LETTER Communicated by Laurence Aitchison

Limitations of Proposed Signatures of Bayesian Confidence

William T. Adler
will@wtadler.com
Center for Neural Science, New York University, New York, NY 10003, U.S.A.

Wei Ji Ma
weijima@nyu.edu
Center for Neural Science and Department of Psychology, New York University, New York, NY 10003, U.S.A.

The Bayesian model of confidence posits that confidence reflects the observer's posterior probability that the decision is correct. Hangya, Sanders, and Kepecs (2016) have proposed that researchers can test the Bayesian model by deriving qualitative signatures of Bayesian confidence (i.e., patterns that one would expect to see if an observer were Bayesian) and looking for those signatures in human or animal data. We examine two proposed signatures, showing that their derivations contain hidden assumptions that limit their applicability and that they are neither necessary nor sufficient conditions for Bayesian confidence. One signature is an average confidence of 0.75 on trials with neutral evidence. This signature holds only when class-conditioned stimulus distributions do not overlap and when internal noise is very low. Another signature is that as stimulus magnitude increases, confidence increases on correct trials but decreases on incorrect trials. This divergence signature holds only when stimulus distributions do not overlap or when noise is high. Navajas et al. (2017) have proposed an alternative form of this signature; we find no indication that this alternative form is expected under Bayesian confidence. Our observations give us pause about the usefulness of the qualitative signatures of Bayesian confidence. To determine the nature of the computations underlying confidence reports, there may be no shortcut to quantitative model comparison.

1 Introduction

Humans possess a sense of confidence about decisions they make, and asking human subjects for their decision confidence has been a common psychophysical method for over a century (Peirce & Jastrow, 1884). But despite the long history of confidence reports, it is still unknown how the brain computes confidence reports from sensory evidence. The leading proposal has been that observers' confidence reports are a function of only their

Neural Computation **30, 3327–3354** (2018)
doi:10.1162/neco_a_01141

posterior probability that their decision is correct (Drugowitsch, Moreno-Bote, & Pouget, 2014; Hangya, Sanders, & Kepecs, 2016; Kepecs & Mainen, 2012; Meyniel, Sigman, & Mainen, 2015; Pouget, Drugowitsch, & Kepecs, 2016), a hypothesis that we call the Bayesian confidence hypothesis (BCH) (Adler & Ma, 2018).

In recent years, some researchers have tested the BCH by formally comparing Bayesian confidence models to other models (Adler & Ma, 2018; Aitchison, Bang, Bahrami, & Latham, 2015). Although this is the most thorough method to test the BCH, it can be laborious in practice. One could instead try to describe signatures of the BCH—qualitative patterns that should theoretically emerge from Bayesian confidence—and then look for those patterns in real data. Hangya et al. (2016) propose four signatures, some of which have been observed in behavior (Kepecs, Uchida, Zariwala, & Mainen, 2008; Lak et al., 2014; Sanders, Hangya, & Kepecs, 2016) and in neural activity (Kepecs et al., 2008; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013).

These signatures are not unique to the Bayesian model; they are expected under a number of other models. Kepecs and Mainen (2012) argue that this is an advantage for a confidence researcher who is not interested in the precise algorithmic underpinnings of confidence. A researcher may observe these signatures in behavior, reasonably conclude that she has evidence that the observer is computing some form of confidence, and probe more deeply into, for instance, neural activity (Kepecs et al., 2008). In this letter, however, we consider the researcher concerned with understanding the computations underlying an observer's sense of confidence. We, along with Insabato, Pannunzi, and Deco (2016) and Fleming and Daw (2017), argue that for such a researcher, the fact that these signatures emerge from multiple models poses a problem. These signatures are not sufficient conditions for any particular model of confidence, including the Bayesian model. In other words, observation of these signatures does not constitute strong evidence in favor of any particular model. Because of this insufficiency, we view with skepticism any research that uses observation of these signatures as the basis for a claim that an observer uses a Bayesian (Navajas et al., 2017), "statistical" (Sanders et al., 2016), or any other specific form of confidence.

