Comments on "Demystifying inferential models: a fiducial perspective" by Y. Cui and J. Hannig

Ryan Martin[*]

June 18, 2022

1 Introduction

When it comes to the foundations of statistics, there are still lots of unanswered questions. One thing I think is relatively safe to say is that, if one of the currently dominant schools of thought—Bayesian or frequentist—were objectively correct, then that fact would've been demonstrated by now. It's most likely that both frequentist and Bayes are correct but only in certain cases/senses, so, therefore, neither is correct in the big-picture sense that we're considering. If the goal is to blend the uncertainty quantification aspect of Bayesian inference with the calibration and error rate control properties offered by frequentist methods, then something different is needed. The *inferential model* (IM) framework that I've developed with Chuanhai Liu and others is, I think, a strong contender for being the "something different" that can meet the community's needs.

The IM framework differs fundamentally from other proposals because it explicitly works in the domain of imprecise probability as opposed to the traditional (precise) probability theory we learn as students. My view is that this imprecision is necessary to achieve the goals so, rather than downplay this aspect to avoid the inevitable criticism, I've ramped up my efforts recently to shed even more light on it. My hope was that, if I draw attention to it, perhaps even challenge those who have differing views, then they'd feel obligated to try and prove me wrong. It's through engaging in this critical back-and-forth that progress can bes be made. Unfortunately, I think most have ignored my challenges. So I was delighted to learn that Professors Cui and Hannig have not ignored me, and that they've offered their own perspectives on how IMs relate to fiducial and other more traditional forms of probabilistic uncertainty quantification.

What follows is my reaction to the results and conclusions of their paper (Cui and Hannig 2022). I can't say that my reaction is positive, but not because their perspective isn't interesting or because they've done something wrong. My feeling is that there's a general misunderstanding of what IMs are aiming to achieve. So, in the spirit of having an open dialog that can lead to real progress, I offer the detailed comments below. I sincerely hope that this is just the start of a conversation.

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

2 Main comments

Section 2. Proposition 2.1 states that, when the random set $\mathcal{S}$ is nested, the IM's lower and upper probability output, i.e., $\mathsf{bel}_y(\cdot)$ and $\mathsf{pl}_y(\cdot)$, are such that, for every A, either $\mathsf{bel}_y(A) = 0$ or $\mathsf{pl}_y(A) = 1$. This is a general fact that applies to all possibility measures[1] or, equivalently, consonant belief functions; it has nothing specifically to do with IMs. The connection to IMs is made through the choice to use a nested random set $\mathcal{S}$. The authors later argue that the choice of nested random sets is, in some sense, "right," so I find their criticism of IMs based on the above fact to be both contradictory to claims they make later and, more importantly, without warrant. I'll have more to say below about the nested random set stuff, but here I'll explain why I think the authors' IM-critical conclusion is without warrant.

1. Based on the above result, the authors claim that the gap, $\mathsf{pl}_y(A) - \mathsf{bel}_y(A)$, between the two is "perhaps too large." That claim is based on an implicit comparison, so my question is: "too large" compared to what? My guess is the authors are interpreting the pair $\{\mathsf{bel}_y(A), \mathsf{pl}_y(A)\}$ as bounds for some "true y-dependent probability of A," in which case a wide interval is arguably less informative than a narrow one. However, no "true y-dependent probability of A" exists, so $\mathsf{bel}_y(A)$ and $\mathsf{pl}_y(A)$ are not to be interpreted as bounds for such a quantity. I understand that the terminology "lower" and "upper probability" (which isn't mine) can be misleading, so I've recently tried to explain this different to avoid confusion.[2] Anyway, when the authors state "at most one of the probability bounds $(\mathsf{bel}_y(A), \mathsf{pl}_y(A))$ provides any useful information," their conclusion is (basically) correct, but the interpretation is wrong: the point is that the IM output $\{\mathsf{bel}_y(\cdot), \mathsf{pl}_y(\cdot)\}$ is the *pair*, not bel_y and pl_y as separate pieces, so if you focus on one and ignore the other, then some information is expected to be lost.

