An efficient Monte Carlo method for valid prior-free possibilistic statistical inference

Ryan Martin*

February 8, 2025

Abstract

Inferential models (IMs) offer prior-free, Bayesian-like posterior degrees of belief designed for statistical inference, which feature a frequentist-like calibration property that ensures reliability of said inferences. The catch is that IMs' degrees of belief are possibilistic rather than probabilistic and, since the familiar Monte Carlo methods approximate probabilistic quantities, there are computational challenges associated with putting this framework into practice. The present paper addresses these challenges by developing a new Monte Carlo method designed specifically to approximate the IM's possibilistic output. The proposal is based on a characterization of the possibilistic IM's credal set, which identifies the "best probabilistic approximation" of the IM as a mixture distribution that can be readily approximated and sampled from. These samples can then be transformed into an approximation of the possibilistic IM. Numerical results are presented highlighting the proposed approximation's accuracy and computational efficiency.

Keywords and phrases: confidence distribution; credal set; Gaussian possibility; inferential model; variational inference.

1 Introduction

Numerous efforts to advance Fisher's vision of fiducial inference have been made over the years. Among these efforts, arguably "one of the original statistical innovations of the 2010s" (Cui and Hannig 2024) is the *inferential model* (IM) framework put forward in Martin and Liu (2013, 2015a,c) and later synthesized in Martin and Liu (2015b). What's unique about the IM framework is that its output takes the form of an imprecise probability—specifically, a possibility measure—and, consequently, inference is based on possibilistic reasoning: hypotheses assigned small possibility are refuted by the data while hypotheses whose complements are assigned small possibility are corroborated by the data. This shift from probability to possibility theory has principled motivations: the IM's possibilities satisfy a frequentist-style *validity* property that Bayesian/fiducial probabilities do not satisfy, i.e., the IM's possibility assigned to true hypotheses tends to be not small, hence it's a provably rare event that the data-driven IM output refutes a true

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

hypothesis or corroborates a false hypothesis. More details about IMs and their properties are provided in Section 2 below. Despite all that the IM framework has to offer, it has been slow to gain traction in the statistical community, largely due to computational challenges. Indeed, as shown recently in Jacob et al. (2021), the familiar Monte Carlo methods used to approximate ordinary probabilities simply aren't enough to approximate imprecise probabilities; something more is needed. The goal of this paper is to identify the aforementioned "something more" for the case of possibilistic IMs and to develop the corresponding Monte Carlo machinery needed to make these computations, and the IM framework more generally, readily available for everyday use in applications.

The jumping off point here is the fact that all (coherent) imprecise probabilities, including the possibilistic IM's output, correspond to a (non-empty, closed, and convex) set of ordinary or precise probabilities, called the *credal set*. Different brands of imprecise probabilities have their own mathematical properties, and one way this distinction manifests is in the kinds of constraints imposed on their associated credal sets. Possibility measures are among the simplest imprecise probabilities and, in turn, their credal sets have a relatively simple characterization. In particular, results like that in Theorem 1 below make it relatively easy to identify probability distributions in the credal set that "best approximate" the IM's possibilistic output. The best approximation is relatively simple mathematically but might still be difficult to compute; fortunately, further approximations can be made. Such second-level approximations can take various forms but, inspired by the asymptotic results in Martin and Williams (2024), here I'll use a family of Gaussian variational approximations as proposed in Cella and Martin (2024). These variational approximations on their own are not fully satisfactory, but here I show they can be stitched together in a satisfactorily way. Specifically, my proposal is to follow the guidelines suggested in the characterization result in Theorem 1 to construct a probability distribution, with certain Gaussian-like features, that is no more concentrated than this best probabilistic approximation of the IM's possibilistic output. The aforementioned Gaussian-like features make this easy and fast to sample from, thereby facilitating principled probabilistic approximations to the IM's output. I also show how this probabilistic approximation can be readily transformed into a possibility measure that is both easy to compute and accurately approximates the target IM output.

The remainder of this paper is organized as follows. Section 2 provides the necessary background on possibilistic IMs, highlighting the existing computational bottlenecks. The credal set characterization and proposed Monte Carlo strategy are presented in Section 3. Three illustrative examples, including logistic regression, are presented in Section 4. These are cases where direct-but-naive IM computations are expensive but not out of reach, so the accuracy of the proposed approximation can be assessed. As a more challenging application, Section 5 presents a non-trivial analysis of a small, censored data set involving a semiparametric model. For this application, the direct-but-naive IM solution is too expensive for practitioners to take seriously but the proposed Monte Carlo-based solution can be carried out in a matter of seconds. The insights and results presented here make up an important first step in a series of developments needed to create an IM toolbox for practitioners. Among other things, the concluding remarks in Section 6 describe my vision of where these developments can and need to go. Some relevant technical details are given in an appendix, including a description of the variational approximation developed in Cella and Martin (2024).

2 Background

The first IM developments relied on random sets and their corresponding belief functions. More recent developments in Martin (2022b), building on Martin (2015, 2018), define a possibilistic IM by applying a version of the probability-to-possibility transform to the model's relative likelihood. This shift is philosophically important, but the present review focuses on the details, properties, and computation of this possibilistic IM.

Consider a parametric model $\{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$ consisting of probability distribution supported on a sample space $\mathbb{Z}$, indexed by a parameter space $\mathbb{T} \subseteq \mathbb{R}^d$. Suppose that the observable data Z, taking values in $\mathbb{Z}$, is a sample from the distribution P_Θ, where $\Theta \in \mathbb{T}$ is the unknown/uncertain "true value." The model and observed data $Z = z$ together determine a likelihood function $\theta \mapsto L_z(\theta)$ and a corresponding relative likelihood

$$R(z, \theta) = \frac{L_z(\theta)}{\sup_\vartheta L_z(\vartheta)}.$$

I will implicitly assume here that the denominator is finite for almost all z. As is typical in the literature, I will also assume that prior information about Θ is vacuous.

The relative likelihood itself defines a data-dependent possibility contour, i.e., it is a non-negative function such that $\sup_\theta R(z, \theta) = 1$ for almost all z. This contour, in turn, determines a possibility measure that can be used for uncertainty quantification about Θ, given $Z = z$, which has been extensively studied in the literature (e.g., Denœux 2006, 2014; Shafer 1982; Wasserman 1990a). In particular, these references suggest assigning data-driven possibility values to hypotheses H about Θ via the rule

$$H \mapsto \sup_{\theta \in H} R(z, \theta), \quad H \subseteq \mathbb{T}.$$

This purely likelihood-driven possibility has a number of desirable properties, which I'll not get into here. What it lacks, however, is a justification for why the "possibilities" assigned to hypotheses about Θ have belief-forming inferential weight. With vacuous prior information, there's no Bayesian justification behind these possibility assignments, so justification can only come from a frequentist-like calibration property—i.e., assigning small possibilities to true hypotheses is a provably rare event—but the purely likelihood-based possibility assignment doesn't meet this requirement in a practically useful way. So, while the relative likelihood provides a natural, data-driven parameter ranking in terms of model fit, this is insufficient for principled and reliable statistical inference.

Fortunately, however, it is conceptually straightforward to achieve the desired calibration by applying what Martin (2022a) calls "validification"—a version of the probability-to-possibility transform (e.g., Dubois et al. 2004; Hose 2022). In particular, for observed data $Z = z$, the possibilistic IM's contour is defined as

$$\pi_z(\theta) = \mathsf{P}_\theta\big\{R(Z, \theta) \leq R(z, \theta)\big\}, \quad \theta \in \mathbb{T}, \tag{1}$$

and the possibility measure—or upper probability—is likewise defined as

$$\overline{\Pi}_z(H) = \sup_{\theta \in H} \pi_z(\theta), \quad H \subseteq \mathbb{T}. \tag{2}$$

It won't be needed in the present paper, but there's a corresponding necessity measure, or lower probability, defined via conjugacy: $\underline{\Pi}_z(H) = 1 - \overline{\Pi}_z(H^c)$. An essential feature of this IM construction is its so-called *validity property*:

$$\sup_{\Theta \in \mathbb{T}} \mathsf{P}_\Theta\big\{\pi_Z(\Theta) \leq \alpha\big\} \leq \alpha, \quad \text{for all } \alpha \in [0,1]. \tag{3}$$

This has a number of important consequences. First, (3) immediately implies that

$$C_\alpha(z) = \{\theta \in \mathbb{T} : \pi_z(\theta) \geq \alpha\}, \quad \alpha \in [0,1] \tag{4}$$

is a $100(1-\alpha)\%$ frequentist confidence set, i.e., $\sup_{\Theta \in \mathbb{T}} \mathsf{P}_\Theta\big\{C_\alpha(Z) \not\ni \Theta\big\} \leq \alpha$. Second, from (2) and (3), it readily follows that

$$\sup_{\Theta \in H} \mathsf{P}_\Theta\big\{\overline{\Pi}_Z(H) \leq \alpha\big\} \leq \alpha, \quad \text{all } \alpha \in [0,1], \text{ all } H \subseteq \mathbb{T}. \tag{5}$$

In words, a valid IM assigns possibility $\leq \alpha$ to true hypotheses at rate $\leq \alpha$ as a function of data Z. This gives the IM its "inferential weight"—(5) implies that $\overline{\Pi}_z(H)$ is not expected to be small when H is true, so one is inclined to doubt the truthfulness of a hypothesis H if $\overline{\Pi}_z(H)$ is small. Third, the above property ensures that the possibilistic IM is safe from false confidence (Balch et al. 2019; Martin 2019, 2024b), unlike all default-prior Bayes and fiducial solutions. An even stronger, *uniform-in-hypotheses* version of (5) holds, which offers opportunities for the data analyst to do more with the IM than test prespecified hypotheses (Cella and Martin 2023). For further details about possibilistic IMs' properties, including de Finetti-style no-sure-loss properties, its connection to Bayesian/fiducial inference, etc, see Martin (2022b, 2023a,b).

