Belief functions and valid statistical inference\(^1\)

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin
www.researchers.one

BELIEF/SMPS 2018
University of Compiègne
September 20, 2018

\(^1\)Thanks to friends and collaborators M. Balch, H. Crane, C. Liu...
Belief functions have a wide range of applications. Just look at this conference’s impressive program! My focus here is on classical statistical inference, what Dempster was thinking about when he started all this. Despite Dempster’s genius, additive beliefs have cornered the market on uncertainty quantification in statistics. In this talk, I hope to convince you that, for certain concrete reasons, non-additivity is actually necessary.
Statistical problem:
- Observable data is $X$;
- Model is $X \sim P_\theta$, where $\theta \in \Theta$ is unknown.

Goal is to quantify uncertainty about $\theta$ based on $X = x$.

Various ways to approach construction of an additive belief or “posterior distribution” for $\theta$, given $x$:
- Bayes
- Fiducial/confidence

Can also construct non-additive belief functions for $\theta$, given $x$.

To be or not to be (additive)?
Comparison of additive and non-additive beliefs is most often treated *philosophically*, i.e., which is “right”? Different philosophies can co-exist, but it’s perhaps easier to compare in terms of statistical properties. From this perspective, there is a difference!

Take-away messages:
- *certain statistical properties fail under additivity,*
- *those properties can be attained under non-additivity,*
- *and, in fact, basically anything that satisfies these good properties corresponds to a non-additive belief*
This talk

- The price of additivity:
  - satellite collision example
  - false confidence theorem
- Overcoming false confidence with valid beliefs
  - validity property
  - achieving validity via *inferential models* (IMs)
- A sort of converse to the IM validity theorem
- Concluding remarks, etc.
Satellite collision problem

- A satellite orbiting Earth could collide with another object.
- Potential big mess, so navigators try to avoid collision.
- Data on position, velocity, etc is used, along with physical and statistical models, to compute a collision probability.
- If collision probability is low, then satellite is judged safe; otherwise, some action is taken.
- An unusual phenomenon has been observed: noisier data makes collision probability small...
- Systematically misleading conclusions – bad!
Satellite collision, cont.

- **Illustration:**
  - $\|x\|$ denotes measured distance between satellite and object.
  - True distance $\leq 1$ implies collision.
  - Measurement error variance is $\sigma^2$.
  - $p_x(\sigma) =$ probability of non-collision, given $x$.

- When $\sigma$ is large, $p_x(\sigma)$ is large, for any $x$.
- And large $\sigma$ is the reality in this application.
- Therefore, collision probability is not a reliable metric!
What’s going on here?

Apparently, data might not be sufficiently informative with respect to questions about collision/non-collision.

*Additivity* forces the probability to go somewhere, happens that it goes to non-collision no matter what data says.

Even if the satellite is on a direct collision course, the probability calculations will tell navigators that it’s safe.

*False confidence*:\(^2\) a hypothesis tending to be assigned large probability even though data does not support it.

Is this general, or just something weird about this problem?

\(^2\)Balch, M., and Ferson, arXiv:1706.08565
False confidence theorem.

Let $\Pi_X$ be any additive belief on $\Theta$, depending on data $X$. Then, for any $\alpha$ and any $t$, there exists $A \subset \Theta$ such that

$$A \not\ni \theta \quad \text{and} \quad P_{\theta}\{\Pi_X(A) > t\} \geq \alpha.$$

- In words, there always exists a false hypothesis that tends to be assigned high posterior probability.
- If judgments about the plausibility of a hypothesis are made based on the magnitude of its probability, then the theorem says there’s risk of systematic error.
- Reveals a practical price attached to additivity.
- Theorem doesn’t say which hypotheses are problematic, or to what extent, only that they exist.
False confidence, cont.

- Toy version of the satellite example.\(^3\)
  - Let \( A \) denote the non-collision hypothesis.
  - Then \( \Pi_X(A) \) as a random variable, with a CDF
  - Plot CDF when data are generated under a collision course.

- **False confidence**: \( \Pi_X(A) \) is almost always large!

\(^3\)More examples in Carmichael and Williams, arXiv:1807.06217
Theorem says *every additive belief function* is afflicted with false confidence.

To avoid false confidence, we must go non-additive.

- Let $\text{bel}_x$ be a data-dependent belief function.
  - $\text{bel}_x(\emptyset) = 0$, $\text{bel}_x(\Theta) = 1$, and completely monotone.
  - In particular, $\text{bel}_x(A) + \text{bel}_x(A^c) \leq 1$.

But not every non-additive belief would be satisfactory.

