iid}}{\sim} \mathsf{Unif}(0,4)$, $i = 1, \ldots, n$, with $n = 200$, and let $Y_i = \mu(X_i) + \varepsilon(X_i)$, where $\mu(x) = 4 + 0.1x$, and $\varepsilon(x) \sim \mathsf{N}\big(0, (0.1 + 0.1x)^2\big)$. Suppose the interest is $\theta = \theta_\tau$ for $\tau = 0.75$. Figure 3(b) displays the data, the estimated quantile regression line corresponding to the Z-estimator $\hat{\theta}_{z^n}$. Plausibility contours are obtained for θ based on the loss function above and $B = 500$, and the plot shows the marginal plausibility contours for μ at selected values of x. The corresponding 95% marginal plausibility band for μ is shown in Figure 3(a). Approximate validity of the plausibility bands is implied by the approximate validity of the generalized IM. To check this claim empirically, we simulate 1000 data sets according the above scheme and calculated $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$, in each replication. Figure 3(c) shows the empirical distribution of these values over replications and it is clear this closely matches a uniform distribution, confirming Theorem 1. The same plots for $\tau = 0.25$ and $\tau = 0.50$ are included and all suggest the uniform approximation for the distribution of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$ is accurate across a range of quantile levels. Figure 3(d) is analogous to Figure 3(c), but considering $n = 100$. As in the previous examples, the proposed solution achieves approximately validity even in finite-sample scenarios.

5 Dynamic treatment regimes

5.1 Introduction

Compared to the common one-size-fits-all approach in medicine, where treatment decisions are developed for the "average" patient, precision medicine focuses on tailoring treatment decisions to individual patients based on certain characteristics of their profile. *Dynamic treatment regimes* provide a formal precision medicine framework, where the individualization of treatments is dictated by a sequence of decision rules, one per stage of intervention, that are based on the patient's "history," which includes both covariates and past treatments (Chakraborty and Murphy 2014).



(a) Data and fitted quantile curve

(b) Plausibility contours on top of Panel (a)

(c) Distribution function of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta_\tau)$, $n = 200$

(d) Distribution function of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta_\tau)$, $n = 100$

Figure 3: Results for the illustration in Section 4.3. Panel (a) shows the fitted quantile curve, for $\tau = 0.75$, with the 95% plausibility band.

In this section, we aim to provide a generalized IM solution to relevant problems that arise when considering dynamic treatment regimes, or regimes for short. For example, given a specific regime, a first basic problem is quantifying uncertainty about the expected outcome if the population under study were to receive treatment according to its rules. This expected outcome is also referred to as the *value* of a regime. In Section 5.3, we construct a generalized IM for this purpose. A more challenging problem is where there is a large class of candidate decision rules and the goal is to identify an optimal one, i.e., the regime that maximizes the expected benefit to patients based on their history. In Section 5.4 we make the notion of an optimal regime precise, and develop a generalized IM to quantify uncertainty about its value.

Before tackling the above relevant problems, we provide a short background on dynamic treatment regimes and setup the basic notation. The details presented throughout

this Section are largely based on Tsiatis et al. (2020).

5.2 Background and notation

For simplicity, our focus throughout this Section will be on the so called *single decision* problem, where there is only one stage at which a treatment must be selected from among a given set of available options.[3] In this situation, a dynamic treatment regime consists of a single rule that takes as input the available patient information and returns as output one of the treatment options. Denote this set of possible treatment action options as $\mathcal{A}$. Formally, a dynamic treatment regime is defined as a *decision rule* $d(x)$, a function that maps an individual's covariates to a treatment option in $\mathcal{A}$, that is, $d : \mathcal{X} \to \mathcal{A}$, where $\mathcal{X}$ is the support of the covariates X. Here, we explore only the simplest case where $\mathcal{A}$ contains two treatment options, e.g., a control/active treatment scenario, so d takes $X \in \mathcal{X}$ as input and returns either 0 or 1.

