

Confidence in confidence distributions!

March 2020

Céline Cunen¹, Nils Lid Hjort¹, Tore Schweder²

¹Department of Mathematics, University of Oslo

²Department of Economics, University of Oslo

Abstract

The recent article ‘Satellite conjunction analysis and the false confidence theorem’ (Balch, Martin, and Ferson, 2019, this journal) points to certain difficulties with Bayesian analysis when used for models for satellite conjunction and ensuing operative decisions. Here we supplement these previous analyses and findings with further insights, uncovering what we perceive of as being the crucial points, explained in a prototype setup where exact analysis is attainable. We also show that a different and frequentist method, involving confidence distributions, is free of the false confidence syndrome.

Key words: confidence distributions; false confidence; inference for lengths; satellite conjunction probabilities

Bayesian analysis is often employed to calculate the collision probability of satellites based on noisy measurement of their position and speed. Balch, Martin, Ferson (2019), below referred to as BMF, have demonstrated that these analyses can be highly misleading, and they argue that this is an instance of a more general phenomenon, which they call *false confidence*. BMF find that the confidence in the event of no collision tends to get larger with noisier data even when the satellites actually are on a collision course. Such unfortunate properties are known to be present in some Bayesian analyses, cf. Schweder and Hjort (2016, e.g. Chapters 14, 15). We argue that such probability dilution is avoided when frequentist analysis by way of confidence distributions is employed. We demonstrate this below by revisiting the satellite collision problem. The purpose of this note is also to make clearer what ‘the heart of the matter’ is: in certain setups, even a sound-looking Bayesian analysis might have seriously unfortunate frequentist properties. Such consequences are particularly drastic when the methods are put to repeated use, as for monitoring of satellite collision probabilities.

Confidence distributions (CDs) are a type of inferential summary whose outward appearance is similar to a Bayesian posterior distribution. Just like a posterior for some parameter of interest δ , after having observed data y , a CD $C(\delta|y)$ defines a probability distribution on the space of possible δ values. The function $C(\delta|y)$ is a data-dependent cumulative distribution function. Unlike a posterior, however, a CD is required to have correct frequentist coverage, i.e. $C(\delta_0|y)$ is uniformly distributed when data are regarded as stochastic and δ_0 is the true value of the parameter.

The full confidence curve $C(\delta | y)$ is a practical summary of the uncertainty associated with the estimated parameter; for more on CDs see Schweder and Hjort (2016), Hjort and Schweder (2018). CDs and the related concept confidence curves are strongly related to confidence intervals. Since properly calibrated confidence intervals have guaranteed coverage, they are free of false confidence, and so are CDs, at least with respect to all interval statements, including confidence intervals and p-values.

CDs yield pictures of the shape and extent of the inferential uncertainty associated with a parameter of interest. These pictures are collections of confidence intervals, and are thus free of false confidence. CDs are less useful for obtaining confidence for all possible statements that can be made concerning the full parameter, as in a sense is the goal of the Inferential Model and the Bayesian method, where false confidence might occur at least in the Bayesian case.

After these brief general points we proceed to present an analysis of a simplified version of BMF’s satellite example. This simplification allows explicit expressions of the functions involved, but retains what we see as the heart of the matter, a crucial and problematic aspect of the Bayesian solution. Note that sophisticated satellite conjunction risk assessors are aware of the false confidence phenomenon, and the related issue of probability dilution, which affects the Bayesian calculation of collision probability. In practice, they therefore use this calculation not as a guarantee of safety but rather as a way to identify conjunctions that are likely to be dangerous. Their null hypothesis is that a collision will not occur and when the estimated collision probability is high the conjunction is identified as ‘dangerous’. See Hejduk, Snow, Newman (2019) for an extended treatment of these issues.

Following Martin (2019), assuming that the two satellites have constant and identical speed we can reduce the problem to two dimensions, with $\theta = (\theta_1, \theta_2)$ being the true differences between the unknown positions of the satellites along each axis on a plane. The interest lies in inference on the distance between the satellites, i.e. $\delta = \|\theta\| = (\theta_1^2 + \theta_2^2)^{1/2}$. We observe a single pair $y = (y_1, y_2)$ and assume that these two are independent and normal with means (θ_1, θ_2) and equal known variances σ^2 .