In a Bayesian analysis, inference is based on summaries of the posterior distribution, e.g., posterior probabilities of scientifically relevant hypotheses, expectations of loss/utility functions, etc. All of these summaries boil down to integration involving the probability density function that determines the posterior. Virtually the same statement holds for the possibilistic IM: the lower–upper probability pairs for scientifically relevant hypotheses, lower–upper expectations of loss/utility functions (Martin 2021), etc. involve optimization of the possibility contour π_z. More precisely, the proper calculus for possibilistic reasoning is Choquet integration (e.g. Troffaes and de Cooman 2014, App. C), which goes as follows. If $h : \mathbb{T} \to \mathbb{R}$ is a non-negative function, then the Choquet integral with respect to the possibility measure $\overline{\Pi}_z$ is defined as

$$\overline{\Pi}_z h := \int_{\inf h}^{\sup h} \Big\{ \sup_{\theta : h(\theta) > s} \pi_z(\theta) \Big\} \, ds. \tag{6}$$

As an important special case, suppose that $h(\theta) = 1(\theta \in H)$, where $H = \{\theta : k(\theta) \in K\}$ for some function k and some subset $K \subseteq k(\mathbb{T})$. Then

$$\overline{\Pi}_z h = \sup_{\theta : h(\theta) = 1} \pi_z(\theta) = \sup_{\theta \in H} \pi_z(\theta) = \sup_{\kappa \in K} \sup_{\theta : k(\theta) = \kappa} \pi_z(\theta).$$

There's a clear analogy that can be made between the right-hand side of the above display and the more familiar Bayesian calculus: the degree of possibility/probability assigned to "$\Theta \in H$" or, equivalently, to "$k(\Theta) \in K$" is obtained by first getting the marginal

contour/density for $k(\Theta)$ at κ by optimizing/integrating over $\{\theta : k(\theta) = \kappa\}$ and then optimizing/integrating over $\kappa \in K$. Unlike with Bayesian integration, however, the IM's optimization operation ensures that the validity property inherent in π_z is transferred to the marginal IM for $k(\Theta)$, which implies that it's safe from false confidence.

While the IM construction is conceptually simple and its properties are quite strong, computation can be a challenge. The key observation is that the sampling distribution of the relative likelihood $R(Z, \theta)$, under P_θ, is rarely available in closed-form to facilitate exact computation of π_z. So, instead, the go-to strategy is to approximate that sampling distribution using Monte Carlo at each value of θ on a sufficiently fine grid (e.g., Hose et al. 2022; Martin 2022b). That is, the possibility contour is approximated as

$$\pi_z(\theta) \approx \frac{1}{M} \sum_{m=1}^{M} 1\{R(Z_{m,\theta}, \theta) \leq R(z, \theta)\}, \quad \theta \in \mathbb{T}, \tag{7}$$

where $Z_{m,\theta}$ are independent copies of the data Z, specifically drawn from P_θ, for $m = 1, \ldots, M$. The above computation is feasible at one or a few different θ values, but this often needs to be carried out over a fine grid covering the relevant portion of the parameter space $\mathbb{T}$. For example, identifying the confidence set in (4) requires finding which θ's satisfy $\pi_z(\theta) \geq \alpha$, and a naive approach is to compute the contour over a huge grid and then keep those that (approximately) meet the aforementioned condition. This amounts to lots of wasted and expensive computations. More generally, the relevant summaries of the IM output involve optimization, as in (6), and doing so numerically requires many contour function evaluations. This is a serious bottleneck, so new and not-so-naive computational strategies are desperately needed.

3 Monte Carlo methods for IMs

3.1 Key insights

Given a possibilistic IM $z \mapsto (\underline{\Pi}_z, \overline{\Pi}_z)$ with contour function π_z, I'll refer to the level sets $C_\alpha(z) = \{\theta \in \mathbb{T} : \pi_z(\theta) \geq \alpha\}$ defined in (4) as α-cuts. According to the theory reviewed in Section 2, the α-cut $C_\alpha(z)$ is a nominal $100(1-\alpha)\%$ confidence region for Θ. Aside from the practical utility of confidence regions, there's much more information in the IM's α-cuts. Towards this, the credal set $\mathscr{C}(\overline{\Pi}_z)$ corresponding to the IM's $\overline{\Pi}_z$ is the set of (possibly data-dependent) probability measures that it dominates, i.e.,

$$\mathscr{C}(\overline{\Pi}_z) = \{\mathsf{Q}_z \in \text{probs}(\mathbb{T}) : \mathsf{Q}_z(H) \leq \overline{\Pi}_z(H) \text{ for all measurable } H\},$$

where probs($\mathbb{T}$) is the set of probability measures on $\mathbb{T}$. Then the claim is that the α-cuts determine the IM's credal set. Indeed, there's the following well-known characterization (e.g., Couso et al. 2001; Destercke and Dubois 2014) of the credal set $\mathscr{C}(\overline{\Pi}_z)$:

$$\mathsf{Q}_z \in \mathscr{C}(\overline{\Pi}_z) \iff \mathsf{Q}_z\{C_\alpha(z)\} \geq 1-\alpha \quad \text{for all } \alpha \in [0,1]. \tag{8}$$

That is, a probability distribution Q_z belongs to the IM's credal set if and only if it's a confidence distribution, i.e., it assigns probability $\geq 1-\alpha$ to each of the IM's α-cuts. More about this notion of "confidence distributions" and when Bayesian/fiducial distributions

achieve equality in (8) can be found in Martin (2023a). At a soon-to-be-demonstrated practical level, elements of the IM's credal set have a relatively simple characterization, as the following theorem demonstrates. The result is presented in a more general, not-IM-specific context because this simplifies the notation.

Theorem 1. *Let $\overline{\Pi}$ be a possibility measure on a space $\mathbb{T}$, with corresponding contour π and α-cuts $C_\alpha = \{\theta \in \mathbb{T} : \pi(\theta) \geq \alpha\}$, $\alpha \in (0,1)$. Then $\mathsf{Q} \in \mathscr{C}(\overline{\Pi})$ if and only if*

$$\mathsf{Q}(\cdot) = \int_0^1 \mathsf{K}^\beta(\cdot)\, \mathsf{M}(d\beta), \tag{9}$$

for some Markov kernel K^β indexed by $\beta \in [0,1]$ such that K^β is fully supported on C_β, i.e., $\mathsf{K}^\beta(C_\beta) = 1$ for each $\beta \in [0,1]$ and some probability measure M on $[0,1]$ such that a random variable with distribution M is stochastically no smaller than $\mathsf{Unif}(0,1)$.

Proof. See Appendix A.1. □

I would be surprised if the characterization in Theorem 1 is genuinely new, but I've not found a reference for this result exactly. A similar result is presented in Wasserman (1990b, Theorem 2.1) characterizing the credal set of a belief function determined by a given distribution and set-valued mapping. Hose (2022, Eq. 2.44) also gives a special case of the right-hand side of (9) but makes no claims that his version is a complete characterization of $\mathscr{C}(\overline{\Pi})$; what I specifically propose to do in (12) below closely matches the Hose's formula. As an aside, choosing the kernel K^β to be a uniform distribution supported on C_β is an idea that commonly appears in the literature, e.g., the *pignistic probability* (Smets and Kennes 1994) or the *Shapley value* (Shapley 1953) of a game.

The above theorem describes the contents of the credal set $\mathscr{C}(\overline{\Pi})$ corresponding to a possibility measure $\overline{\Pi}$. A relevant follow-up question is if there exists an element in $\mathscr{C}(\overline{\Pi})$ that's "maximally consistent" with $\overline{\Pi}$, i.e., a distribution $\mathsf{Q}^\star$ such that

$$\text{equality in (8) holds: } \mathsf{Q}^\star(C_\alpha) = 1-\alpha \text{ for each } \alpha \in [0,1]. \tag{10}$$

Such a probability distribution will be called an *inner probabilistic approximation* of $\overline{\Pi}$. The theorem's proof offers some insights that help to answer this question. Indeed, the two properties needed to achieve the equality in (10) are, first, that the kernel satisfies $\mathsf{K}^\beta(C_\alpha) = 0$ for all pairs (α, β) with $\beta < \alpha$ and, second, that $\mathsf{M} = \mathsf{Unif}(0,1)$.

As an illustration, which will prove to be useful in Section 3.3 below, consider the Gaussian possibility measure $\overline{\Pi}$, with corresponding mean vector m and covariance matrix V, defined in Martin and Williams (2024) with contour

$$\pi(\theta) = 1 - F_d\{(\theta - m)^\top V^{-1}(\theta - m)\}, \quad \theta \in \mathbb{T} = \mathbb{R}^d,$$

where F_d is the $\mathsf{ChiSq}(d)$ distribution function. Then the corresponding α-cuts are given by the d-dimensional ellipsoids

$$C_\alpha = \{\theta \in \mathbb{R}^d : (\theta - m)^\top V^{-1}(\theta - m) \leq F_d^{-1}(1-\alpha)\}, \quad \alpha \in [0,1].$$

Of course, the normal distribution, $\mathsf{N}_d(m, V)$, assigns probability $1-\alpha$ to each α-cut, so that must be the inner probabilistic approximation; in fact, this is a tautology since the

Gaussian possibility measure above is defined as the outer possibilistic approximation of the Gaussian probability distribution (Martin and Williams 2024). Following the guidance offered in the previous paragraph, one first finds that the marginal distribution M of $\pi(\Theta)$ is $\mathsf{Unif}(0,1)$ under $\Theta \sim \mathsf{N}_d(m, V)$. Second, since a Gaussian random vector—or any elliptically symmetric random vector for that matter (e.g., Hult and Lindskog 2002, Theorem 3.1)—can be written as $\Theta = m + RV^{1/2}U$, where $V^{1/2}$ is the Cholesky factor of V, (R, U) are independent, with U a uniform random vector on the unit sphere in $\mathbb{R}^d$ and R a non-negative random variable, the corresponding kernel K^β is the distribution of $m + \{F_d^{-1}(1-\beta)\}^{1/2} V^{1/2} U$, which is fully supported on the boundary ∂C_β of C_β.