Although they do not claim that the signatures are sufficient conditions, Hangya et al. (2016) do claim that the signatures are necessary conditions for the BCH—that if confidence is Bayesian, these patterns will be present in behavior. Observation of a single necessary but not sufficient signature does not imply that the BCH is true; one would need to observe several signatures in order to gain confidence in the nature of confidence.[1]

[1] Restating this logic in probabilistic terms, a signature being a necessary condition for the BCH implies that $p(\text{signature observed} \mid \text{BCH is true}) = 1$. A signature being an insufficient condition implies that $p(\text{signature observed} \mid \text{BCH is false}) > 0$. By

The main contribution of this letter is to show that three signatures are not necessary conditions of Bayesian confidence, which reduces the overall value of the qualitative signature method for testing the BCH. We describe conditions under which these signatures are expected or not expected under the BCH. Researchers interested in Bayesian confidence should be aware of these conditions in order to avoid making one of two mistakes. First, a researcher who incorrectly believes that a signature is expected under the BCH will then incorrectly interpret the observation of a signature as positive evidence in favor of the Bayesian model. Conversely, if such a researcher fails to observe that signature, they will incorrectly rule out Bayesian confidence.

One signature is a mean confidence (i.e., the observer's estimated probability of being correct) of 0.75 on trials with neutral evidence. In section 3, we show that under the BCH, this signature will be observed only when stimulus distributions do not overlap and when noise is very low. Another signature is that as stimulus magnitude increases, mean confidence increases on correct trials but decreases on incorrect trials. In section 4, we show that under the BCH, this signature will be observed only when stimulus distributions do not overlap or when noise is high. (Readers who are interested only in nonoverlapping categories may skip section 4 or read it for intuition's sake.) For completeness, we briefly discuss insufficiency for both signatures. In section 5, we consider an alternative divergence signature recently proposed by Navajas et al. (2017). We show that this signature is not expected under the BCH. All code used for simulation and plotting is available at github.com/wtadler/confidence/signatures.

We hope that this letter will contribute some clarity and intuition to the study of Bayesian confidence.

2 Binary Categorization Tasks

We restrict ourselves to the following widely used family of binary perceptual categorization tasks (Green & Swets, 1966). On each trial, a category $C \in \{-1, 1\}$ is randomly drawn with equal probability. Each category corresponds to a category-conditioned stimulus distribution (CCSD) $p(s \mid C)$, where s could be, for example, an odor mixture (Kepecs et al., 2008), the net motion energy of a random dot kinematogram (Kiani & Shadlen, 2009; Newsome, Britten, & Movshon, 1989), the orientation of a Gabor (Adler & Ma, 2018; Denison, Adler, Carrasco, & Ma, 2018; Qamar et al., 2013), or the mean orientation of a series of Gabors (Navajas et al., 2017). The CCSDs are mirrored across $s = 0$: $p(s \mid C = -1) = p(-s \mid C = 1)$. Additionally, they are chosen such that a stimulus s is at least as likely to be drawn from category $C = 1$ as $C = -1$: $p(s \mid C = 1) \geq p(s \mid C = -1)$ for all $s \geq 0$.

Bayes's rule, for signatures that are both necessary and insufficient, p(BCH is true | signature(s) observed) will increase with the observation of each signature but will never reach 1.

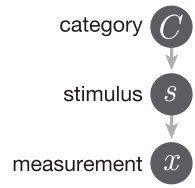

Figure 1: Generative model of the task.

A stimulus s is drawn from the chosen CCSD and presented to the observer. Observers do not have direct access to the value of s; instead, they take a noisy measurement x, drawn from the distribution $p(x \mid s, \sigma) = \mathcal{N}(x; s, \sigma)$, which denotes a gaussian distribution over x with mean s and standard deviation σ (see Figure 1).

If an observer's choice behavior is Bayesian (i.e., minimizes expected loss, which, in a task where each category has equal reward, is equivalent to maximizing accuracy), he computes the posterior probability of each category by marginalizing over all possible values of s: $q(C \mid x, \sigma) = \int q(C \mid s) q(s \mid x, \sigma) \mathrm{d}s$. In this letter, we use $p(\ldots)$ to refer to the true probability distributions used to, for example, generate stimuli and measurements and $q(\ldots)$ to refer to the observer's belief about such distributions. In some cases, $q(\ldots)$ may not equal $p(\ldots)$, a situation known as model mismatch (Acerbi, Vijayakumar, & Wolpert, 2014; Beck, Ma, Pitkow, Latham, & Pouget, 2012; Orhan & Jacobs, 2014).