Let $Z_i = (X_i, A_i, Y_i)$, for $i = 1, \ldots, n$, represent the observed data from n patients, where A_i, Y_i and X_i denote, respectively, the treatment received (either 0 or 1), the observed outcome under treatment A_i, and the covariates collected, for the i^{th} patient. Central to both the definition of the value of a given regime and the notion of the optimal regime, to be explored, respectively, in Sections 5.3 and 5.4 below, is the concept of *potential outcome* for any regime $d \in \mathcal{D}$. Generally speaking, a potential outcome (e.g., Rubin 1974, 2005) is the outcome for an individual under a potential treatment. In our context, the random variables $Y^*(0)$ and $Y^*(1)$ represent the outcome that would be achieved by a randomly chosen individual with covariates X in the population of interest if she were to receive treatment 0 or 1, respectively. Note that potential outcomes are hypothetical constructs, since a patient receives only one of the treatments, not both. The idea is to consider what this outcome would have been had the patient received the other treatment option. Now, if treatment is assigned according to regime d, the potential outcome of a patient is defined as

$$Y^*(d) = Y^*(1)\, d(X) + Y^*(0)\, \{1 - d(X)\}. \tag{15}$$

We end this subsection by pointing out that the essential results to be explored in the upcoming subsections depend on three fundamental assumptions that are common in the causal inference literature.

- *Stable unit treatment value assumption:* the outcome Y of a patient who received treatment A is the same as her potential outcome for that treatment, i.e.,

$$Y = Y^*(1)\, A + Y^*(0)\, (1 - A).$$

- *No unmeasured confounders assumption:* all of the information used to make treatment decisions is captured by the observed covariates X, so that

$$\left[\{Y^*(0), Y^*(1)\} \perp\!\!\!\perp A\right] \mid X.$$

- *Positivity assumption:* For any $X = x$, there are individuals receiving both treatment options, that is, $P(A = a \mid X = x) > 0$ for $a = 0, 1$.

[3] We adopt the convention in Tsiatis et al. (2020) that any treatment regime, single or multistage, whose decision rules potentially vary according to baseline and evolving patient information, is dynamic. However, other authors (e.g., Murphy et al. 2001) consider single decision regimes as "non-dynamic."

5.3 Value of a regime

5.3.1 Overview

When considering a specific regime $d \in \mathcal{D}$, a fundamental question is how its use in the entire population would affect the outcome of interest, on average. With the definition of a potential outcomes in (15), the value of any regime $d \in \mathcal{D}$ is defined as

$$\mathcal{V}(d) = E\{Y^*(d)\}.$$

Towards uncertainty quantification of $\mathcal{V}(d)$ based on observed data $Z^n = z^n$, the challenge is deducing the distribution of $Y^*(d)$, which depends on that of $(X, Y^*(1), Y^*(0))$, from the distribution of the observable (X, A, Y). Under the assumptions stated in the end of Section 5.2, it can be shown that

$$E\{Y^*(d)\} = E[E(Y \mid X, A = 1)\, d(X) + E[E(Y \mid X, A = 0)\, \{1 - d(X)\}], \qquad (16)$$

where the outer expectation is with respect to the marginal distribution of X. If we introduce an outcome regression relationship—or *Q-function*—for the conditional mean,

$$Q_{x,a}(\theta) = E(Y \mid X = x, A = a), \qquad (17)$$

depending on a parameter θ; see (19). Then (16) becomes

$$E\{Y^*(d)\} = E[Q_{X,1}(\theta)\, d(X) + Q_{X,0}(\theta)\, \{1 - d(X)\}]. \qquad (18)$$