3.2 Probabilistic approximation to a possibilistic IM

Suppose I have at my disposal a collection of data-dependent probability distributions, $\{\mathsf{Q}_z^\alpha : \alpha \in [0,1]\}$, supported on $\mathbb{T}$, with the property

$$\mathsf{Q}_z^\alpha\{C_\alpha(z)\} \geq 1 - \alpha, \quad \alpha \in [0,1], \tag{11}$$

where $C_\alpha(z)$ is the possibilistic IM's α-cut. Where specifically these Q_z^α's might come from will be discussed in Section 3.3 below and in more detail in Appendix A.2. Note that the Q_z^α's aren't confidence distributions because no guarantees are offered concerning the Q_z^α-probability assigned to $C_\beta(z)$ for $\beta \neq \alpha$. Following the intuition provided by Theorem 1 and the subsequent example, I propose to stitch these α-dependent probability distributions together into a single probability distribution, $\mathsf{Q}_z^\star$, using the mixture operation in (9). More specifically, I propose to take the marginal distribution M to be $\mathsf{Unif}(0,1)$ and the kernel/conditional distribution as follows: let $\Theta \sim \mathsf{Q}_z^\alpha$ and set the kernel $\mathsf{K}^\alpha \equiv \mathsf{K}_z^\alpha$, which now depends on data z, to be (a version of) the conditional distribution of Θ, given $\Theta \in \partial C_\alpha(z)$, i.e.,

$$\mathsf{K}_z^\alpha(H) := \mathsf{Q}_z^\alpha\{\Theta \in H \cap C_\alpha(z) \mid \Theta \in \partial C_\alpha(z)\}.$$

Then my proposed probabilistic approximation of the possibilistic IM is based on stitching the α-specific probabilities according to the following rule:

$$\mathsf{Q}_z^\star(\cdot) = \int_0^1 \mathsf{K}_z^\alpha(\cdot)\, d\alpha. \tag{12}$$

Of course, evaluating features of or simulating from the distribution $\mathsf{Q}_z^\star$ in (12) could be challenging in general but, fortunately, this is relatively easy to do in the case of my recommended approach described in Section 3.3 below. At least intuitively, however, it's clear how to proceed from (12) to evaluate the proposed probabilistic approximation of the IM: samples from $\mathsf{Q}_z^\star$ can be obtained by first sampling $A \sim \mathsf{Unif}(0,1)$ and then by sampling Θ from the conditional distribution Q_z^A that's concentrated on the boundary $\partial C_A(z)$. If there's a specific feature of $\mathsf{Q}_z^\star$ to be evaluated, then it may be advantageous to employ a Rao–Blackwellization strategy to reduce the Monte Carlo variance. That is, suppose the goal is to evaluate the expected value, say, $\mathsf{Q}_z^\star h$, of some real-valued function h defined on $\mathbb{T}$, such as $h = 1_H$ for a relevant hypothesis H. Then first get $\alpha \mapsto \mathsf{K}_z^\alpha h$, the expected value of $h(\Theta)$ relative to the conditional distribution K_z^α using Monte Carlo, and then integrate over α numerically, e.g., using R's `integrate` function.

Keep in mind that the possibilistic IM is the target, and even the "best" probabilistic approximation can be far off from the target in its answers to certain questions. In general, all that's guaranteed is $\mathsf{Q}_z^\star(H) \leq \overline{\Pi}_z(H)$, and equality holds effectively only when H is the complement of an α-cut. Similar statements can be made for expectations of general functions h, but it's a bit more complicated: basically, the probabilistic $\mathsf{Q}_z^\star$-expectations agree with the possibilistic $\overline{\Pi}_z$-upper expectations when h has level sets that match the IM's collection of α-cuts. Section 3.5 below offers an approach to overcome the inherent limitations of any probabilistic approximation of the possibilistic IM.

3.3 Implementation details

Cella and Martin (2024) proposed introducing a parametric family, say, $\{\mathsf{R}_z^\xi : \xi \in \Xi\}$ of probability distributions on $\mathbb{T}$ and, akin to variational Bayes analysis (e.g., Blei et al. 2017), choosing a particular value $\xi = \xi(z, \alpha)$ of the parameter—depending on data z and level α—such that $\mathsf{Q}_z^\alpha := \mathsf{R}_z^{\xi(z,\alpha)}$ assigns at least $1-\alpha$ probability to the α-cut $C_\alpha(z)$, as required in (11). Intuitively, if ξ controls the spread of R_z^ξ, and if this distribution is properly centered and oriented, then there exists a value $\xi(z, \alpha)$ such that the Q_z^α defined in the previous sentence has the stated property, at least approximately. As the reader might expect, some problem-specific considerations are required to efficiently implement this strategy, but it can be done accurately in a wide range of applications as demonstrated next and in the reference cited above.

For the relative likelihood-based possibilistic IM described in Section 2, under the usual regularity conditions, Martin and Williams (2024) established an asymptotic Gaussianity result that says, if the sample size is large, then

$$\pi_z(\theta) \approx 1 - F_d\big((\theta - \hat{\theta}_z)^\top J_z (\theta - \hat{\theta}_z)\big), \quad \theta \in \mathbb{T} \subseteq \mathbb{R}^d, \tag{13}$$

where $\hat{\theta}_z$ is the maximum likelihood estimator, J_z is the $d \times d$ observed Fisher information matrix, and, again, F_d is the $\mathsf{ChiSq}(d)$ distribution function. This Gaussian approximation also holds in alternative parametrizations and, in fact, it might be more accurate in certain parametrizations than in others. For example, if there are non-negativity constraints on θ, then the Gaussian approximation might be more appropriate/accurate when applied on the $\log\theta$ scale than on the θ scale. My general presentation below works on the θ scale, but I will apply these transformations freely in the examples that follow.

Since the best probabilistic approximation of the Gaussian possibility measure is the familiar Gaussian distribution, a reasonable choice of the variational family is

$$\mathsf{R}_z^\xi = \mathsf{N}_d\big(\hat{\theta}_z,\, J_z^{-1}(\xi)\big), \quad \xi \in \Xi = (0, \infty)^d \tag{14}$$

where, if $J_z = E \Lambda E^\top$ is the spectral decomposition of J_z, then

$$J_z(\xi) = E \,\mathrm{diag}(\xi^{-1})\, \Lambda \,\mathrm{diag}(\xi^{-1})\, E^\top,$$

for ξ^{-1} the entry-wise reciprocal of ξ, and $\mathrm{diag}(\cdot)$ the operator that takes its vector argument to a diagonal matrix with that vector on the diagonal. Now it should be clear what was meant when I said above that "ξ controls the spread of R_z^ξ." With this family of candidate probabilistic approximations to the possibilistic IM, Cella and Martin (2024)

proposed a stochastic approximation algorithm (see Appendix A.2 below) that, for a given α, selects $\xi = \xi(z, \alpha)$ such that

$$\mathsf{R}_z^{\xi(z,\alpha)}\{C_\alpha(z)\} \geq 1 - \alpha, \quad \text{at least approximately,}$$

where $\mathsf{R}_z^{\xi(z,\alpha)}$ is the Gaussian distribution in (14) with $\xi = \xi(z, \alpha)$ plugged into the variance and $C_\alpha(z)$ is still the original IM's α-cut. More specifically, if the variational family's α-cuts are defined as

$$\begin{aligned} C_\alpha^\xi(z) &= \{\theta : 1 - F_d\big((\theta - \hat{\theta}_z)^\top J_z(\xi)(\theta - \hat{\theta}_z)\big) \geq \alpha\} \\ &= \{\theta : (\theta - \hat{\theta}_z)^\top J_z(\xi)(\theta - \hat{\theta}_z) \leq F_d^{-1}(1 - \alpha)\}, \end{aligned} \tag{15}$$

then Cella and Martin's algorithm returns $\xi = \xi(z, \alpha)$ such that

$$K_\alpha^{\xi(z,\alpha)}(z) \supseteq C_\alpha(z).$$

They basically stopped here, suggesting $\mathsf{Q}_z^\alpha = \mathsf{R}_z^{\xi(z,\alpha)}$ as a limited-but-simple probabilistic approximation of the possibilistic IM's output $\overline{\Pi}_z$ with the property of being a confidence distribution at least for the specified confidence level α.

To move beyond the insights in Cella and Martin, the key two-fold observation here is as follows: first, $C_\alpha^\xi(z)$ has a simple expression and geometric form (ellipsoid) and, second, it is easy to condition the Gaussian distribution Q_z^α to the boundary $\partial C_\alpha^{\xi(z,\alpha)}(z)$—this is exactly what was shown in the Gaussian example at the end of Section 3.1. Of course, bigger α-cuts imply more conservative inference, but empirical evidence in Cella and Martin (2024) and Section 4 below suggests that the approximation is quite accurate, i.e., the two α-cuts in the above display are nearly the same. Just a slight loss of statistical efficiency is a small price to pay for replacing the impossible task of conditioning on $\partial C_\alpha(z)$ with the easy task of conditioning on $\partial C_\alpha^{\xi(z,\alpha)}(z)$.

To summarize, I propose to implement the probabilistic approximation $\mathsf{Q}_z^\star$ in (12) by iterating the following two steps:

1. Sample $A \sim \mathsf{Unif}(0, 1)$;
2. Given A, evaluate $\xi(z, A)$, and then sample Θ just as described in the Gaussian example at the end of Section 3.1, i.e.,

$$(\Theta \mid A) \overset{\text{dist}}{=} \hat{\theta}_z + \{F_d^{-1}(1 - A)\}^{1/2} \, J_z^{-1}(\xi(z, A))^{1/2} \, U,$$

   where half-power of a matrix means the Cholesky factor and U is uniformly distributed on the sphere in $\mathbb{R}^d$.