BFM present a Bayesian framework with flat priors, and in this simplified setup we get the simple posterior distribution where $\theta_i | y_i$ is a normal $N(y_i, \sigma^2)$ for the two components. From this posterior one can obtain a posterior for δ by integration, which is most conveniently presented in the form of a cumulative distribution function on the space of δ ,

$$B(\delta | y_1, y_2) = \Gamma_2\left(\frac{\delta^2}{\sigma^2}, \frac{\|y\|^2}{\sigma^2}\right), \quad (0.1)$$

writing $\Gamma_2(\cdot, \nu)$ for the cumulative distribution function of a non-central χ^2 with two degrees of freedom and noncentrality parameter ν . The component posterior noted above is also the natural CD for θ_1, θ_2 separately, but one cannot derive the confidence distribution of δ by manipulating the joint CD based on these two independent CDs. Instead, the CD of δ is found by considering the sampling distribution of $\hat{\delta} = (y_1^2 + y_2^2)^{1/2}$, which is a sufficient statistic for δ . One finds

$$C(\delta | y_1, y_2) = 1 - \Gamma_2\left(\frac{\|y\|^2}{\sigma^2}, \frac{\delta^2}{\sigma^2}\right). \quad (0.2)$$

Incidentally, this agrees with the belief in $[0, \delta]$, found by the Inferential Model methods of Martin and Liu (2015), for this situation. Note also that the marginalisation to δ is carried out in the observation space and not in the epistemic probability space (which is what is done in the Bayesian analysis).

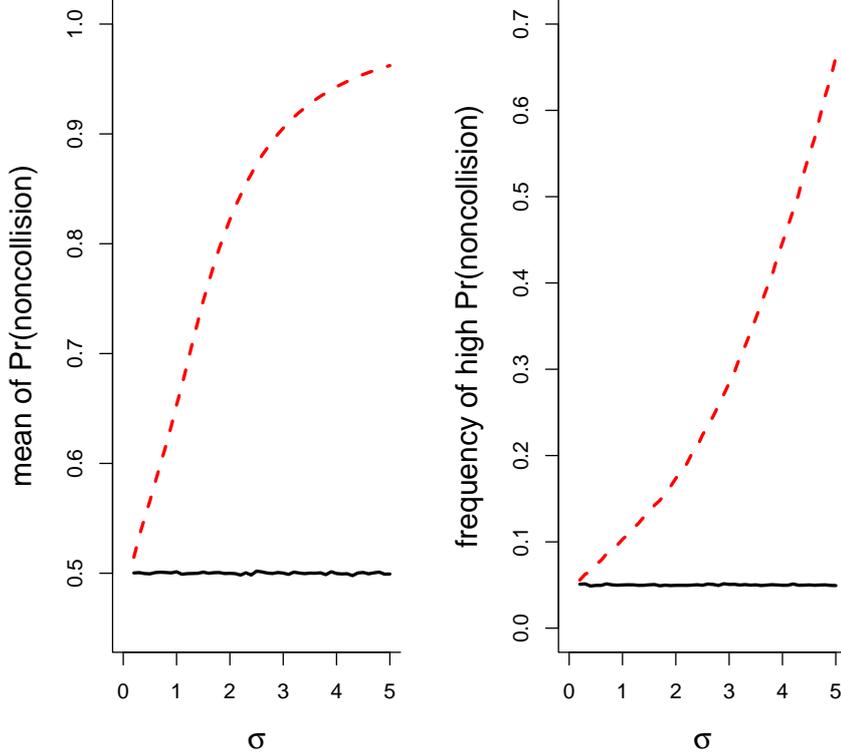


Figure 0.1: For each value of σ , we have computed the distribution of non-collision probabilities $1 - B(2.00 | y_1, y_2)$ and $1 - C(2.00 | y_1, y_2)$, with the Bayesian (slanted, red curves) and frequentist CD (full, black curves) methods. This is for a setup with $R = 2.00$ the threshold for collision and true value $\delta = 1.99$. Left panel: the means of these probabilities; right panel: the frequency of high probabilities, those above 0.95.