2. It can also be argued that the phenomenon highlighted in Proposition 2.1 is desirable, or at least not bad. In the present setting, there is some value (or values) of θ that the data specifically points to. These are the values we tend to choose as *estimates* of θ based on data y. For concreteness, let $\hat\theta_y$ denote the maximum likelihood estimator and consider testing hypotheses $H_0 : \theta \in A$ using the usual likelihood ratio test. Mathematically, there is no significance level at which we could reject H_0 when the A in question contains $\hat\theta_y$. If one aims to quantify uncertainty about θ, given y, in some probabilistic or distributional kind of way, while simultaneously being consistent with this intuition coming from the behavior of the likelihood ratio test, how else could one do it other than having $\mathsf{pl}_y(A) = 1$ for all $A \ni \hat\theta_y$? This makes perfect sense to me, as it did to Shafer (1976, Ch. 10–11), so I don't agree with Cui & Hannig's conclusion that the phenomenon described in Proposition 2.1 should be interpreted negatively or as a sign of inefficiency in the IM.

[1] The earliest mention of this fact that I'm aware of is in Shackle (1961, p. 74), which pre-dates the work of Dempster and Shafer on the topic.

[2] I've recently been describing the IM's lower and upper probabilities as lower and upper bounds on the prices one would be willing to pay for certain bets. The advantage of this interpretation is that it's easier to see that there's no "objectively correct price" for which these are intended to bound.

Lemma 2.2 in the present manuscript looks very similar, if not virtually the same, as Theorem 4.3 in Martin and Liu (2015). Indeed, Theorem 4.3 in the book is precisely why our IM developments always focus on the case of nested random sets: that choice we made was not about convenience or being compelled by the consonance-encouraging arguments given by Shackle or Shafer, it's about efficiency for us. That there's some overlap between the authors' results and ours doesn't mean the former aren't interesting, but I think the authors should comment on if the conclusions here are different than those in the book and, if so, then how.

Section 3. First, Theorems 3.1 and 3.2 in the paper are very closely related to the results in Proposition 4.1 and Theorem 4.4 in Martin and Liu (2015). There are some differences, however, the main one being that our discussion of "fiducial" in the book centered around the classical version of Fisher, Dempster, and others. The authors' generalized fiducial framework is different, of course, but it's not clear to me how this affects the details of the present analysis. This would help a lot if the authors could explain here how the *generalized* nature of their fiducial framework distinguishes their results from those in the book. In Theorem 3.2 in particular, we get virtually the same conclusion in Theorem 4.4 in the book, but only for certain kinds of assertions. What's different about the analysis here that allows for a more general conclusion? Is it the generalized nature of the fiducial formulation, just a more sophisticated analysis, or something else?

I have a few issues with the authors' paragraph following the proof of their Theorem 3.2. Each is relatively short, so I'll just list them.

1. The statement "the bound can be achieved for any A using some random set $\mathcal{S}$" is technically correct, but the crucial fact that the random set $\mathcal{S} = \mathcal{S}_A$ that achieves the bound would generally depend on A gets swept under the rug. By not emphasizing this point, the authors give the impression that it's possible to get a valid IM that agrees with the fiducial distribution, which is not true. The IM achieves its desirable properties with a fixed random set $\mathcal{S}$, so if you have to break the valid IM by introducing assertion-dependent random sets that ruin these properties, then that's neither a shortcoming of the IM nor a selling point for fiducial.

2. Again, "beliefs are perhaps too small and plausibilities too high" is not a sign of the IM being inefficient. To say "too small" or "too high" requires a comparison to something else that achieves the same validity properties, but there's nothing to compare to, at least not yet.

3. This statement about "data snooping" is weird. Who said anything about data snooping, what motivated making such a claim? This has nothing to do with IMs—tailoring the choice of method based on peeking at the data almost always ruins whatever frequentist properties the method satisfies without tailoring.

To conclude my discussion of this section of the paper, I have one comment and one question concerning the result in Theorem 3.3.