Of course, both sampling steps above are trivial; the only expensive operation—"evaluate $\xi(z, A)$"—is almost hidden in the description. The evaluation of $\xi(z, A)$ involves an A-specific run of the stochastic approximation algorithm summarized in Appendix A.2 which, fortunately, typically only involves a few updates and a few Monte Carlo evaluations of the IM contour π_z per update. So, this is fast and easy to compute, and it scales linearly in dimension. Moreover, it's easy to parallelize the evaluation of $\xi(z, A)$ over multiple A's, dramatically reducing computation time.

A practical adjustment to the above procedure, which allows for larger sample sizes from $\mathsf{Q}_z^\star$ with even less computation time, is as follows. Start with a fixed grid $\mathcal{A}$ of, say 100 α values; in my examples that follow, I use an equally-spaced grid between 0.001 and 0.999. Then evaluate $\xi(z, \alpha)$ for all $\alpha \in \mathcal{A}$, which sets a fixed limit on the amount of computational investment required. Now, for each $A \sim \mathsf{Unif}(0, 1)$ drawn in Step 1 above, find the two α values in $\mathcal{A}$ that sandwich A and get $\xi(z, A)$ via linear interpolation, and then do Step 2 as described above. Once the 100 values of $\xi(z, \cdot)$ are obtained, which takes roughly 5 seconds in my examples below, the samples from $\mathsf{Q}_z^\star$ are virtually free.

Finally, the procedure proposed here has certain aspects in common with the *calibrated bootstrap* algorithm developed in Jiang et al. (2023); in fact, the results in this paper and in Cella and Martin (2024) were inspired by the calibrated bootstrap efforts. On the one hand, the calibrated bootstrap proposal is more general than that here because it doesn't require specification of a parametric/Gaussian family of approximate distributions—it uses bootstrap to learn a full distribution nonparametrically. On the other hand, learning the full distribution requires computational effort and, furthermore, that full distribution is not well-suited for the condition-to-the-boundary step. Consequently, the calibrated bootstrap will spend considerable computational resources to evaluate the contour π_z at points which will ultimately be discarded from the sample that describes $\mathsf{Q}_z^\star$.

3.4 Asymptotic analysis

Here I offer a brief large-sample analysis to justify the claimed overall accuracy of the proposed probabilistic approximation. For the sake of brevity, and without loss of persuasiveness, I give only a heuristic argument building on the asymptotic convergence results rigorously demonstrated in Martin and Williams (2024).

As mentioned above, Martin and Williams (2024) show that, under the regularity conditions sufficient for asymptotic normality and efficiency of the maximum likelihood estimator, the relative likelihood-based possibilistic IM enjoys a large-sample Gaussianity property, akin to the classical Bernstein–von Mises theorem fundamental to Bayesian analysis; see (13). The uniform mode of their convergence result implies, first, that the IM's α-cuts, $C_\alpha(z)$, will merge with the α-cuts, $C_\alpha^1(z)$, of the variational approximation, with $\xi = 1$, as the size of the sample z increases to ∞. Second, since $\xi(z, \alpha)$ is designed to make the variational family's α-cuts agree with the original IM's, one fully expects that the components of $\xi(z, \alpha)$ are converging to 1 as the sample size increases.

Putting everything together, the original IM's α-cuts merge with the corresponding Gaussian α-cuts asymptotically, and the components of the variational family index $\xi(z, \alpha)$ identified by Cella and Martin's algorithm are converging to 1. Therefore, the symmetric difference $C_\alpha(z) \triangle C_\alpha^{\xi(z,\alpha)}(z)$ is converging to $\varnothing$, which implies that the proposed sampling algorithm exactly recovers the IM's limiting inner probabilistic approximation, which is the Gaussian distribution just like in the example in Section 3.1.

3.5 Back to a possibilistic IM

The starting point was a possibilistic IM with contour π_z that generally can only be evaluated using Monte Carlo by sampling artificial data sets from the underlying statistical model. This is cheap and easy to do at a few parameter values, but becomes overwhelm-

ingly expensive when the contour evaluations are needed at all parameter points on a sufficiently fine grid. Unfortunately, virtually every relevant summary of the IM output requires this evaluation on a fine grid. What was proposed above was a probabilistic approximation $\mathsf{Q}_z^\star$ of the possibilistic IM $\overline{\Pi}_z$, which could be sampled with relative ease. Then those relevant summaries, e.g., expectations, could be readily approximated using the samples from $\mathsf{Q}_z^\star$, and subsequently used to approximate the IM's summaries. The challenge is that, for each summary, there are mathematical limits to how well the probabilistic approximation can match the target possibilistic one; for some summaries, the approximation is accurate while for others it's not. The question here is if there's a way to convert the probabilistic approximation $\mathsf{Q}_z^\star$ to a possibilistic approximation that better represents the target IM while retaining the computational efficiency.

The answer to this question is *Yes*. The "inner probabilistic approximation" just described can be interpreted as a possibility-to-probability transform, but a probability-to-possibility transform goes the other direction. Let $r_z : \mathbb{T} \to \mathbb{R}$ be a (possibly data-dependent) ranking function on the parameter space, where larger values of $r_z(\theta)$ indicate that θ is "higher ranked" in some meaningful sense. Good examples of ranking functions are, first, the density function $q_z^\star$ corresponding to the distribution $\mathsf{Q}_z^\star$ and, second, the likelihood function L_z of the underlying model. With both of these choices, the meaningfulness of the ranking is clear. Taking the ranking function to be the density $q_z^\star$ gives the tightest contours, similar to how highest density sets have smallest volume among those with a fixed probability content, but "tightest contours" isn't the objective. Moreover, there may be computational challenges associated with use of the density-based ranking. After all, $q_z^\star$ would need to be estimated using, say, kernel methods applied to the samples from $\mathsf{Q}_z^\star$ as described above, which is non-trivial in the multiparameter case. See below for more on how the ranking function affects the approximation accuracy.

In any case, for a given ranking function r_z, the probability-to-possibility transform of $\mathsf{Q}_z^\star$ returns the possibility contour

$$\omega_z(\theta) = \mathsf{Q}_z^\star\big\{r_z(\Theta) \leq r_z(\theta)\big\}, \quad \theta \in \mathbb{T}. \tag{16}$$

To see that $\sup_\theta \omega_z(\theta) = 1$ and, hence, that this transformation defines a genuine possibility contour, just take θ equal (or converging) to a "highest ranked" value according to r_z. Given that ω_z is a genuine possibility contour, it makes sense to define the corresponding possibility measure via optimization as before, i.e., $\overline{\Omega}_z(H) = \sup_{\theta \in H} \omega_z(\theta)$. In the examples that follow, I'll refer to the approximation ω_z of π as the *stitched IM contour*, since it's obtained by stitching together the α-cut approximations.

Computationally, the contour ω_z in (16) is far more efficient than that in (7). The reason being that the Monte Carlo samples in the former are fixed whereas, in the latter, the samples generally depend on the θ at which the contour is being evaluated. It's also worth asking if ω_z is a good approximation of π_z. Towards this, if $\mathsf{Q}_z^\star$ is a genuine inner probabilistic approximation of $\overline{\Pi}_z$, i.e., if (10) holds exactly, then it's easy to see that

$$\mathsf{Q}_z^\star\{\pi_z(\Theta) \leq \pi_z(\theta)\} = \pi_z(\theta), \quad \theta \in \mathbb{T},$$

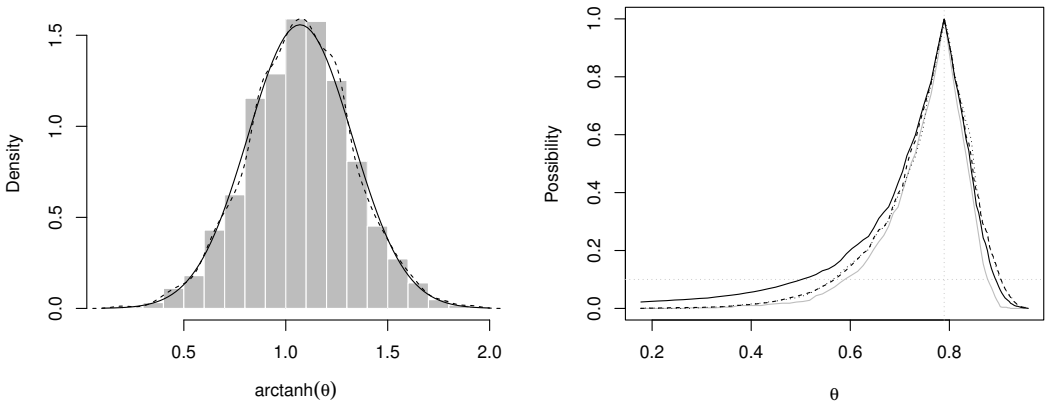
where the left-hand side above is $\omega_z(\theta)$ corresponding to the ranking function $r_z = \pi_z$. So, with a proper choice of ranking function, it's theoretically possible to recover the original IM from this. The downside is that, of course, if π_z was available to serve as a ranking

function, then there'd be no need for any of these approximations. So, it's unrealistic to expect that the stitched IM with contour ω_z will exactly recover the original IM. That said, however, since the original IM is driven by the relative likelihood function, it follows that the π_z-ranking and the L_z-ranking are very similar. In fact, they are exactly the same if the relative likelihood $R(Z, \theta)$ is a pivot under P_θ, which is the case asymptotically for all regular models. To summarize, there is a sense in which the likelihood-based ranking is preferred, but this doesn't mean the likelihood-based version of the stitched IM contour ω_z in (16) exactly matches that of the original IM. In the large-sample context discussed in Section 3.4 above, however, one can expect ω_z and π_z to merge asymptotically, for both the likelihood- and density-based rankings.

4 Examples

The illustrative examples below are low-dimensional and, therefore, the naive approximation (7) of the IM contour is computationally feasible. So, my goal here is to offer some further computational details concerning the new proposal in Section 3 and to highlight its approximation quality, efficiency, etc. R code for the gamma illustration in Example 2 below is available at `https://www4.stat.ncsu.edu/~rgmarti3/software`; some more versatile and user-friendly software is currently in the works.