That is, to avoid false confidence, the belief must satisfy some additional conditions...
First a question: *Why should we avoid false confidence?*

What about all the success stories of, say, Bayesian inference?

Key point: a Bayesian posterior is good for some tasks and not for others.\(^4\)

But statisticians are in the business of developing *methods*.

If my method is touted as giving a full posterior distribution, then people will want to use it as such.

If I can’t control what hypotheses a user might plug in, then I better be sure that there’s no false confidence.

\(^4\)See Grünwald’s recent work on *safe probability*, arXiv:1604.01785.
Validity property

Definition.

A data-dependent belief function \( \text{bel}_x \) is said to be *valid* if

\[
\sup_{\theta \not\in A} P_{\theta}\{\text{bel}_x(A) > 1 - \alpha\} \leq \alpha, \quad \forall \ A \subseteq \Theta, \quad \forall \ \alpha \in (0, 1).
\]

- Validity property\(^5\) implies that assigning high belief to a false hypothesis is a rare event — *no false confidence*.
- \( \alpha \) on inside and outside calibrates beliefs.
- Definitions are fine, but how to achieve validity?

---

\(^5\)First version of this is in M. and Liu, arXiv:1206.4091; see, also, the “calibrated beliefs” in Denoeux & Li (IJAR 2018).
Inferential models

- The *inferential model* (IM) approach relies on random sets.
- Express the statistical model as

\[ X = a(\theta, U), \quad U \sim P_U, \quad P_U \text{ known.} \]

- Difference between IMs and fiducial:
  - Fiducial converts distribution of \( U \) to a distribution of \( \theta \).
  - IMs care about the *particular value* of \( U \) corresponding to observed data \( X = x \).
  - This special value of \( U \) is unobservable, but we can “guess” its value with a random set \( S \sim P_S \) in the \( U \)-space.

- There is theory to guide the choice of \( P_S \).
Given $S \sim P_S$, push it forward to the $\theta$ space:

$$\Theta_x(S) = \bigcup_{u \in S} \{\theta : x = a(\theta, u)\}. $$

Then compute the random set probability$^6$

$$\text{bel}_x(A) = P_S\{\Theta_x(S) \subseteq A\}. $$

This gives a genuinely non-additive belief function, along with the corresponding plausibility function

$$\text{pl}_x(A) = P_S\{\Theta_x(S) \cap A \neq \emptyset\}. $$

$^6$Here I’m ignoring the possibility that $\Theta_x(S) = \emptyset$. 

IMs, cont.

**IM validity theorem.**

With a “suitable choice of random set $S$,” the IM satisfies the validity property and, therefore, avoids false confidence.

- Re: “suitable choice of $S$.”
  - Let $f(u) = P_S(S \ni u)$, contour function.
  - Need $f(U) \geq_{st} \text{Unif}(0, 1)$ when $U \sim P_U$.

- This is actually easy to arrange...

- Consequences of validity:
  - “Reject $H_0 : \theta \in A$ if $\text{pl}_x(A) \leq \alpha$” is a size $\alpha$ test.
  - Nominal coverage probability of $100(1 - \alpha)\%$ plausibility region

\[
P_\alpha(x) = \{ \vartheta : \text{pl}_x(\{\vartheta\}) \geq \alpha \}.
\]
Let $X \sim \text{Pois}(\theta)$, distribution function $F_\theta$.

Construct an IM for $\theta$:

A. $F_\theta(X - 1) \leq U < F_\theta(X)$, $U \sim \text{Unif}(0, 1)$.

P. “Default” predictive random set

$$S = \{ u : |u - 0.5| \leq |\tilde{U} - 0.5|, \tilde{U} \sim \text{Unif}(0, 1) \}.$$ 

C. Combine to get

$$\Theta_x(S) = \bigcup_{u \in S} \{ \theta : F_\theta(x - 1) \leq u < F_\theta(x) \}$$

$$= \left[ G_x^{-1}(0.5 - |\tilde{U} - 0.5|), G_{x+1}^{-1}(0.5 + |\tilde{U} - 0.5|) \right].$$

Can evaluate belief and plausibility based on $\tilde{U} \sim \text{Unif}(0, 1)$.

---

7Use fact that $F_\theta(x) = 1 - G_{x+1}(\theta)$, gamma distribution.
Observation $X = 5$.

Plot of $p_{X}(\{\theta\})$ over a range of $\theta$. 