5.3.2 Generalized IM construction

First, the simple connection between the Q-function—which depends on a parameter θ—and the mean of the response Y allows for a straightforward IM construction for θ interpreted as a risk minimizer. For a given form describing the Q-function's dependence on the parameter θ, we can define a loss function as

$$\ell_\vartheta(z) = \{y - Q_{x,a}(\vartheta)\}^2, \quad z = (x, a, y).$$

Then the generalized IM construction for the minimizer of the expected loss proceeds exactly as in, say, the quantile regression application above. Asymptotic normality of the corresponding M-estimator and consistency of the bootstrap hold for very general Q-function specifications. Note that this generalized IM construction for inference on the risk minimizer θ does not require that the posited functional form of Q to be correct. That is, the existence of the risk minimizer does not require that $Q_{x,a}(\theta)$ be the *true* conditional mean of Y, given $X = x$ and $A = a$; moreover, as we discussed in Section 1, if the risk minimizer exists, the it is a "real" inferential target, so drawing inference on the risk minimizer is meaningful whether there is a correctly-specfied model or not.

We are not primarily interested in the aforementioned risk minimizer since, typically, the inferential target is some other characteristic of the problem. Fortunately, these other characteristics can often be expressed as functions of θ, i.e., as features $\phi = \phi(\theta)$; we will consider two such features below. From the previously-described generalized IM for θ, uncertainty quantification about the value of a given regime is readily obtained through

marginalization as discussed in Section 3.2. There is a catch, however, that deserves to be emphasized: these features have their desired interpretation only as functions of *the parameter θ on which the true Q-function depends*. So, in order for marginal inference about these particular features, derived from a generalized IM for θ, to be meaningful, it is required to assume that the posited form of the Q-function is correctly specified; that is, θ is not just the risk minimizer but determines the true conditional mean function. Note, however, that this does not require correct specification of a statistical model—which includes distributional forms—for the observable data $Z_i = (X_i, A_i, Y_i)$. The situation we are describing here falls under the general umbrella of *semiparametric inference*, where only a part of the model is assumed to be correctly specified.

We present the details of our proposed generalized IM in the context of an example. Consider the simulated observational study presented in `https://laber-labs.com/dtr-book/booktoc.html`,[4] whose objective is to assess the effectiveness of a fictitious medication developed for the treatment of hypertension. Each patient in this study either received the new medication ($A = 1$) or received no treatment ($A = 0$) based on patient/physician discretion. The outcome of interest Y is the change in systolic blood pressure (mmHg) after six months of treatment, i.e., $Y = Y_0 - Y_6$. The covariates $X = (X_1, X_2)$ are, respectively, the total cholesterol (mg/dl) and the potassium level (mg/dl). Here is how the data are generated. Let $Y_{0,i} \overset{\text{iid}}{\sim} \mathsf{N}(160, 12^2)$, $i = 1, \ldots, n$, with $n = 1000$ and the constraint $140 < Y_{0,i} \leq 200$. Let $X_{1,i} \overset{\text{iid}}{\sim} \mathsf{N}(211, 45^2)$, $X_{2,i} \overset{\text{iid}}{\sim} \mathsf{N}(4.2, 0.35^2)$, $A_i \sim \mathsf{Ber}(\pi(x_{1,i}, y_{0,i}))$, where

$$\pi(x, y) = \frac{\exp\{-16.348 + 0.078y + 0.017x\}}{1 + \exp\{-16.348 + 0.078y + 0.017x\}},$$

and $Y_{6,i} = Y_{0,i} - \mathsf{N}(\mu(x_i, a_i), 3^2)$, where

$$\mu(x, a) = -15 - 0.2x_1 + 12x_2 + a(-65 + 0.5x_1 - 5.5x_2).$$

This implies an outcome regression relationship $Q_{x,a}(\theta)$ in (17) given by

$$Q_{x,a}(\theta) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 a + \theta_4 a x_1 + \theta_5 a x_2, \tag{19}$$

depending on $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$. From the definition of $\mu(x, a)$ above, the true values of θ_0, θ_1, θ_2, θ_3, θ_4, and θ_5 are -15, -0.2, 12, -65, 0.5 and -5.5, respectively. Note that the only aspect of the above description that the generalized IM assumes as "true" is the Q-function specification in (19); the statements concerning the *distributions* of the observables are not used at all in the generalized IM formulation nor are they assumed true in the supporting theory presented in Section 3.4.