Example 1. One of the more challenging inference problems involving a one-parameter model is the bivariate normal with known means and variances but unknown correlation. Suppose that $Z = (Z_1, \ldots, Z_n)$ are iid, with $Z_i = (X_i, Y_i)$ a bivariate normal random vector with zero means, unit standard deviations, and correlation $\Theta \in \mathbb{T} = (-1, 1)$ to be inferred. Interestingly, that the means and variances are known makes the problem more difficult—it's a curved exponential family so the minimal sufficient statistic is not complete and there are various ancillary statistics available to condition on (e.g., Basu 1964). How this affects asymptotic inference is detailed in Reid (2003). It is straightforward to construct a possibilistic IM (Martin 2024a, Example 2), which is exactly valid for all sample sizes and asymptotically efficient, but computation of the naive approximation (7) is relatively expensive because there's no pivotal structure and no closed-form expression for the maximum likelihood estimator. The variational approximation proposed in Cella and Martin (2024) is fast and easy, but it loses the exact validity of the original IM. Here I apply the proposed Monte Carlo sampling strategy on Fisher's transformation scale, i.e., on $\Psi = \text{arctanh}(\Theta)$, the inverse hyperbolic tangent. Figure 1(a) shows a histogram of 5000 samples from the distribution $\mathsf{Q}_z^\star$ of Ψ as described above, based on (a centered and scaled version of) the law school admissions data analyzed in Efron (1982), where $n = 15$ and the maximum likelihood estimator of Θ is $\hat{\theta}_z = 0.789$. Overlaid on this plot is, first, a normal density with mean $\text{arctanh}(0.789) = 1.07$ and estimated standard deviation and, second, the kernel density estimate; the difference between these two estimates is negligible. Panel (b) shows the exact IM contour and three stitched IM approximations as in (16) corresponding to three different ranking functions r_z. All three approximations have roughly the same shape as the true contour, as expected. The two based on density estimates—Gaussian and kernel—are quite similar and closely agree with the exact contour. The likelihood-based approximation is very accurate on the right-hand side but a bit conservative on the left-hand side; this is because, as in Figure 1(a), the distribution



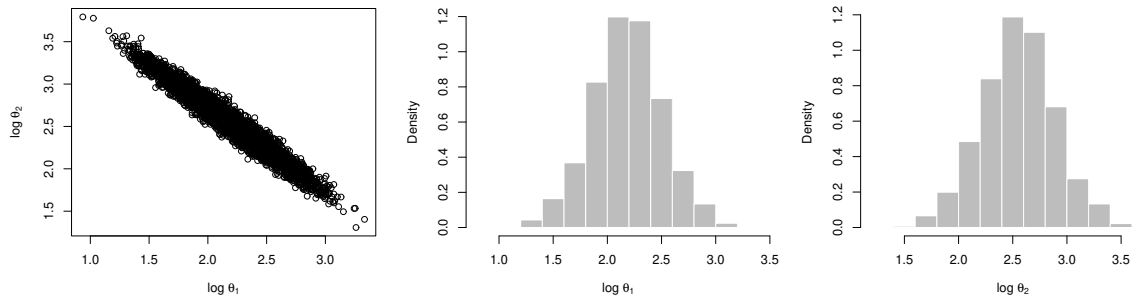
(a) Samples from $\mathsf{Q}_z^\star$ (b) Possibility contours

Figure 1: Summary of the results in Example 1 for the bivariate normal correlation. In Panel (a), the solid line is the Gaussian density and dashed line is the kernel density estimate from R's `density`. In Panel (b), the gray line is the exact IM contour, the dashed and dotted lines are approximations using the Gaussian and kernel density rankings, respectively, and the solid line is the approximation based on the likelihood ranking.

$\mathsf{Q}_z^\star$ of the Fisher-transformed Θ is symmetric but the likelihood-based ranking is not.

Example 2. Let $Z = (Z_1, \ldots, Z_n)$ denote an iid sample from a gamma distribution with unknown parameter $\Theta = (\Theta_1, \Theta_2)$, where $\Theta_1 > 0$ and $\Theta_2 > 0$ are the unknown shape and scale parameters, respectively. It is straightforward to get the naive approximation (7) at any particular value of the parameter, but prohibitively expensive to carry this out over a sufficiently fine grid that spans the plausible pairs (θ_1, θ_2). So, there's a need for more computationally efficient approximation methods, and the Monte Carlo strategy of Section 3 above fits the bill. For this illustration, I'll work with the data presented in Example 3 of Fraser et al. (1997), which consists of the survival time (in weeks) for $n = 20$ rats exposed to a certain amount of radiation, modeled by a gamma distribution with unknown shape and scale parameters. Figure 2 summarizes the 5000 samples of $(\log \Theta_1, \log \Theta_2)$ from the proposed probabilistic approximation $\mathsf{Q}_z^\star$ developed in Section 3. This reveals that the distribution is at least approximately Gaussian. Figure 3(a) shows the exact joint contour function for (Θ_1, Θ_2) along with two stitched IM approximations: one takes the ranking function to be a bivariate Gaussian density function and the other takes it to be the gamma likelihood function. Notice that the likelihood-ranking-based approximation almost perfectly matches the exact contour. And for comparison, the exact contour requires more than 10 minutes of computation time, while the new Monte Carlo approximation is completed in roughly 10 seconds.

A practically relevant and surprisingly challenging follow-up question concerns inference on the mean $\Phi = \Theta_1\Theta_2$ of the gamma distribution. IM solutions that offer exactly valid inference on Φ are presented in Martin and Liu (2015c) and Martin (2023b), but the relevant computations can be a burden. An obvious idea is that, given a sample of $(\log \Theta_1, \log \Theta_2)$ values from the probabilistic approximation, it's straightforward to get a corresponding sample of $\Phi = \exp(\log \Theta_1 + \log \Theta_2)$ values, but would this provide a



(a) Samples from $\mathsf{Q}_z^\star$ (b) Histogram of $\log \Theta_1$ (c) Histogram of $\log \Theta_2$

Figure 2: Summary of the samples $(\log \Theta_1, \log \Theta_2)$ from the probabilistic approximation $\mathsf{Q}_z^\star$ of the IM in Example 2.

good approximation of the marginal IM for Φ presented in Martin (2023b)? To answer this question, Figure 3(b) shows the latter exact, profile likelihood-based, marginal IM possibility contour for the mean Φ and two approximations thereof:

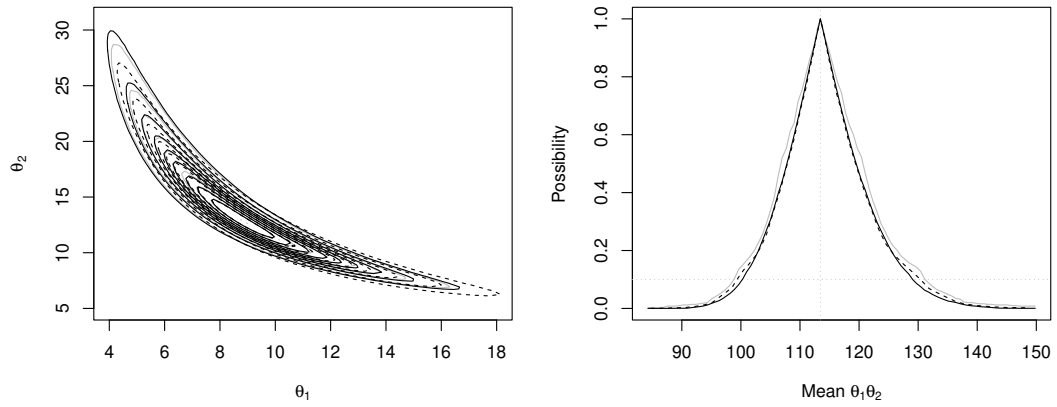
- an *indirect* approximation based on the samples of $(\log \Theta_1, \log \Theta_2)$ from $\mathsf{Q}_z^\star$ as described above, using the profile likelihood as the ranking function, and
- a *direct* approximation wherein the mean Φ is sampled from the inner probabilistic approximation of the target marginal IM.

From the context alone, the reader should expect that the direct approximation would be superior to the indirect approximation, but this superiority doesn't come for free: the former is tailored to the particular choice of feature Φ and, therefore, is not as simple to obtain as the latter. From the plot in Figure 3(b), it's clear that both approximations are quite accurate and, indeed, the direct approximation is more accurate overall, with the indirect approximation being a bit too narrow compared to the target. This narrowness is inevitable because the marginal distribution for Φ derived from $\mathsf{Q}_z^\star$ generally would not be the inner probabilistic approximation of the marginal IM. Whatever this approximation might lack in terms of accuracy, it makes up for in simplicity.

Example 3. The data presented in Table 8.4 of Ghosh et al. (2006) concerns the relationship between exposure to chloracetic acid and mouse mortality. A simple logistic regression model can be fit to relate the binary death indicator (y) with the levels of exposure (x) to chloracetic acid for the dataset's $n = 120$ mice. That is, data Z consists of pairs $Z_i = (X_i, Y_i)$, for $i = 1, \ldots, n$, and a conditionally Bernoulli model for Y_i, given X_i, with mass function

$$p_\theta(y \mid x) = F(\theta_1 + \theta_2 x)^y \, \{1 - F(\theta_1 + \theta_2 x)\}^{1-y}, \quad \theta = (\theta_1, \theta_2) \in \mathbb{R}^2,$$

where $F(u) = (1 + e^{-u})^{-1}$ is the logistic distribution function. The corresponding likelihood cannot be maximized in closed-form, but this is easy to do numerically, and the maximum likelihood estimator and the corresponding observed information matrix lead to the asymptotically valid inference reported by standard statistical software. Figure 4 shows the data and the fitted "death probability" curve. For exact inference, however,



(a) Joint contours for (Θ_1, Θ_2) (b) Marginal contours for the mean $\Theta_1\Theta_2$

Figure 3: Results for gamma model fit in Example 2. Panel (a) shows the exact IM contour (gray) for the gamma shape and scale parameters (Θ_1, Θ_2) and two approximations: Gaussian density-based ranking (dashed) and likelihood-based ranking (solid). Panel (b) shows the exact marginal IM contour (gray) for the gamma mean and two approximations as described in the text: direct (dashed) and indirect (solid).

the computational burden is heavier: evaluating the exact IM contour for Θ over a sufficiently fine grid of θ values again is prohibitively expensive. As an alternative, the Monte Carlo sampling method presented in Section 3 is easy to implement and runs in a matter of seconds. Figure 5(a) shows the 5000 Monte Carlo samples of (Θ_1, Θ_2) from $\mathsf{Q}_z^\star$ along with the stitched IM contour based on a Gaussian density ranking, closely agreeing with the approximations in Cella and Martin (2024) and Martin and Williams (2024).