1. The first part of the theorem is not surprising. The IM output is determined by the distribution of a nested, data-dependent random set $\Theta_y(\mathcal{S})$, which depends implicitly on the statistical model and the association. To know the distribution

of $\Theta_y(\mathcal{S})$ amounts to knowing the sets on which it's support and the probabilities assigned to those sets. These probabilities are determined by the validity condition, as the authors have observed, so then all that's left is knowing the support. In other words, the IM output ought to be determined by a collection of nested, data-dependent subsets of Θ. So the fact that what Cui & Hannig refer to as *principle assertions* "carry all the information available in the IM" is not a surprise and not a criticism of the IM (if that's how the authors intended it).

Incidentally, the "best" random set $\mathcal{S}$ seems to be that analyzed in Theorem 4.1 of Martin and Liu (2015), which is entirely determined by its support (and other features of the problem which are generally taken as given). So if $\mathcal{S}$ is itself determined by a nested collection of subsets, then $\Theta_y(\mathcal{S})$ clearly would be too.

2. About the second part of the theorem, on the one hand, of course the sets $A_{\alpha,y}$ defined there depend on the IM's choice of $\mathcal{S}$. On the other hand, I don't see any obvious reason why the fiducial distribution would depend on $\mathcal{S}$. Then the theorem states that those principle assertions (that depend on $\mathcal{S}$) also effectively determine (a version of) the fiducial distribution. Does that mean the choice between "different versions" of the fiducial distribution is more-or-less equivalent the IM's choice of $\mathcal{S}$? Furthermore, if the fiducial distribution is effectively determined by the IM, but lacks the strong validity property of the latter, then isn't the appropriate conclusion that the fiducial distribution is a sort of "quick and dirty" IM?

Section 4. A confidence curve (CC) is equivalent to a collection of confidence regions, so it offers nothing more than what's offered by the set of confidence regions associated with it. That is, without introducing more structure, they can only answer questions that can be answered with confidence regions. If one wants to get more from a CC, e.g., like the kind of uncertainty quantification that I'm after, then this additional structure needs to be spelled out. I did this in Martin (2021a).

In particular, Cui & Hannig's Lemma 4.1 should be compared to Theorem 3 in Martin (2021a). There I also provide a detailed discussion of how confidence distributions are related to IM; see, also, Martin et al. (2021). Furthermore, Theorem 4.1 in Cui & Hannig should be compared to Theorem 6 in Martin (2021a); I also do something similar for tests, not just for confidence regions, and I allow for the possibility that the given test/confidence region is only for a feature of θ, not for θ in its entirety. Furthermore, I think Cui & Hannig's Theorem 4.2 is saying basically the same thing as my Theorem 6 mentioned above. I can't follow the definition of what they call cc_y'—what's the infimum taken with respect to?—but my result also shows how efficiency can be gained compared to the original confidence regions through the IM construction. I also use that potential efficiency gain to improve upon some existing methods developed recently for some non-trivial problems (Martin 2021a, Sec. 8–9).

To follow up on an earlier point, but in more detail, there's a crucial issue that's being entirely overlooked in this paper—the fiducial, CC, and IM calculi. For example, given a CC for θ, how does one get a CC for $\phi = \phi(\theta)$? Example 2 in the paper is a good illustration of this: one can write CCs for μ_x, for μ_y, and even for (μ_x, μ_y), but the CC theory doesn't say how you could get the CC for μ_x/μ_y from these; in fact, as far as I know, the theory says don't even try to do so, because it doesn't work. To be clear, given a CD for θ, the marginal distribution for $\phi = \phi(\theta)$ obtained by following the probability calculus is *not* generally a CD for ϕ. This is why the CC presented in Example 2 for μ_x/μ_y is constructed separate from the CC for (μ_x, μ_y) rather than derived from it. There is a calculus one can use for manipulating CCs/CDs to get conclusions that aren't immediately available from the given confidence regions, e.g., marginal inference, but it's not the probability calculus. As I show in Martin (2021a), the calculus you need is the possibility calculus that IMs use explicitly. So, it's wrong to interpret the result in Theorem 4.1 of this paper, or Theorem 6 in Martin (2021a), as saying that IMs can be understood/viewed as CCs; instead, since IMs provide a detailed, mathematical prescription for how to carry out valid uncertainty quantification, these results actually say that the IM framework is the right way to use CCs.