As a first check, we empirically verify the approximate validity claim in Theorem 1 for inference on θ, the true parameters of the regression function. Figure 4 shows the distribution function of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$ based on repeated sampling from the data-generating process described above, with θ the true values. As the theory predicts, we see that the distribution of $\hat{\pi}^{\text{boot}}_{Z^n}(\theta)$ is almost exactly $\mathsf{Unif}(0, 1)$, which means that inference drawn on θ in this setting is approximately—and almost exactly—valid.

[4]This is Tsiatis et al. (2020)'s companion website, where several examples are provided. The particular example we explore here can be found under the "Chapter 3" tab.

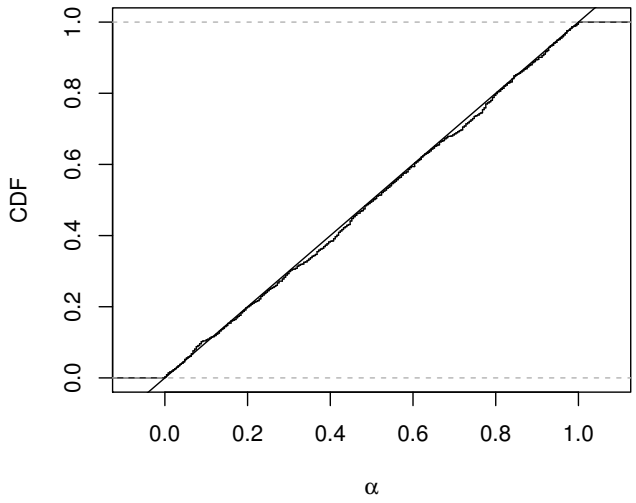


Figure 4: The plot shows the empirical distribution function of $\hat{\pi}_{Z^n}^{\text{boot}}(\theta)$ under repeated sampling from the data-generating process described in the text.

Next, we consider three different marginalization examples. First, in treatment/control scenarios, it is common that the first enquiry performed by the data analyst concerns the presence of a *treatment effect*. Here the treatment effect is the average change in systolic blood pressure, after six months, if everyone in the population took the new medication compared to if everyone took the old medication. For the Q-function in (19), the treatment effect is $\phi := \theta_3 + \theta_4 E(X_1) + \theta_5 E(X_2)$, so interest is in the assertion "$\phi = 0$." For the particular data z^n, the upper probability assigned to this assertion is approximately 0; this is not particularly surprising, given that the true treatment effect is equal to $-65 + 0.5 \times 211 - 5.5 \times 4.2 = 17.4$ mmHg.

Second, we investigate marginal inference on the value of a fixed regime. In this case, we consider two such regimes:

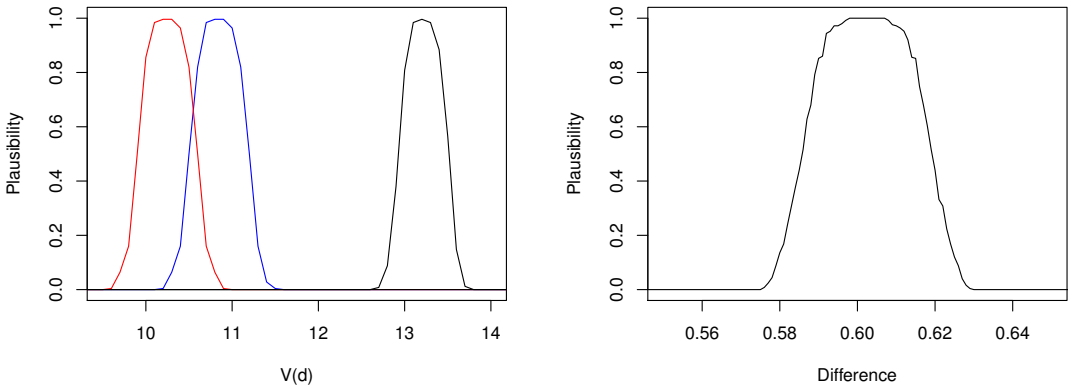
- the *static* regime where all individuals are recommended to receive the new medication, i.e., $d(X) \equiv 1$;
- the *covariate-dependent* regime that assigns patients to receive the treatment if their cholesterol level exceeds a certain threshold, i.e.,