A specific and practically relevant question concerns the chloracetic acid exposure level required to make the death probability 0.5. This amounts to setting $F(\theta_1 + \theta_2 x) = 0.5$ and solving for x; the solution is $\lambda = -\theta_1/\theta_2$, and the true value $\Lambda = -\Theta_1/\Theta_2$ is called the *median lethal dose*, or LD50. For inference on Λ, I use the previously-obtained samples (Θ_1, Θ_2) from $\mathsf{Q}_z^\star$ and use a Gaussian density ranking to construct and approximate possibility contour for Λ. This curve is shown in Figure 5(b) and, again, it closely agrees with the asymptotic approximations in Martin and Williams (2024).

5 Application

The examples above are non-trivial when it comes to exact (marginal) inference, but they still fall under the umbrella of "standard" models. As model complexity increases, obviously so too does the computational burden of exact inference. One example of a relatively complex model is that required for modeling censored data: on the one hand, data from the population under investigation are corrupted, which adds complexity, and, on the other hand, little is known about the corruption process so this should be modeled nonparametrically, which also adds complexity. More specifically, I consider time-to-death data on ovarian cancer patients from a clinical trial that took place from 1974 to 1977 (Edmonson et al. 1979); this data is contained in the `ovarian` data set in the R

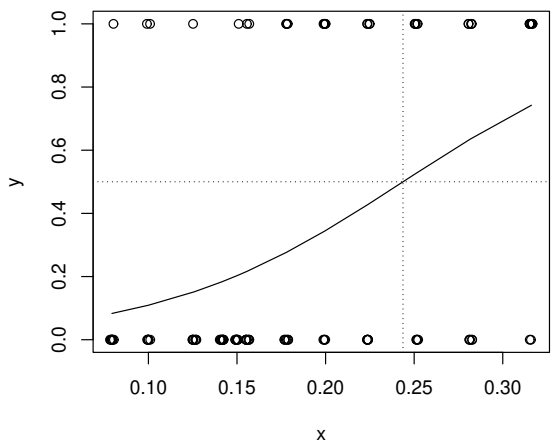


Figure 4: Plot of the data described in Example 3 with the fitted "death probability" curve. The vertical line at $\hat{\lambda}_z = 0.244$ marks the estimated *median lethal dose* (LD50), which is the exposure level at which the death probability is 0.5.

package `survival` (Therneau 2024). Among other things, this data set includes survival times for $n = 26$ patients that entered the study with Stage II or IIIA cancer and were treated with either cyclophosphamide alone or cyclophosphamide with adriamycin. Of these patients, 14 survived to the end of the study—so their survival times were right censored—and 12 unfortunately died. In such cases, it's common to have a parametric model for the survival times, which is adjusted in a nonparametric way to accommodate censoring. Let Y_i denote the actual survival time of patient i, which may or may not be observed. Assign these survival times a statistical model $\{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$, where the true-but-unknown value Θ of the model parameter is the target. Let C_i denote the censoring time for patient i, which is a random variable. Then, under the assumption of random right censoring, the observed data Z consists of n iid pairs $Z_i = (X_i, T_i)$, where

$$X_i = \min(Y_i, C_i) \quad \text{and} \quad T_i = 1(Y_i \leq C_i), \quad i = 1, \ldots, n, \tag{17}$$

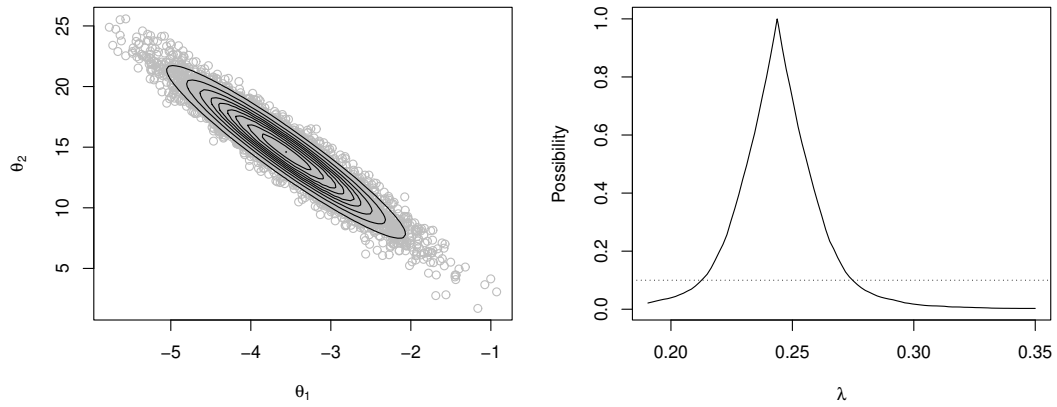
where $T_i = 1$ if the observation is an event time and $T_i = 0$ if it's a censoring time. The goal is to infer the unknown Θ, but with the censoring-corrupted survival times. The likelihood function for the observed data is

$$L_z(\theta, G) = \prod_{i=1}^{n} g(x_i)^{1-t_i} \{1 - G(x_i)\}^{t_i} \times \prod_{i=1}^{n} p_\theta(x_i)^{t_i} \{1 - P_\theta(x_i)\}^{1-t_i},$$

which depends on both the generic value θ of the true unknown model parameter Θ for the concentrations and on the generic value G of the true unknown censoring level distribution G. In the above expression, g and p_θ are density functions for the censoring and concentration distributions, and G and P_θ are the corresponding distribution functions.

Following the review in Section 2, given that G is a nuisance parameter, the natural strategy is to work with a relative profile likelihood for Θ which, in this case, is given by

$$R^{\text{PR}}(z, \theta) = \frac{\prod_{i=1}^{n} p_\theta(x_i)^{t_i} \{1 - P_\theta(x_i)\}^{1-t_i}}{\prod_{i=1}^{n} p_{\hat{\theta}_z}(x_i)^{t_i} \{1 - P_{\hat{\theta}}(x_i)\}^{1-t_i}}, \quad \theta \in \mathbb{T},$$



(a) Joint contour for (Θ_1, Θ_2) (b) Marginal contour for Λ

Figure 5: Summary of the logistic regression results results in Example 3. Panel (a) shows the samples of (Θ_1, Θ_2) from $\mathsf{Q}_z^\star$ and the corresponding approximate possibility contour. Panel (b) shows the approximate possibility contour for the median lethal dose Λ based on the samples in Panel (a).

where $\hat{\theta}_z$ is the maximum likelihood estimator of Θ. The *distribution* of the relative profile likelihood still depends on the nuisance parameter G, so when we define the possibilistic IM contour by validifying the relative profile likelihood, we get

$$\pi_z(\theta) = \sup_G \mathsf{P}_{\theta,G}\big\{R^{\text{PR}}(Z,\theta) \leq R^{\text{PR}}(z,\theta)\big\}, \quad \theta \in \mathbb{T}.$$

Cahoon and Martin (2021) proposed a novel strategy wherein a variation on the Kaplan–Meier estimator (e.g., Kaplan and Meier 1958; Klein and Moeschberger 2003) is used to obtain a $\widehat{G}$, and then the contour above is approximated by

$$\hat{\pi}_z(\theta) = \mathsf{P}_{\theta,\widehat{G}}\big\{R^{\text{PR}}(Z,\theta) \leq R^{\text{PR}}(z,\theta)\big\}, \quad \theta \in \mathbb{T}. \tag{18}$$

Evaluation of the right-hand side via Monte Carlo boils down to sampling censoring levels from $\widehat{G}$, sampling concentration levels from P_θ, and then constructing new data sets according to (17). While this procedure is conceptually relatively simple, naive implementation over a sufficiently fine grid of θ values is very expensive. Fortunately, the proposed Monte Carlo strategy can be readily applied to sample from an inner probabilistic approximation of $\hat{\pi}_z$, from which approximately valid inference on Θ can be obtained. While only first-order asymptotic validity of this IM could be established in Cahoon and Martin (2021), the empirical results presented there are quite striking in that they suggest the possibility of a higher-order accuracy similar to that associated with bootstrap.

