Limitations of the "Limitations of proposed signatures of Bayesian confidence"

Balázs Hangya[1,2], Joshua I. Sanders[2,3] and Adam Kepecs[2]

[1]Institute of Experimental Medicine, Hungarian Academy of Sciences, Budapest, Hungary

[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States

[3]Sanworks LLC, Stony Brook, New York, United States

Abstract

Adler and Ma re-examine the proposed signatures of statistical confidence and argue that hidden assumptions in the mathematical derivations limit their applicability. We agree that if the conditions explicitly stated in our Theorems 2 and 4 are violated then the specific predictions differ. The counter-examples provided, however, create new assumptions, defining stimulus noise in a manner such that stimuli sometimes convey *opposite* evidence about the categories. Under these new assumptions, even if the decision maker is correct about the observed stimulus category, the choice may be labeled as incorrect. Thus, choice and confidence are not represented in a consistent way, and hence it is not an appropriate model for most decision tasks used in the field. Their other argument that our results do not provide sufficient conditions for the 'Bayesian Confidence Hypothesis' (BCH) is perplexing, since we made no mention of either BCH or sufficiency. In fact, we showed that the proposed signatures could be also produced by frequentist and bootstrap statistics, providing a direct demonstration that these do not imply BCH. Our framework surpasses previous algorithmic models not because the relationship between confidence and other variables would necessarily look different, but because of the generality of the statistical assumptions. Nevertheless, we appreciate the detailed examinations of the mathematical boundary conditions of our framework, which we see as complementary and not controversial.

Confidence is a variable internal to a decision maker, one that cannot be directly controlled or observed in experiments. Thus, it is challenging to appropriately identify confidence as a computation and to determine whether a particular behavioral response relies on this computation. To resolve this issue, numerous studies in neuroscience and psychology have employed quantitative models to formalize the decision maker's internal variable of "confidence" in terms of observable and quantifiable parameters. This definition enabled the field to study confidence reports in non-human animals (Drugowitsch et al., 2014; Kepecs et al., 2008; Kiani and Shadlen, 2009; Lak et al., 2014; Van Den Berg et al., 2016; Zylberberg et al., 2012), neural activity (Kepecs et al., 2008; Kiani and Shadlen, 2009; Komura et al., 2013) and even in pre-linguistic children (Goupil and Kouider, 2016).

We and others have observed distinct interrelationships among variables quantifying aspects of decision confidence. Curiously, these appeared to be very general across the types of confidence measurements we were conducting that spanned from perceptual to knowledge-based decisions and from rats to humans (Kepecs et al., 2008; Sanders et al., 2016). Furthermore, we found that several algorithmic models including signal detection theory (SDT) and some 'race' models (while not others Kiani et al., 2014) could reproduce these relationships. Therefore we wondered whether there was an underlying *law* that dictated these relationships. In Hangya et al. 2016 (HSK2016; Hangya et al., 2016) we showed that a statistical definition of decision confidence successfully explained these observations, providing a fairly general and, importantly, normative framework.

In their note, Adler and Ma (AM2017; Adler and Ma, 2017) show that under specific conditions, it is possible to violate the assumptions of our theorems 2 and 4 to generate different outcomes. We find this a useful analysis, since it explores the necessity of our mathematical boundary conditions in more depth, even if

restricted to specific examples. Their analysis expands on the necessity of the assumptions, which we discussed at the end of sections 2.3 and 2.5 in HSK2016. We also appreciate the shorter proof to our Lemma 1. In HSK2016, we pointed out the necessity of assumption #1 of theorem 2: 'belief independence' (p.1849). Similar observations can be noted for assumption #2 'percept monotonicity', which Adler and Ma explore in greater detail. They constructed an example by introducing 'stimulus noise' that violates this assumption (p.4 point 3.3) and therefore leads to a difference in the relationship of stimulus difficulty and confidence for incorrect choices. Exploring the necessity of the underlying assumptions and results under alternative conditions should always follow the formalization of a general model, which may eventually lead to even more general models – this is yet to be achieved for decision confidence. Nonetheless, below we point out some misinterpretations in the note by Adler and Ma.

1. Noise assumptions for the categorization task

The definition of stimulus and percept noise in AM2017's version of a categorization task is problematic. The examples rely on situations where stimulus noise can result in stimuli that convey the *opposite* evidence about the categories. Therefore, even if the decision maker is correct about the stimulus category, the choice may be labeled incorrect. This is not a rare exception, but rather these types of errors dominate in the examples presented (see Fig. 2E). The problem with this definition of 'stimulus noise' is that these choices cannot be correct by definition and hence choice and confidence no longer model mechanisms internal to the decision-maker. We deliberately avoided this issue in our framing of decision tasks: the randomly generated evidence was re-categorized after completion of the task so that choice correctness is assigned based on the actual evidence (stimulus plus stimulus noise). In other words, the examples presented by Adler and Ma do not represent choice and confidence in a consistent manner.