To summarize, an attempt to connect IMs to fiducial distributions, CCs, etc. is necessarily incomplete if the calculus that would be used to manipulate them isn't considered. A CC on its own is equivalent to a collection of confidence regions, so it contributes nothing beyond what is contained in the confidence regions. Similarly, a fiducial distribution evaluated at various assertions also isn't especially meaningful because this can't be manipulated in a way that provides a complete quantification of uncertainty about θ and its relevant features. Upon considering the manipulation of these objects, it becomes almost immediately clear that the probability calculus won't do. The possibility calculus works in the sense of preserving the validity/confidence properties, but then one is basing inference on the IM, not really the CC. If IMs already do everything that one can hope for in the context of probabilistic uncertainty quantification in statistical inference, and if imprecise probability considerations are inevitable, then why the resistance to IMs?

Section 5. I disagree with almost all of the points/conclusions made in this section. Let me go through each of them line-by-line.

1. Line 1: "the key concept of validity is innate to CCs." I don't see how this could be true when CCs only allow you to read off confidence regions. IMs, on the other hand, offer validity guarantees at every assertion, including marginal assertions about $\phi = \phi(\theta)$. To make a CC similarly valid, one needs to view it through the lens of possibility theory, but that's not "innate." If anything, the conclusion is that if you want a CC that's valid, then you need to use an IM.

2. Line 2: "IMs are valid when they produce valid CCs." IMs are always valid and, in particular, they produce CCs from which genuine confidence regions can be read off and, moreover, by explicitly identifying that CC as a possibility contour, one immediately knows how this should be manipulated for valid marginal inference, etc.; see Martin (2021a) and Martin et al. (2021).

3. Line 4: "the big advantage of IMs..." IMs are aiming to do more than just construct confidence regions. The objective is valid probabilistic uncertainty quantification about θ which, among other things, includes the construction of confidence regions.

4. Line 6: "main link between IMs and CCs." Even if you have the sets $A_{\alpha,y}$, the CC theory doesn't tell you how to do anything besides read off the sets $A_{\alpha,y}$. So I agree that there's a link, but it only goes in one direction: IM $\to$ CC but CC $\not\to$ IM.

5. Line 11: "IM answers an old question, when are fiducial credible intervals confidence intervals?" I don't think that's what IMs do, but Pitman (1957) talks about this briefly (and points to an earlier paper with details). I'll have a few other things to say about this below. To me, the question isn't about confidence regions at all, at least not directly. We have lots of different ways to construct approximate confidence regions, the question is how to do more with that construction, how to get a more complete uncertainty quantification for θ (and its relevant features) while retaining the frequentist-style calibration properties.

6. Line 27: "IMs can be viewed as fiducial distribution based confidence curves." While I'm not entirely sure what the authors mean by this, I'm inclined to 100% disagree. Again, IMs are about valid uncertainty quantification, not just about confidence regions. If all that is desired is a confidence region, then there's no point in any of these considerations, there are so many classical strategies that can be used to directly get a confidence region. These considerations are relevant only because the objective is to do more than construct confidence regions. The results in this paper—and in Martin (2021a)—actually say that, to achieve more with a CC/CD, one needs the possibility-theoretic perspective that the IM explicitly brings. The more appropriate take-away message, in my view, is that "CCs are quick and dirty IMs"—they're summaries that perform one specific (and important) task, but don't fully achieve the uncertainty quantification objective without the possibility calculus, which is clearly justified through the underlying IM.

3 Some relatively minor things

1. Clearly I don't especially like "demystifying" in the title. It suggests that (a) my explanations haven't made the developments clear and (b) that the connection to fiducial inference will make it more clear, both of which I personally disagree with. I'd prefer if the title could say more specifically what the paper contributes, but the authors are free to do as they please.

2. Page 7, line 3: $\theta_y(\mathcal{S})$ should be $\Theta_y(\mathcal{S})$.

3. Page 7, line 8: The notion of "measurable" being mentioned here is with respect to the random set $\mathcal{S}$, not y. This is obviously much more complicated and, if the authors choose to bring this up, then I think some explanation of what that actually means should be given. I've generally chosen not to bring it up in my papers because it's a distraction and not in the spirit of "demystifying." A related point is that it might help the reader if the authors somehow make clear that "P" in Equation (4) and elsewhere is with respect to the distribution of $\mathcal{S}$, not y. In my writing on the subject, I've chosen to make explicit in the notation which quantity is random in each probability statement, but the authors are free to adopt a different style.