As is common in the time-to-event data analysis literature, I'll take a Weibull model for the survival times, where the density function is

$$p_\theta(y) = (\theta_1/\theta_2)\,(y/\theta_2)^{\theta_1-1}\,\exp\{-(y/\theta_2)^{\theta_1}\}, \quad y > 0,$$

where $\theta = (\theta_1, \theta_2)$ is the unknown parameter, with $\theta_1 > 0$ and $\theta_2 > 0$ the shape and scale parameters, respectively. Figure 6 shows the samples of (Θ_1, Θ_2) from the inner

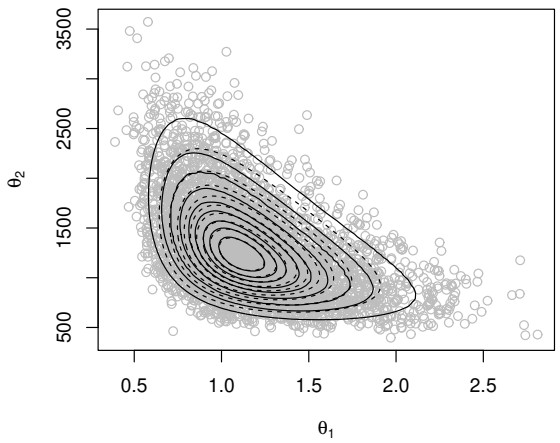


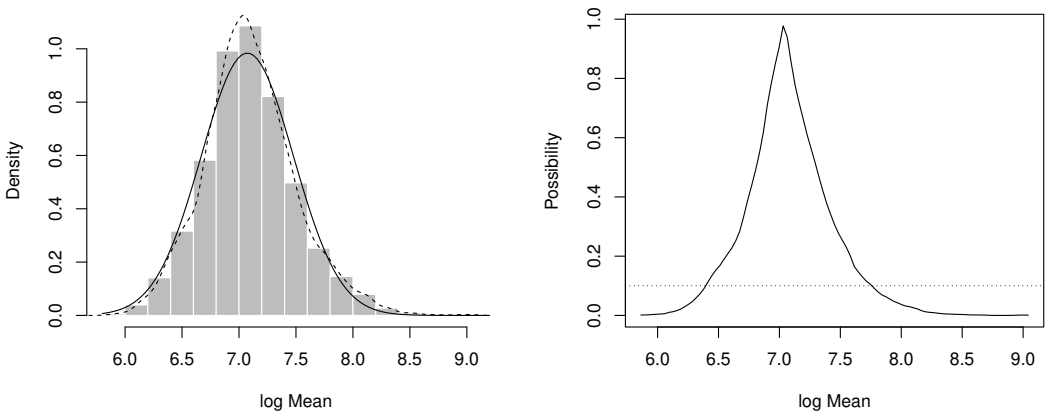
Figure 6: Plot of samples (Θ_1, Θ_2) of the Weibull shape and scale parameters from the inner probabilistic approximation $\mathsf{Q}_z^\star$ along with two corresponding joint contours: Gaussian density ranking (solid) and the asymptotic Gaussian possibility contour (dashed) as developed in Martin and Williams (2024).

probabilistic approximation $\mathsf{Q}_z^\star$ of the IM defined by (18). This plot also shows two approximate contours for Θ: one is based solely on the asymptotic Gaussianity of the possibilistic IM as demonstrated in Martin and Williams (2024) and the other is based on the samples from $\mathsf{Q}_z^\star$ and the Gaussian density ranking. The former asymptotic contour assumes "$n \approx \infty$" which is difficult to justify with $n = 26$ observations, so the latter contour, which is more diffuse in this application, offers more trustworthy inference. It's also very similar to the results presented in Cahoon and Martin (2021), but with only a tiny fraction of the computational cost.

While the Weibull model is common in applications, the Weibull model parameters themselves are difficult to interpret. A relevant and interpretable feature is the mean of the Weibull distribution, which is given by $\Phi = \Theta_2 \, \Gamma(1+\Theta_1^{-1})$, where Γ is the usual gamma function. Like in the gamma and logistic regression examples above, it is straightforward to obtain an approximate marginal IM for the relevant feature Φ or, equivalently $\log \Phi$, based on the samples of (Θ_1, Θ_2) from $\mathsf{Q}_z^\star$. Figure 7(a) shows a histogram of the samples of $\log \Phi$ and, since there's a slight sign of asymmetry, I use the kernel density estimate-based ranking function to construct the marginal contour for $\log \Phi$ in Figure 7(b). From here, a nominal 90% confidence interval for $\log \Phi$ can be immediately read off, i.e., $(6.41, 7.74)$. Exponentiating the endpoints gives a corresponding nominal 90% confidence interval for the mean survival time Φ as $(610.7, 2309.0)$.

6 Conclusion

According to Cui and Hannig (2024), the IM framework is a fundamental advancement in statistical inference. To date, these advances have mostly been foundational, theoretical, and methodological, with practically important computational advances unfortunately dragging behind. The present paper breaks that trend by offering new, simple,



(a) Samples of $\log \Phi$ (b) Marginal contour for $\log \Phi$

Figure 7: Results on the mean survival time $\Phi = \Theta_2 \, \Gamma(1 + \Theta_1^{-1})$ of the Weibull model. Panel (a) shows the distribution of $\log \Phi$ under $\mathsf{Q}_z^\star$ and Panel (b) shows the corresponding marginal IM contour for $\log \Phi$.

and computationally Monte Carlo sampling-driven procedures for approximating various summaries of the IM's possibilistic output. The non-trivial numerical examples presented here highlight the accuracy of the proposed Monte Carlo approximations which are achieved at just a tiny fraction—seconds versus minutes—of the time required to evaluate the IM output directly using naive, brute-force strategies.

The key innovation, motivated in part by Jiang et al. (2023), is to stitch together a collection of simple but individually-inadequate approximations into a superior meta-approximation. Here, for the individual approximations, I'm using the variational-like strategy put forward recently in Cella and Martin (2024), rather than the bootstrap-based strategy offered in Jiang et al. (2023), because the former is computationally simpler than the latter. There may, however, be applications in which the additional flexibility offered by the bootstrap-based strategy is worth the extra computational costs.

The results here mark a first and important step in a series of developments leading to an IM toolbox for more-or-less off-the-shelf use by practitioners. The obvious next step is the development of general-purpose software for those common models—like the ones considered as examples in Sections 4 and 5 above—often encountered in applications. A natural next step would be extending the methods proposed here in various directions. This includes extensions to general non- and semiparametric models, where there may not be a likelihood function (e.g., Cella and Martin 2022), which is well within reach since the application in Section 5 involved a semiparametric model. This would also include problems that involve "model uncertainty," i.e., where all or part of what's unknown and to be learned is the underlying model structure. A good example of this is mixture models, where the number of mixture components is unknown. Structure learning problems often require regularization, since the data alone cannot rule out overly-complex models, so this leads naturally into extensions of the proposed computational framework to accommodate cases with what Martin (2022a,b) has referred to as "partial prior information." Finally, while the new proposal here, and the ideas in Cella and Martin (2024) and elsewhere that it's built on, aren't specific to models involving low-dimensional unknowns, there are

sure to be challenges associated with scaling up this proposal to handle high-dimensional problems. My current suggestion is to focus this effort on improving the efficiency and/or flexibility of the variational approximation in Section 3.3.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation, under grants SES–2051225 and DMS–2412628.

A Technical details

A.1 Proof of Theorem 1

To prove sufficiency, according to (8), it's enough to check that, if Q is as defined in (9), then $\mathsf{Q}(C_\alpha) \geq 1-\alpha$ for each α. Using the fact that the α-cuts are nested, this follows from the simple manipulation,

$$\begin{aligned}
\mathsf{Q}(C_\alpha) &= \int_0^1 \mathsf{K}^\beta(C_\alpha)\, \mathsf{M}(d\beta) \\
&= \int_0^\alpha \underbrace{\mathsf{K}^\beta(C_\alpha)}_{\geq 0}\, \mathsf{M}(d\beta) + \int_\alpha^1 \underbrace{\mathsf{K}^\beta(C_\beta)}_{=1}\, \mathsf{M}(d\beta) \\
&\geq \mathsf{M}([\alpha, 1]) \\
&\geq 1-\alpha,
\end{aligned}$$

where the last inequality follows by the stochastically-no-smaller-than-$\mathsf{Unif}(0,1)$ property of M. To prove necessity, take the given $\mathsf{Q} \in \mathscr{C}(\overline{\Pi})$ and consider a random element $\Theta \sim \mathsf{Q}$. Note that $\Theta \in \partial C_{\pi(\Theta)}$ with Q-probability 1. For some intuition, imagine partitioning $\mathbb{T}$ based on these level sets; then Θ itself is determined by the level set it's on together with its position on the level set, and hence Q corresponds to an average of the conditional distribution of Θ, given $\pi(\Theta)$, with respect to the marginal distribution of $\pi(\Theta)$. It follows from (8) that the random variable $\pi(\Theta)$ is stochastically no smaller than $\mathsf{Unif}(0,1)$, and let M be its marginal distribution relative to Q. Similarly, take K^β to be (a version of) the conditional distribution of Θ, given $\pi(\Theta) = \beta$, relative to Q—note that K^β is fully supported on $\partial C_\beta \subset C_\beta$, as required. Then the equality (9) follows from the law of iterated expectation, completing the proof.

A.2 IMs and variational approximations

The goal here is to review some of the relevant details presented in Cella and Martin (2024) concerning the choice of $\xi = \xi(z, \alpha)$ employed in Section 3.3 above. The focus here, as above, is on the case where the variational family is Gaussian, but similar things can surely be done with other distributional families. The key features of the Gaussian are that its credible sets can be described in closed-form (ellipsoids) and that it's straightforward to sample on the boundary of these credible sets.