$$d(X) = 1\{X_1 > 120 \text{ mg/dl}\}. \tag{20}$$

From (18), the values of the static and covariate dependent regimes are, respectively,

$$\begin{aligned} \phi_{\text{ST}} &:= E\{\theta_0 + \theta_3 + (\theta_1 + \theta_4)X_1 + (\theta_2 + \theta_5)X_2\} \\ \phi_{\text{CD}} &:= E[\theta_0 + \theta_1 X_1 + \theta_2 X_2 + (\theta_3 + \theta_4 X_1 + \theta_5 X_2) 1\{X_1 > 120\}], \end{aligned}$$

and Figure 5(a) shows the corresponding marginal plausibility contours for each, both obtained from the generalized IM for θ. The plot suggests that, not unexpectedly, the covariate-dependent regime is no worse than the static regime.



(a) Plausibility contour for $\phi = \mathcal{V}(d)$

(b) Plausibility contour for $\phi_{\text{CD}} - \phi_{\text{ST}}$

Figure 5: In Panel (a), the red curve is the plausibility contour for $\phi = \mathcal{V}(d)$ under the static regime, $d(X) \equiv 1$, blue is for the covariate-dependent regime (20), and black is the optimal regime discussed in Section 5.4. Panel (b) shows the plausibility contour for the difference between the covariate-dependent and static regimes.

Third, a relevant question in practice might be whether there is a difference between two fixed regimes. For example, given the overlap of the plausibility contours in Figure 5(a), one may wonder if the covariate-dependent regime is in fact better than the static regime. More specifically, interest is whether $\phi_{\text{CD}} > \phi_{\text{ST}}$. Figure 5(b) shows the marginal plausibility contour for the difference $\phi_{\text{CD}} - \phi_{\text{ST}}$, obtained, once again, from the generalized IM for θ. Despite the differences being small, the assertion "$\phi_{\text{CD}} = \phi_{\text{ST}}$" has zero plausibility, confirming the superiority of the covariate-dependent regime. Whether these small differences are practically relevant is a separate question.

5.4 Optimal regime

Estimating the value of a specific decision rule, d, may be of some interest in applications, but a more challenging problem is to identify the optimal regime within a given set $\mathcal{D}$, that is, a regime that leads to the best benefit on average if used to select treatment in the population. For situations where larger values of the response variable mean greater benefit to the patient, like in the example in Section 5.3.2 above, the optimal regime d^* is defined as one that leads to the maximum value among all $d \in \mathcal{D}$, i.e.,

$$d^* = \arg\max_{d \in \mathcal{D}} \mathcal{V}(d)$$

or, equivalently,

$$E\{Y^*(d^*)\} \geq E\{Y^*(d)\} \quad \text{for all } d \in \mathcal{D}. \tag{21}$$

It is possible that more than one regime satisfies (21), but we will not concern ourselves with this technicality here. Recall that our focus is on a single stage treatment regime

with $\mathcal{A} = \{0, 1\}$. For such a case, Tsiatis et al. (2020) prove that (21) is satisfied by

$$\begin{aligned} d^*(x) &= \arg\max_{a \in \mathcal{A}} E\{Y^*(a) \mid X = x\} \\ &= 1[E\{Y^*(1) \mid X = x\} > E\{Y^*(0) \mid X = x\}]. \end{aligned} \tag{22}$$