Let us briefly investigate the non-collision probabilities that these two methods produce. Collision is defined as the event that δ is smaller than R , the combined radius of the two satellites. We set $R = 2.00$ and $\delta = 1.99$ – a setup where the satellites really are on a collision course, but just barely so. For each of a set of values of σ we simulate 10^5 realisations of normal pairs (y_1, y_2) . For each such dataset we then calculate the probability of non-collision, according to each method, i.e. $P(\delta > 2.00 | y_1, y_2)$, which is equal to $1 - B(2.00 | y_1, y_2)$ for the Bayesian method, and $1 - C(2.00 | y_1, y_2)$ for the CD. Figure 0.1 displays frequentist properties of the Bayesian and the frequentist CD curves for two aspects of these non-collision probabilities; the left panel shows their mean non-collision probabilities, the right panel the frequency of high values, above 0.95. Exact calculations are incidentally also possible here.

The misleading behaviour of the Bayesian solution is apparent: as the observation noise increases, the Bayesian method reports higher and higher confidence in non-collision. In particular, the Bayesian method is *biased* in repeated use. The CD, on the other hand, reports probabilities that are correctly calibrated, in the sense that it wrongly indicates a high confidence in non-collision for only 5% of the datasets. A mean probability of non-collision of around 50% might seem high, but keep in mind that the data are drawn from a setup where the true distance is such that the

satellites barely collide. When the observation error is high there is a high chance of observing data that (wrongly) indicate a large estimated distance. This misleading property of the Bayesian probability of the satellites to not collide was identified by BMF and called false confidence.

The frequentist confidence $1 - C(R | y_1, y_2)$ is free of such false confidence. More formally we actually have that $1 - C(R | y_1, y_2)$ is uniformly distributed if the true minimal distance is R , i.e. if $\delta = R$. If however $\delta < R$ the distribution of $1 - C(R | y_1, y_2)$ will be shifted to the left of the uniform, and towards larger values if $\delta > R$. Note furthermore that $C(R | y_1, y_2)$ is the confidence we have in $\delta \in [0, R]$, i.e. in collision, after having observed our data; also, $1 - C(R | y_1, y_2)$ is the p-value for testing the null hypothesis that there will be no collision (see Schweder, 2018). Again the frequentist CD is fully calibrated, whereas the Bayesian machine provides misleading answers, the more so for increasing noise level.

It is also illuminating to see how the Bayesian cumulative $B(\delta | y_1, y_2)$ and frequentist $C(\delta | y_1, y_2)$ of (0.1) and (0.2) pan out in practice. In Figure 0.2 we display these cumulatives (left panel), in a situation with observed length $\|y\| = 5.00$, with assumed $\sigma = 2.50$. The right panel then shows the useful Bayesian and confidence curves, $|1 - 2B(\delta | y_1, y_2)|$ and $|1 - 2C(\delta | y_1, y_2)|$. These are convenient data summaries, for the most pertinent parameter, the δ . They ‘point’ to the median confidence estimates, 4.29 for the CD and 5.61 for Bayes, and also make it easy to read off confidence intervals. Here, with coverage level 0.90, the Bayesian interval $[2.01, 9.57]$ is unfortunate, in that it misses the true value $\delta = 1.99$; the $cc(\delta | y_1, y_2)$ fares rather better with its $[0.00, 8.63]$ interval.

The frequentist properties, also for a Bayesian statistical method, are of interest in contexts like the satellite collision problem, where potential collision events occur frequently. When the model parameter is given a vague prior, as there, but the basic parameter of interest is a nonlinear function of this underlying model parameter, problems like false confidence and bias might be present. When confidence can be measured in the frequentist spirit of Fisher and Neyman, as laid out in Schweder and Hjort (2016), there will never be any false confidence, and we can trust the obtained confidence!