![Graph showing the Poisson distribution with observation $X = 5$.]
Curved normal example

- \( Y = (Y_1, \ldots, Y_n) \) iid \( \text{N}(\theta, \theta^2) \).\(^8\)
- A classical non-regular problem, \( \dim(\theta) < \dim(\text{suff stat}) \).
- Obvious association with sufficient statistics:

\[
\bar{X} = \theta + |\theta| U_1 \quad S = |\theta| U_2.
\]

- Two auxiliary variables for one parameter, inefficient.\(^9\)
- Rewrite association as

\[
\bar{X} = \theta + S \frac{U_1}{U_2} \quad \text{and} \quad \frac{\bar{X}}{S} = \frac{\text{sign}(\theta) + U_1}{U_2}.
\]

- Function of \((U_1, U_2)\) is observed.

---

\(^8\)Without (much) loss of generality, assume sign(\(\theta\)) is known.

Curved normal example, cont.

- Make a change-of-variable and condition on what’s observed.
- New association, A-step,

\[ \bar{X} = \theta + S V, \]

where \( V \) has the conditional distribution of \( U_1/U_2 \), given...

- For P-step, let \( f \) be (conditional) density of \( V \), and take \( S \) to be the random level sets

\[ S = \{v : f(v) \geq f(\tilde{V}), \tilde{V} \sim f\} \]

- C-step is just like before and, e.g.,

\[ \text{pl}_{X}(\theta) \approx \frac{1}{M} \sum_{m=1}^{M} 1\{f(\frac{\bar{X} - \theta}{S}) \geq f(\tilde{V}_m)\}, \quad \tilde{V}_m \sim f. \]
Curved normal example, cont.

- Experiment: \( n = 10 \) and \( \theta = 2 \).
- Data: \( \bar{X} = 1.97 \) and \( S = 2.48 \).
- Plausibility function \( p_{\bar{X}}(\theta) \) and 95% plausibility interval.
IM is valid by construction, but what about efficient?

Compare to a (generalized) fiducial solution which is (at least) first-order accurate asymptotically.

Repeat experiment above; table shows estimated coverage probability and mean lengths of 95% intervals.

<table>
<thead>
<tr>
<th>Coverage probability</th>
<th>Mean length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiducial</td>
<td>0.932</td>
</tr>
<tr>
<td>IM</td>
<td>0.956</td>
</tr>
</tbody>
</table>

Fiducial seems to be slightly shorter, but it under-covers.
Let $X = (X_1, \ldots, X_n)$ be iid $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma)$. 

Goal is inference on $\psi = \sigma / \mu$, coefficient of variation. 

$\mu$ in the denominator makes this a difficult problem. 

IM approach: 

- builds a reduced-association:

$$\frac{n^{1/2} \bar{X}}{S} = F_{n,1/\psi}^{-1}(U), \quad U \sim \text{Unif}(0,1).$$

- this yields a (marginal) IM for $\psi$ 
- automatically valid and gives very reasonable results

Full details are slightly too lengthy to present here...

---

Normal CV example, cont.

- Simulate data of size \( n = 30 \) from \( N(\mu, \sigma^2) \), with \( \sigma = 1 \).
- Plot of plausibility function for \( \psi \); based on “default” \( S \).
- Left: \( \mu = 1 \); right: \( \mu = 0 \).
Simulation: \( n = 10 \) from \( \text{N}(0.1, 1) \), so that \( \psi = 10 \).

Consider \( A = (\infty, 9] \), which does not contain \( \psi = 10 \).

Compute CDFs of Bayes, fiducial, CD, and IM beliefs.\(^\text{11}\)

\(^{11}\)My fiducial and CD approaches here are “naive”...
Valid beliefs yield nominal confidence regions.

Interesting reversal:
- Given a nested family of confidence regions, create a function by stacking up its contours
- Function has the shape a plausibility function and, in fact, it corresponds to a valid belief function.

So, valid belief functions are fundamental to statistics.
- We always tell students that confidence isn’t probability, but we never tell them what confidence IS.
- Confidence is belief/plausibility: a confidence set contains all parameter values that are sufficiently and justifiably plausible based on the observed data.
- I teach my students this explanation.\(^{12}\)

\(^{12}\text{M., arXiv:1606.02352}\)
So far, we have seen that

- the IM approach is a way to construct valid beliefs,
- and confidence regions correspond to valid beliefs.

Not clear, however, if every confidence region corresponds to a valid belief that can be constructed via an IM.

That is, there could be a gap between “good” beliefs and those that can be reached via IMs.

But it turns out that there is no gap:

- (basically) every nominal confidence region can be derived from a (“glued”) IM,
- and there’s an algorithm for constructing the IM.