Just like in the previous subsection, uncertainty quantification about d^* based on observed data z^n requires us to rewrite (22) in terms of Z^n. Under the assumptions stated at the end of Section 5.2, (22) can be equivalently written as

$$d^*(x) = \arg\max_{a \in \mathcal{A}} E\{Y \mid X = x, A = a\}, \tag{23}$$

which, under the outcome regression model formulation can itself be rewritten as

$$d^*(x) = \arg\max_{a \in \mathcal{A}} Q_{x,a}(\theta) = 1\{Q_{x,1}(\theta) > Q_{x,0}(\theta)\}. \tag{24}$$

Moreover, the value of this optimal treatment is given by

$$\mathcal{V}(d^*) = E\Big\{\max_{a \in \mathcal{A}} Q_{X,a}(\theta)\Big\}, \tag{25}$$

where, again, the outer expectation is with respect to the distribution of X. The right-hand side of the above display is, again, a function $\phi = \phi(\theta)$ of θ, so if we have a generalized IM for θ, then we can readily obtain a marginal IM for ϕ.

As an illustration, consider again the example explored in Section 5.3.2, where $Q_{x,a}(\theta)$ is linear as in (19). In this case, it is clear that (24) becomes

$$d^*(x) = 1\{\theta_3 + \theta_4 x_1 + \theta_5 x_2 > 0\}.$$

From (25), the value of this optimal regime is given by

$$\mathcal{V}(d^*) = E\{\theta_0 + \theta_1 X_1 + \theta_2 X_2 + (\theta_3 + \theta_4 X_1 + \theta_5 X_2) 1\{\theta_3 + \theta_4 X_1 + \theta_5 X_2 > 0\}\}. \tag{26}$$

Uncertainty quantification about the value $\mathcal{V}(d^*)$ above is obtained through marginalization, as it was the case for the fixed regimes considered earlier. For example, to obtain the plausibility contour in Figure 5(a), one starts with the generalized IM for θ and marginalize to the corresponding $\mathcal{V}(d^*)$ in (26). Note how the value of d^* is significantly greater than the values of the two fixed regimes also shown in Figure 5.

6 Conclusion

Here we focused on direct, data-driven uncertainty quantification for unknowns defined as risk minimizers or solutions to estimating equations rather than parameters of a statistical model. We presented a new generalized IM that not only avoids the explicit description of the data generating process, but does not require a statistical model at all. We showed that this construction leads to approximately valid uncertainty quantification in the sense of Theorem 1. This provides guarantees beyond those from classical confidence regions. That is, the IM's validity property applies to belief assignments to all assertions about the inferential target—even marginal inference about features of the original inferential

target. To our knowledge, this is the first paper providing direct and valid probabilistic uncertainty quantification in this practically relevant class of learning problems.

Applications in cases beyond the simple, low-dimensional problems above will be reported elsewhere. Of course, larger dimension creates computational challenges, so getting marginal plausibility contours for each component of the high-dimensional inferential target in an efficient way remains an open question. Since evaluation of $\hat{\pi}_{z^n}^{\text{boot}}$ is based on sampling (bootstrap or Monte Carlo), and marginalization is optimization in this imprecise probability setting, techniques like *stochastic approximation* or *stochastic gradient descent* seem especially promising; see, e.g., Syring and Martin (2021).

We end this section with a brief discussion of some open questions. First, it is well-known that the bootstrap often enjoys a certain higher-order accuracy, that is, the coverage probability of bootstrap-based confidence regions converge to the nominal level at a rate faster than the expected root-n; see, e.g., Hall (1992) and Lehmann (1999). Similarly, in other settings, with simpler versions of the generalized IM framework developed here, it was observed empirically that the uniform limit distribution approximation for "$\hat{\pi}_{Z^n}(\theta)$" was quite accurate, even for small samples, suggesting some higher-order accuracy. The proposed generalized IM in this paper borrows aspects of these two approaches that (at least empirically) enjoy higher-order accuracy. Then the question is if this combination of two higher-order accurate methods is also higher-order accurate?