4. Concerning the validity property in Equation (5), I'd suggest that the authors emphasize more the "for all A" part of it. For example, as stated, if they refer to Equation (5) as the definition of validity, then it doesn't include "for all A," so the

reader could get the impression that the validity property achieved by the IM is for a particular A when, in fact, it's achieved for all A.

5. It's just a matter of taste I suppose, but I had hoped to see an example other than one that appeared in the very first IM publication. This is a nice example, due to its simplicity, but it's arguably too simple to really communicate what's happening. It could also be nice to see a non-trivial example that compares the IM and generalized fiducial solutions, e.g., one that distinguishes the validity property achieved by the IM from the confidence-related properties achieved by fiducial. Having an example where the sets $A_{\alpha,y}$ are written explicitly could also be helpful to the reader.

6. Finally, and this might only be relevant to me, but the presentation of IM-related material here is based entirely on work that was done roughly 10 years ago. I've done a lot of work in the last 3+ years or so (e.g., Cella and Martin 2021a,b, 2022; Liu and Martin 2021; Martin 2019, 2021a,b, 2022) to modernize, simplify, and clarify some of those details that the authors aim to "demystify." I understand that it might be difficult to keep up with all of these developments, so of course it's fine to stick with the older stuff, but at least some references to the more recent work would be appreciated.

4 Moving forward

To me, the most important fundamental question concerning the relationship between IMs and Bayes/fiducial/CDs/etc. is the following. The *false confidence theorem* (Balch et al. 2019) states that, no matter what kind of data-dependent distribution one uses, there are always false assertions that will tend to be assigned high posterior probability—this is an issue about uncertainty quantification, not about confidence regions, etc. This misalignment of the posterior probability creates a risk for erroneous inference. *What do those problematic assertions look like?* This is relevant to the Bayes/fiducial/CD communities because, if it ends up that these assertions are rather extreme, impractical, or whatever, then there would be essentially no risk in ignoring false confidence altogether. Of course, the problematic assertions depend on the context of the problem, so I doubt that one can say, across the board, there's no reason for concern; the satellite collision example in Balch et al. (2019) makes it clear that the class of problematic assertions isn't vacuous. Anyway, my point is that we currently have very little understand of what those problematic assertions look like, which means we don't understand how great the risk is. The results from Pitman (1957) I mentioned above seem relevant but so far I've had trouble wrapping my mind around what he shows. I'm obviously interested in showing that the problematic assertions are plentiful, whereas Bayes/fiducial/CD folks want to show that they're rare or extreme. So it seems like the best way to get to an answer to this question, one way or the other, is for both groups to try. Cui & Hannig have obviously thought more about generalized fiducial inference than I have, so I wonder if they have any insights on this question they can share.

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.

Cella, L. and Martin, R. (2021a). Approximately valid probabilistic inference on a class of statistical functionals. `arXiv:2112.10232`.

Cella, L. and Martin, R. (2021b). Valid inferential models for prediction in supervised learning problems. `arXiv:2112.10234`.

Cella, L. and Martin, R. (2022). Validity, consonant plausibility measures, and conformal prediction. *Internat. J. Approx. Reason.*, 141:110–130.

Cui, Y. and Hannig, J. (2022). Demystifying inferential moels: A fiducial approach. `arXiv:2205.05612`.

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

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

Martin, R. (2021b). Inferential models and the decision-theoretic implications of the validity property. `https://researchers.one/articles/21.12.00005`.

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., and Ferson, S. (2021). Response to the comment 'Confidence in confidence distributions!'. *Proc. R. Soc. A.*, 477:20200579.

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

Pitman, E. J. G. (1957). Statistics and science. *J. Amer. Statist. Assoc.*, 52:322–330.

Shackle, G. L. S. (1961). *Decision Order and Time in Human Affairs.* Cambridge University Press, Cambridge.

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