After computing the posterior, observers make a category choice $\hat{C}$ by choosing the category with the highest posterior: $\hat{C} = \mathrm{argmax}_C q(C \mid x, \sigma)$. For the conditions described above, that amounts to choosing $\hat{C} = 1$ when $x > 0$, and $\hat{C} = -1$ otherwise (see appendix A).

Furthermore, if the observer's confidence behavior is Bayesian, it will be some function of the believed posterior probability of the chosen category. This probability is $q(C = \hat{C} \mid x, \sigma) = \max_C q(C \mid x, \sigma)$. Because it is a deterministic function of x and σ, we refer to it as $\mathrm{conf}(x, \sigma)$.[2] (See appendix B for derivations of $\mathrm{conf}(x, \sigma)$ for all stimulus distribution types used in this letter.)

3 0.75 Signature: Mean Bayesian Confidence Is 0.75 for Neutral Evidence Trials

Hangya et al. (2016) propose a signature concerning neutral evidence trials—those in which the stimulus s is equal to 0 (i.e., there is equal

[2] Note that our assumption that confidence and category choice are deterministic functions of x amounts to an assumption that there is no noise at the action (i.e., reporting) stage.

evidence for each category) and observer performance is therefore at chance. Bayesian confidence on each individual trial is always at least 0.5. One can intuitively understand why this is. In binary categorization, if the posterior probability of one option is less than 0.5, the observer makes the other choice, which has a posterior probability above 0.5. Therefore, all trials have confidence of at least 0.5, and mean confidence at any value of s is also at least 0.5. Hangya et al. (2016) go beyond these results and provide a proof that, under some assumptions, mean Bayesian confidence on neutral evidence trials is exactly 0.75. We refer to this prediction as the 0.75 signature, and we show that it is not always expected under a normative Bayesian model.

3.1 The 0.75 Signature Is Not a Necessary Condition for Bayesian Confidence. To determine the conditions under which the 0.75 signature is expected under the Bayesian model, we used Monte Carlo simulation with the following procedure. We generated an experiment in which all stimuli s were 0: $p(s \mid C) = \delta(0)$, where δ is the Dirac delta function. (For this analysis, the true generating distribution $p(s \mid C)$ does not matter; we could have instead used other distributions $p(s \mid C)$ and only analyzed trials in which s is very close to 0.) For a range of measurement noise levels σ, we drew measurements x from $p(x \mid s, \sigma) = \mathcal{N}(x; s = 0, \sigma)$. Using gaussian or uniform functions $q(s \mid C)$, we computed Bayesian confidence $\text{conf}(x, \sigma)$ for each measurement. We then took the mean confidence, equal to $\text{E}_{x|s=0}[\text{conf}(x, \sigma)]$.

The 0.75 signature holds only if the SD of the noise is very low relative to the range of the believed CCSD and if the observer has accurate knowledge of the low noise (see appendix D). Additionally, the subject must believe that the CCSDs are nonoverlapping (see Figure 2a, dotted line; any nonoverlapping CCSDs will do). If the observer believes that the CCSDs overlap by even a small amount, mean confidence on neutral evidence trials drops to 0.5. Therefore, in an experiment with overlapping CCSDs, one should not expect a Bayesian observer to produce the 0.75 signature. In experiments with nonoverlapping CCSDs, an observer's false belief might also cause him to not produce the 0.75 signature. We use the example of overlapping uniform CCSDs (see Figure 2a, solid lines) to demonstrate the fragility of this signature, although such distributions are not common in the literature. Overlapping gaussian CCSDs (see Figure 2b), however, are relatively common in the perceptual categorization literature (Adler & Ma, 2018; Ashby & Gott, 1988; Green & Swets, 1966; Norton, Fleming, Daw, & Landy, 2017; Qamar et al., 2013) and arguably more naturalistic (Maddox, 2002). Because the 0.75 signature requires both low measurement noise and the belief of nonoverlapping CCSDs, mean 0.75 confidence at neutral evidence trials is not a necessary condition for Bayesian confidence.