Second, although this did not appear in our illustrations in Section 5, an especially challenging aspect of the dynamic treatment regime problem is non-regularity, resulting from the non-smooth "max" operator in (23), that affects the limit distribution theory of the M/Z-estimators and, in turn, the corresponding bootstrap-based inference. A common remedy for failure of the bootstrap, e.g., due to non-regularity, is to "undersample" with the so-called *m-out-of-n bootstrap*. That is, Bickel et al. (1997) showed that, by taking bootstrap samples of size $m = o(n)$, bootstrap failure could be avoided. A more sophisticated m-out-of-n bootstrap scheme was proposed in Chakraborty et al. (2013) that could prevent bootstrap failure in the dynamic treatment regime setting resulting from the non-regularity induced by the "max" operator. An interesting follow-up project could investigate the reliability of the proposed generalized IM equipped with a m-out-of-n bootstrap strategy under non-regularity.

Acknowledgments

The authors thank the Guest Editor and the two anonymous reviewers for their helpful feedback on a previous version of this manuscript. This work is partially supported by the U.S. National Science Foundation, grants DMS–1811802 and SES–2051225.

References

Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):20180565.

Becker, C., Fried, R., and Kuhnt, S. (2014). *Robustness and Complex Data Structures:*

Festschrift in Honour of Ursula Gather. SpringerLink : Bücher. Springer Berlin Heidelberg.

Bhattacharya, I. and Ghosal, S. (2022). Bayesian inference on multivariate medians and quantiles. *Statistica Sinica*, 32(1):517–538.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica*, 7(1):1–31.

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference*. Springer Texts in Statistics. Springer, New York.

Cahoon, J. and Martin, R. (2020). Generalized inferential models for meta-analyses based on few studies. *Statistics and Applications*, 18(2):299–316.

Cahoon, J. and Martin, R. (2021). Generalized inferential models for censored data. *International Journal of Approximate Reasoning*, 137:51–66.

Cella, L. and Martin, R. (2021a). Approximately valid and model-free possibilistic inference. In Denœux, T., Lefèvre, E., Liu, Z., and Pichon, F., editors, *Belief Functions: Theory and Applications*, pages 127–136, Cham. Springer International Publishing.

Cella, L. and Martin, R. (2021b). Valid inferential models for prediction in supervised learning problems. In Cano, A., De Bock, J., Miranda, E., and Moral, S., editors, *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, pages 72–82. PMLR. Extended version available at `https://researchers.one/articles/21.12.00002v1`.

Cella, L. and Martin, R. (2022). Validity, consonant plausibility measures, and conformal prediction. *International Journal of Approximate Reasoning*, 141:110–130.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, 1(1):447–464.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *The Annals of Statistics*, 33(1):414 – 436.

Cheng, G. and Huang, J. (2010). Bootstrap consistency for general semiparametric M-estimation. *The Annals of Statistics*, 38(5):2884–2915.

Choquet, G. (1953–1954). Theory of capacities. *Université de Grenoble. Annales de l'Institut Fourier*, 5:131–295 (1955).

Conover, W. (1971). *Practical Nonparametric Statistics*. Wiley.

Cui, Y. and Hannig, J. (2019). Nonparametric generalized fiducial inference for survival functions under censoring. *Biometrika*, 106(3):501–518.

Cunen, C., Hjort, N. L., and Schweder, T. (2020). Confidence in confidence distributions! *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2237):20190781.

Dawid, A. P. and Stone, M. (1982). The functional-model basis of fiducial inference. *The Annals of Statistics*, 10(4):1054 – 1067.