Recall the Gaussian family of approximations R_z^ξ in (14), where ξ is a d-vector with positive entries and $J_z(\xi)$ is the ξ-modified version of the observed Fisher information matirx J_z with spectral decomposition $J_z = E\Lambda E^\top$. This is a generalization of a simpler but less flexible approximation that works with a scalar ξ and sets $J_z(\xi) = \xi^{-2}\, J_z$. Also recall the Gaussian family's α-cut in (15):

$$C_\alpha^\xi(z) = \{\theta : (\theta - \hat{\theta}_z)^\top\, J_z(\xi)\, (\theta - \hat{\theta}_z) \leq F_d^{-1}(1-\alpha)\}, \quad \alpha \in [0,1].$$

The key observation is that the Gaussian approximation R_z^ξ assigns probability at least $1-\alpha$ to the original IM's α-cut $C_\alpha(z)$ if $C_\alpha^\xi(z) \supseteq C_\alpha(z)$ or, equivalently, if

$$\sup_{\theta \notin C_\alpha^\xi(z)} \pi_z(\theta) \leq \alpha.$$

Since the contour π_z is itself approximately Gaussian (Martin and Williams 2024) and the maximum likelihood estimator $\hat{\theta}_z$, also the mode of π_z, is in $C_\alpha^\xi(z)$, it follows that the action in the above supremum takes place on the boundary $\partial C_\alpha^\xi(z)$. Moreover, since equality in the above display implies a near-perfect match between the IM's and the posited Gaussian α-cuts, a reasonable goal is to find a root to the function

$$g_\alpha(\xi) := \max_{\theta \in \partial C_\alpha^\xi(z)} \pi_z(\theta) - \alpha.$$

Design of an iterative algorithm to find this root requires care, primarily because evaluating π_z is expensive; so the goal is to evaluate $g_\alpha(\xi)$ with as few π_z evaluations as possible. Cella and Martin (2024) propose to represent the boundary of $C_\alpha^\xi(z)$ by $2d$-many vectors

$$\vartheta_s^{\xi,\pm} := \hat{\theta}_n \pm \left\{F_d^{-1}(1-\alpha)\, \xi_s \,/\, \lambda_s\right\}^{1/2} e_s, \quad s = 1, \ldots, d, \tag{19}$$

where (λ_s, e_s) is the eigenvalue–eigenvector pair corresponding to the s^{th} largest eigenvalue in the spectral decomposition of J_z mentioned above. Then define the vector-valued function $\hat{g}_\alpha$ with components

$$\hat{g}_{\alpha,s}(\xi) = \max\{\pi_z(\vartheta_s^{\xi,+}), \pi_z(\vartheta_s^{\xi,-})\} - \alpha, \quad s = 1, \ldots, d. \tag{20}$$

At least intuitively, negative and positive $\hat{g}_{\alpha,s}(\xi)$ indicate that the α-cut $C_\alpha^\xi(z)$ is too large and too small, respectively in the e_s-direction. From here, Cella and Martin (2024) suggest applying a stochastic approximation algorithm à la Robbins and Monro (1951) and Kushner and Yin (2003) to construct a sequence $(\xi^{(t)} : t \geq 1)$ of d-vectors that converges to a root of $\hat{g}_\alpha$ and, hence, an approximate root of g_α. For an initial guess $\xi^{(0)}$, the specific sequence is defined as

$$\xi_s^{(t+1)} = \xi_s^{(t)} + w_{t+1}\, \hat{g}_{\alpha,s}(\xi^{(t)}), \quad s = 1, \ldots, d, \quad t \geq 0,$$

where (w_t) is a deterministic sequence that satisfies

$$\sum_{t=1}^\infty w_t = \infty \quad \text{and} \quad \sum_{t=1}^\infty w_t^2 < \infty.$$

These steps are iterated until (practical) convergence is achieved, and the limit is what I called $\xi(z, \alpha)$ in Section 3.3. The basic steps are outlined in Algorithm 1, but I refer to Cella and Martin (2024) for further details and discussion.

Algorithm 1: Determining $\xi(z, \alpha)$—from Cella and Martin (2024)

requires: data z, eigen-pairs (λ_s, e_s), and ability to evaluate π_z;
initialize: α-level, guess $\xi^{(0)}$, step size sequence (w_t), and threshold $\varepsilon > 0$;
set: `stop = FALSE`, $t = 0$;
while `!stop` **do**
  construct the representative points $\{\vartheta_s^{\xi^{(t)},\pm} : s = 1, \ldots, d\}$ as in (19);
  evaluate $\hat{g}_{\alpha,s}(\xi_s^{(t)})$ for $s = 1, \ldots, d$ as in (20);
  update $\xi_s^{(t+1)} = \xi_s^{(t)} \pm w_{t+1}\, \hat{g}_{\alpha,s}(\xi_s^{(t)})$ for $s = 1, \ldots, d$;
  if $\max_s |\xi_s^{(t+1)} - \xi_s^{(t)}| < \varepsilon$ **then**
    $\xi(z, \alpha) = \xi^{(t+1)}$;
    `stop = TRUE`;
  else
    $t \leftarrow t + 1$;
  end
end
return $\xi(z, \alpha)$;

References

Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proc. Royal Soc. A*, 475(2227):2018.0565.

Basu, D. (1964). Recovery of ancillary information. *Sankhyā Ser. A*, 26:3–16.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.*, 112(518):859–877.

Cahoon, J. and Martin, R. (2021). Generalized inferential models for censored data. *Internat. J. Approx. Reason.*, 137:51–66.

Cella, L. and Martin, R. (2022). Direct and approximately valid probabilistic inference on a class of statistical functionals. *Internat. J. Approx. Reason.*, 151:205–224.

Cella, L. and Martin, R. (2023). Possibility-theoretic statistical inference offers performance and probativeness assurances. *Internat. J. Approx. Reason.*, 163:109060.

Cella, L. and Martin, R. (2024). Computationally efficient variational-like approximations of possibilistic inferential models. `arXiv:2404.19224`.

Couso, I., Montes, S., and Gil, P. (2001). The necessity of the strong α-cuts of a fuzzy set. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 9(2):249–262.

Cui, Y. and Hannig, J. (2024). Demystifying inferential moels: A fiducial perspective. `arXiv:2205.05612`.

Denœux, T. (2006). Constructing belief functions from sample data using multinomial confidence regions. *Internat. J. of Approx. Reason.*, 42(3):228–252.

Denœux, T. (2014). Likelihood-based belief function: justification and some extensions to low-quality data. *Internat. J. Approx. Reason.*, 55(7):1535–1547.

Destercke, S. and Dubois, D. (2014). Special cases. In *Introduction to Imprecise Probabilities*, Wiley Ser. Probab. Stat., pages 79–92. Wiley, Chichester.

Dubois, D., Foulloy, L., Mauris, G., and Prade, H. (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliab. Comput.*, 10(4):273–297.

Edmonson, J. H., Fleming, T. R., Decker, D. G., Malkasian, G. D., Jorgensen, E. O., Jefferies, J. A., Webb, M. J., and Kvols, L. K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer Treat. Rep.*, 63(2):241–247.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa.

Fraser, D. A. S., Reid, N., and Wong, A. (1997). Simple and accurate inference for the mean of a gamma model. *Canad. J. Statist.*, 25(1):91–99.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer, New York.

Hose, D. (2022). *Possibilistic Reasoning with Imprecise Probabilities: Statistical Inference and Dynamic Filtering*. PhD thesis, University of Stuttgart. `https://dominikhose.github.io/dissertation/diss_dhose.pdf`.

Hose, D., Hanss, M., and Martin, R. (2022). A practical strategy for valid partial prior-dependent possibilistic inference. In Le Hégarat-Mascle, S., Bloch, I., and Aldea, E., editors, *Belief Functions: Theory and Applications (BELIEF 2022)*, volume 13506 of *Lecture Notes in Artificial Intelligence*, pages 197–206. Springer.

Hult, H. and Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Adv. in Appl. Probab.*, 34(3):587–608.

Jacob, P. E., Gong, R., Edlefsen, P. T., and Dempster, A. P. (2021). A Gibbs sampler for a class of random convex polytopes. *J. Amer. Statist. Assoc.*, 116(535):1181–1192.

Jiang, Y., Liu, C., and Zhang, H. (2023). Finite sample valid inference via calibrated bootstrap. `arXiv:2408.16763`.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis*. Springer-Verlag, New York, 2nd edition.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, New York, second edition.

Martin, R. (2015). Plausibility functions and exact frequentist inference. *J. Amer. Statist. Assoc.*, 110(512):1552–1561.

Martin, R. (2018). On an inferential model construction using generalized associations. *J. Statist. Plann. Inference*, 195:105–115.

Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *Internat. J. Approx. Reason.*, 113:39–73.

Martin, R. (2021). Inferential models and the decision-theoretic implications of the validity property. `arXiv:2112.13247`.

Martin, R. (2022a). Valid and efficient imprecise-probabilistic inference with partial priors, I. First results. `arXiv:2203.06703`.

Martin, R. (2022b). Valid and efficient imprecise-probabilistic inference with partial priors, II. General framework. `arXiv:2211.14567`.

Martin, R. (2023a). Fiducial inference viewed through a possibility-theoretic inferential model lens. In Miranda, E., Montes, I., Quaeghebeur, E., and Vantaggi, B., editors, *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*, volume 215 of *Proceedings of Machine Learning Research*, pages 299–310. PMLR.

Martin, R. (2023b). Valid and efficient imprecise-probabilistic inference with partial priors, III. Marginalization. `arXiv:2309.13454`.

Martin, R. (2024a). A possibility-theoretic solution to Basu's Bayesian–frequentist via media. *Sankhya A*, 86:43–70.

Martin, R. (2024b). Which statistical hypotheses are afflicted by false confidence? In Bi, Y., Jousselme, A.-L., and Denoeux, T., editors, *BELIEF 2024*, volume 14909 of *Lecture Notes in Artificial Intelligence*, pages 140–149, Switzerland. Springer Nature.

Martin, R. and Liu, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313.

Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(1):195–217.

Martin, R. and Liu, C. (2015b). *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Martin, R. and Liu, C. (2015c). Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, 110(512):1621–1631.

Martin, R. and Williams, J. P. (2024). Asymptotic efficiency of inferential models and a possibilistic Bernstein–von Mises theorem. `arXiv:2412.15243`.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, 31(6):1695–1731.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407.

Shafer, G. (1982). Belief functions and parametric models. *J. Roy. Statist. Soc. Ser. B*, 44(3):322–352. With discussion.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games, vol. 2*, volume no. 28 of *Ann. of Math. Stud.*, pages 307–317. Princeton Univ. Press, Princeton, NJ.

Smets, P. and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66(2):191–234.

Therneau, T. M. (2024). *A Package for Survival Analysis in R*. R package version 3.5-8.

Troffaes, M. C. M. and de Cooman, G. (2014). *Lower Previsions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Wasserman, L. A. (1990a). Belief functions and statistical inference. *Canad. J. Statist.*, 18(3):183–196.

Wasserman, L. A. (1990b). Prior envelopes based on belief functions. *Ann. Statist.*, 18(1):454–464.