This is a sort of converse to the IM validity theorem.
Quick interlude about random sets

- A wide class of random sets $\mathcal{S}$ that all lead to a valid IM are those we call *admissible*, i.e.,
  - the support $\mathcal{S}$ of $\mathcal{S}$ is nested, and
  - $P_{\mathcal{S}}$ satisfies $P_{\mathcal{S}}(\mathcal{S} \subseteq K) = \sup_{\mathcal{S} \subseteq K} P_U(S)$.

- By restricting to this class, we see that the random set is completely determined by its support $\mathcal{S}$ (and the measure $P_U$).

- So if the goal is to construct a valid IM that satisfies such-and-such condition, then we only need to construct a suitable support.
There is no reason that one has to use a single $S$. Consider $\{(S_\vartheta, P_{S|\vartheta}) : \vartheta \in \Theta\}$, admissible for each $\vartheta$.

Do IM construction to get a $\vartheta$-dependent plausibility

$$p_{x}(\theta \mid \vartheta) = P_{S|\vartheta}\{\Theta_{x}(S) \ni \theta\}, \quad \theta \in \Theta.$$ 

Define a “glued IM” with plausibility

$$p_{x}(A) = \sup_{\vartheta \in A} p_{x}(\vartheta \mid \vartheta).$$

Validity of this glued IM follows immediately from that of each $\vartheta$-specific IM.
This glued IM idea may seem a bit strange.

Has appeared in a different form in some of my papers, under the name of a “local conditional IM.”

This additional flexibility is especially useful for reducing the dimension of $U$ in certain non-regular cases:

- a conditioning operation helps improve efficiency, but may not be able to find a conditioning event free of $\theta$;
- roughly, our idea is to absorb the parameter dependence in the conditioning event into the family of random sets.

No time for an example here... :(
Complete-class theorem

- From a consonant plausibility on $\Theta$, we can get a *marginal plausibility* for $\psi = \psi(\theta)$ via optimization:

$$\text{mpl}_x(\psi) = \sup_{\theta : \psi(\theta) = \psi} \text{pl}_x(\theta).$$

- Marginal plausibility region: $\mathcal{P}_\alpha(x) = \{\psi : \text{mpl}_x(\psi) > \alpha\}$.

**Theorem (M. arXiv:1707.00486).**

Given nominal confidence regions $C_\alpha$ for $\psi = \psi(\theta)$ that are nested and satisfy a *compatibility condition*, there exists a valid (glued) IM on $\Theta$ such that the marginal plausibility region above satisfies

$$\mathcal{P}_\alpha(x) \subseteq C_\alpha(x) \quad \text{for all } (\alpha, x).$$
Complete-class theorem, cont.

- Binomial example: $n = 25$ and $x = 17$.
- Start with Clopper–Pearson intervals, $C_\alpha(x)$.
- Deduced plausibility region from glued IM is smaller!
Complete-class theorem, cont.

- Idea of the proof:
  - use $C_\alpha$’s to construct the random set’s support, i.e.,
    $$S_\alpha(\vartheta) = \{ u : C_\alpha(a(\vartheta, u)) \ni \psi(\vartheta) \}, \quad \alpha \in (0, 1), \quad \vartheta \in \Theta$$
  - yields a (glued) IM on $\Theta$, with consonant plausibility;
  - marginalize to $\psi$ via optimization.

- Compatibility condition:
  - Special case where $C_\alpha$ depends only on minimal sufficient statistic $X$ of the same dimension as $\theta$.
  - Need “$x = a(\theta, u)$” such that $S_\alpha(\vartheta) \neq \emptyset$ for all $(\vartheta, \alpha)$.

- Therefore: every good confidence region corresponds to a valid (glued) IM.
Concluding remarks

- Non-additive beliefs are fundamental to statistics.
- That is, additive beliefs
  - put user at risk of false confidence
  - lack a direct connection to relevant statistical properties.
- I focused on connection to confidence but a similar things can be said about p-values and IM plausibilities.\(^{13}\)
- Main drawback of IM approach is that it’s not always easy to do; we still just don’t understand it that well.
- Lots of interesting questions that remain to be answered:
  - marginalization mysteries
  - “optimal” random sets
  - incorporating prior info without losing validity
  - ...

\(^{13}\) M. and Liu, arXiv:1211.1547.
Thank you!

rgmarti3@ncsu.edu
www4.stat.ncsu.edu/~rmartin
www.twitter.com/@statsmartin

A new open-access publication platform is now open, featuring an *author-driven* peer review process.

**RESEARCHERS.ONE**

For more details, check out

www.researchers.one
www.twitter.com/@ResearchersOne