Dempster, A. P. (1968). A generalization of Bayesian inference. (With discussion). *Journal of the Royal Statistical Society, Series B*, 30:205–247.

Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365 – 377.

Dempster, A. P. (2014). Statistical inference from a Dempster–Shafer perspective. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 24. Chapman & Hall/CRC Press.

Denœux, T. (2009). Extending stochastic ordering to belief functions on the real line. *Information Sciences*, 179(9):1362–1376.

Denœux, T. (2014). Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547.

Dubois, D. and Prade, H. (1986). The principle of minimum specificity as a basis for evidential reasoning. In *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 75–84. Springer.

Dubois, D. and Prade, H. (1988). *Possibility Theory*. Plenum Press, New York.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. Hafner Press, New York, 3rd edition.

Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley & Sons Inc., New York.

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence. *Statistical Science*, 26:299–316.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. Springer.

Godambe, V. (1991). *Estimating Functions*. Oxford Science Publications. Clarendon Press.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, 11(1):105–121.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer-Verlag, New York.

Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361.

Hjort, N. L., Holmes, C. C., Müller, P., and Walker, S. G., editors (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hose, D. and Hanss, M. (2021). A universal approach to imprecise probabilities in possibility theory. *International Journal of Approximate Reasoning*, 133:133–158.

Huber, P. (1981). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer Texts in Statistics. Springer-Verlag, New York.

Liu, C. and Martin, R. (2021). Inferential models and possibility measures. *Handbook of Bayesian, Fiducial, and Frequentist Inference*, to appear; `arXiv:2008.06874`.

Martin, R. (2015). Plausibility functions and exact frequentist inference. *Journal of the American Statistical Association*, 110(512):1552–1561.

Martin, R. (2018). On an inferential model construction using generalized associations. *Journal of Statistical Planning and Inference*, 195:105–115.

Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113:39–73.

Martin, R. (2021). An imprecise-probabilistic characterization of frequentist statistical inference. *Researchers.One*, `https://researchers.one/articles/21.01.00002`.

Martin, R. (2022). Valid and efficient imprecise-probabilistic inference across a spectrum of partial prior information. `https://researchers.one/articles/21.05.00001`.

Martin, R., Balch, M. S., and Ferson, S. (2021). Response to the comment confidence in confidence distributions! *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2250):20200579.

Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108:301–313.

Martin, R. and Liu, C. (2015). *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press.

Martin, R. and Syring, N. (2022). Direct Gibbs posterior inference on risk minimizers: construction, concentration, and calibration. In *Handbook of Statistics: Advancements in Bayesian Methods and Implementation*, to appear; `arXiv:2203.09381`.

Molchanov, I. (2005). *Theory of Random Sets*. Probability and Its Applications (New York). Springer-Verlag London Ltd., London.

Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423. PMID: 20019887.

Nguyen, H. T. (2006). *An Introduction to Random Sets*. Chapman & Hall/CRC, Boca Raton, FL.

Reid, N. and Cox, D. R. (2015). On some principles of statistical inference. *International Statistical Review*, 83(2):293–308.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology.*, 66(5):688–701.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Shao, J. and Tu, D. S. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer-Verlag, New York.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, 9(6):1187–1195.

Syring, N. and Martin, R. (2021). Stochastic optimization for numerical evaluation of imprecise probabilities. In Cano, A., De Bock, J., Miranda, E., and Moral, S., editors, *Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, pages 289–298. PMLR.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2020). *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman & Hall/CRC.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. *Journal of Statistical Planning and Inference*, 105:35–65.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York.

Wellner, J. A. and Zhang, Y. (1996). Bootstrapping Z-estimators. Technical Report 308, University of Washington. `https://stat.uw.edu/research/tech-reports/bootstrapping-z-estimators`.

Xie, M.-g. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39.

Zadeh, L. A. (1975). Fuzzy logic and approximate reasoning. *Synthese*, 30(3-4):407–428.