Note that in nearly all laboratory tasks aiming at understanding neural representations of stimuli and perceptual decisions, noise is not produced this way. For instance, Kiani & Shadlen and Brunton, Botvinick & Brody (Brunton et al., 2013; Kira et al., 2015; Scott et al., 2015) investigated the sources of noise in accumulating evidence during perceptual decisions. Every individual stimulus unit was modeled exactly, since introducing 'stimulus noise' would have limited their ability to understand neural sources of variability. The same principle applies to studies of decision confidence.

2. Is the average confidence in neutral evidence always 0.75?

In our theorem 4 we proved a surprising result: the average confidence in neutral evidence trials is precisely 0.75 under the conditions outlined on p 1850. We pointed out that the exact assumptions may not hold in real-life situations and went on to argue that "[*m]ore generally, this proof points to apparent overconfidence in percepts with neutral evidence in situations when the difficulty of the decisions cannot be determined. The degree of overconfidence will depend on the actual integrals involved*." Indeed, this apparent overconfidence of statistical confidence (and the not the precise value of 0.75) has significant implications for interpreting studies that show overconfidence in low discriminability and underconfidence in high-discriminability conditions, called hard-easy effect (Juslin et al., 2000; Merkle, 2009). This has been also pointed out by Drugowitsch and colleagues (Drugowitsch et al., 2014) for a different model class. Adler and Ma show that for their specific model, average confidence varies between 0.6 and 0.75 (Fig. 3) depending on the 'external noise' level (i.e. choice misclassification). This result further supports our claim that apparent overconfidence can be observed even for perfectly calibrated confidence **estimates.**

3. The Bayesian Confidence Hypothesis

Lastly, we are perplexed by AM2017's statements about sufficiency and the 'Bayesian Confidence Hypothesis'. We made no claims at all about sufficiency nor did we use the term 'Bayesian Confidence Hypothesis'. Although the definition of statistical confidence we use is 'Bayesian posterior probability', this does not reflect a commitment to a statistical school of thought but simply the use of conditional probability (and hence Bayes' Theorem). In fact, we showed that the proposed signatures were also produced by frequentist (t-test) and bootstrap statistics (HSK2016, Fig 3) making it explicit that our theorems do not imply sufficiency. In fact, many of these signatures were previously shown to be generated by other models

for confidence, such as signal detection theory (e.g. Kepecs et al., 2008), which is based on a different theoretical foundation and is considerably less general.

References

Adler, W. T., and Ma, W. J. (2017). Limitations of proposed signatures of Bayesian confidence. *bioRxiv*, 218222. doi:10.1101/218222.

Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and Humans Can Optimally Accumulate Evidence for Decision-Making. *Science (80-.).* 340, 95–98.

Drugowitsch, J., Moreno-Bote, R., and Pouget, A. (2014). Relation between Belief and Performance in Perceptual Decision Making. *PLoS One* 9, e96511.

Goupil, L., and Kouider, S. (2016). Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Curr. Biol.* 26, 3038–3045. doi:10.1016/j.cub.2016.09.004.

Hangya, B., Sanders, J. I., and Kepecs, A. (2016). A Mathematical framework for statistical decision confidence. *Neural Comput.* 28. doi:10.1162/NECO_a_00864.

Juslin, P., Winman, A., and Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychol. Rev.* 107, 384.

Kepecs, A., Uchida, N., Zariwala, H. A., and Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455, 227–231. doi:http://www.nature.com/nature/journal/v455/n7210/suppinfo/nature07200_S1.html.

Kiani, R., Corthell, L., and Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron* 84, 1329–1342.

Kiani, R., and Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science (80-.).* 324, 759–764.

Kira, S., Yang, T., and Shadlen, M. N. (2015). A Neural Implementation of Wald's Sequential Probability Ratio Test. *Neuron* 85, 861–873. doi:10.1016/j.neuron.2015.01.007.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., and Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat. Neurosci.* 16, 749–55. doi:10.1038/nn.3393.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., and Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* 84, 190–201. doi:10.1016/j.neuron.2014.08.039.

Merkle, E. C. (2009). The disutility of the hard-easy effect in choice confidence. *Psychon. Bull. Rev.* 16, 204–213.

Sanders, J. I., Hangya, B., and Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* 90. doi:10.1016/j.neuron.2016.03.025.

Scott, B. B., Constantinople, C. M., Erlich, J. C., Tank, D. W., and Brody, C. D. (2015). Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. *Elife* 4, 1–23. doi:10.7554/eLife.11308.

Van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., and Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *Elife* 5, 1–21. doi:10.7554/eLife.12192.

Zylberberg, A., Barttfeld, P., and Sigman, M. (2012). The construction of confidence in a perceptual decision. *Front. Integr. Neurosci.* 6. doi:10.3389/fnint.2012.00079.