Acknowledgements. We are grateful for comments from both Ryan Martin and from anonymous referees, which have contributed to a clearer presentation.

References

- Balch, M.S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society, A*, 475, issue 2227.
- Hejduk, M.D., Snow, D.E., and Newman, L.K. (2019). Satellite conjunction assessment risk analysis for ‘dilution region’ events: issues and operational approaches. In *Space Traffic Management Conference 28*, Austin, TX, February. commons.erau.edu/cgi/viewcontent.cgi?article=1294&context=stm
- Hjort, N.L. and Schweder, T. (2018). Confidence distributions and related themes. [General introduction article to a Special Issue, dedicated to this topic.] *Journal of Statistical Planning and Inference* **195**, 1–13.
- Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113, 39–73.
- Martin, R. and Liu, C. (2015). *Inferential Models: Reasoning with Uncertainty*. CRS Press, New York.
- Schweder, T. (2018). Confidence is epistemic probability for empirical science. *Journal of Statistical Planning and Inference* **195**, 116–125.

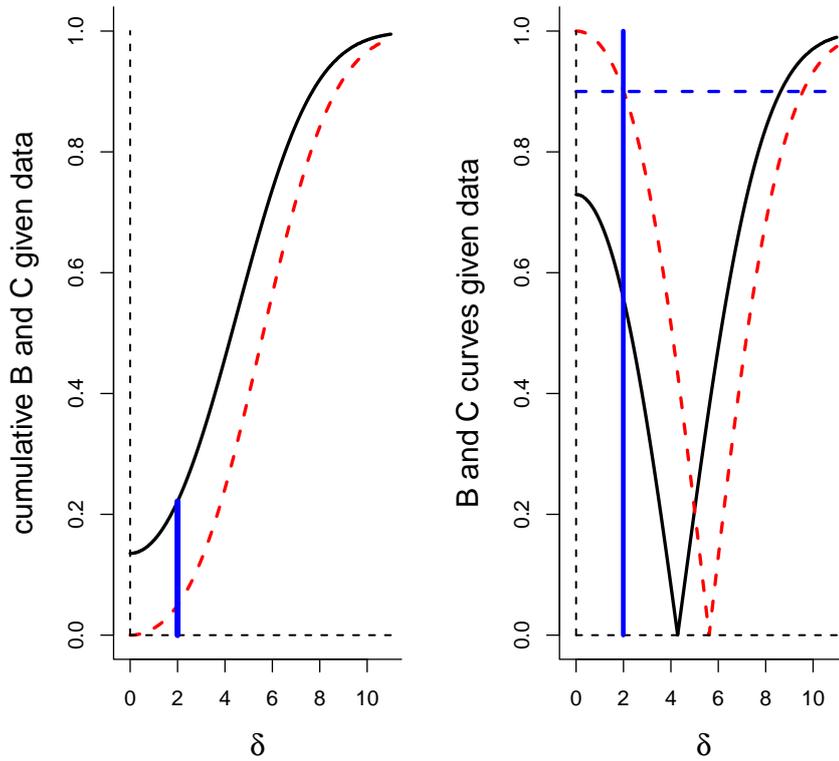


Figure 0.2: Left panel: the CD (black curve) and the Bayesian posterior cumulative for δ , after having observed $\|y\| = 5.00$, with assumed $\sigma = 2.50$; the critical value is $R = 2.00$ (marked blue), where we read off the confidence 0.222 in $[0, R]$. Right panel: the corresponding confidence curve $cc(\delta | y_1, y_2) = |1 - 2C(\delta | y_1, y_2)|$ (black) and the Bayesian credibility curve $|1 - 2B(\delta | y_1, y_2)|$ (red). Reading off 90% confidence and credibility intervals yields $[0.00, 8.63]$ and $[2.01, 9.57]$, respectively. The true $\delta = 1.99$ behind the generation of $\|y\|$ here is indicated by the vertical blue line.

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