Additionally, the 0.75 signature is relevant only in experiments where subjects are specifically asked to report confidence in the form of a perceived

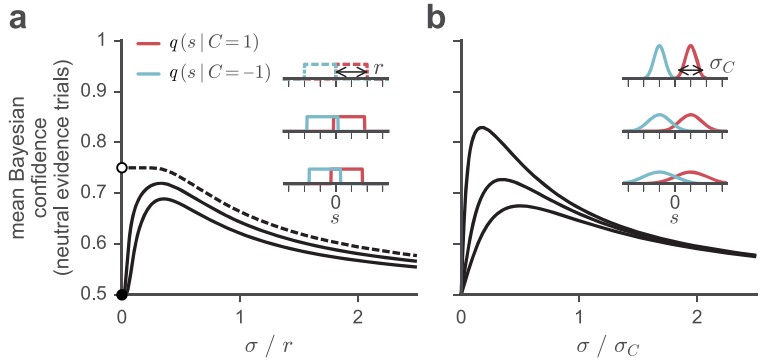

Figure 2: The 0.75 signature is not a necessary condition for Bayesian confidence. The y-axis indicates mean Bayesian confidence on trials for which $s = 0$. Each inset corresponds to a line, in the same top-to-bottom order. Dotted and solid lines indicate, respectively, the nonoverlapping and overlapping CCSDs that go into the observer's computation of confidence. For each value of σ, 50,000 trials were simulated. (a) Trials were simulated using believed uniform CCSDs defined by $q(s \mid C = 1) = \mathcal{U}(s; a, b)$, with $b - a = r = 2$; $q(s \mid C = -1)$ is mirrored across $s = 0$, as described in section 2. When the CCSDs are believed to be nonoverlapping (i.e., with $a = 0$ and $b = 2$, top inset), the 0.75 signature can be observed as measurement noise approaches 0 (dotted black line). However, mean Bayesian confidence decreases as a function of measurement noise. Additionally, when the distributions overlap slightly (bottom two insets), the 0.75 signature will not be observed (solid black lines). (b) Moreover, when the CCSDs are believed to be gaussian distributions defined by $q(s \mid C = 1) = \mathcal{N}(s; \mu_C = 1, \sigma_C)$, the 0.75 signature will not be observed at any σ_C or measurement noise level σ. One can intuitively understand why mean confidence is 0.5 for overlapping categories at very low measurement noise and increases with measurement noise. At very low measurement noise, the observer makes measurements that are very close to zero, which the observer "knows" are associated with a low probability of being correct. However, as noise increases, the observer starts to make measurements that have higher magnitude, leading the observer to believe that they have a higher probability of being correct. At high levels of noise, confidence starts to decrease.

probability of being correct (or are incentivized to do so through a scoring rule (Brier, 1950; Gneiting & Raftery, 2007; Massoni, Gajdos, & Vergnaud, 2014), although in this case, it has been argued (Adler & Ma, 2018; Ma & Jazayeri, 2014) that any Bayesian behavior might simply be a learned mapping). In other words, in an experiment where subjects are asked to report confidence on a scale of 1 to 5, a mean confidence of 3 only corresponds to 0.75 if one makes the a priori assumption that there is a linear mapping

between rating and perceived probability of being correct (Sanders et al., 2016).

3.1.1 Relevant Assumptions in Hangya et al. (2016). Hangya et al. (2016) describe an assumption that is critical for the 0.75 signature: each CCSD is a continuous uniform distribution. However, the 0.75 signature depends on two additional assumptions that they make implicitly. We reproduce their proof, drawing attention to those assumptions. For clarity, we remove σ from $\text{conf}(x, \sigma)$, $p(x \mid s = 0, \sigma)$, and $q(C = 1 \mid x, \sigma)$ as it is not necessary for the proof.

Using the definition of expected value and splitting the integral:

$$\begin{aligned}
\mathop{\mathrm{E}}_{x|s=0}[\text{conf}(x)] &= \int p(x \mid s = 0)\text{conf}(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{0} p(x \mid s = 0)\text{conf}(x)\,\mathrm{d}x + \int_{0}^{\infty} p(x \mid s = 0)\text{conf}(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{0} p(x \mid s = 0)q(C = -1 \mid x)\,\mathrm{d}x \\
&\quad + \int_{0}^{\infty} p(x \mid s = 0)q(C = 1 \mid x)\,\mathrm{d}x,
\end{aligned}$$

where they use the fact that for $x > 0$, confidence is equal to the posterior probability of $C = 1$, and for $x < 0$, confidence is equal to the posterior probability of $C = -1$. Next, they make use of the symmetry of $p(x \mid s = 0)$ about $x = 0$ and of the symmetry $q(C = -1 \mid -x) = q(C = 1 \mid x)$ to find

$$\mathop{\mathrm{E}}_{x|s=0}[\text{conf}(x)] = 2\int_{0}^{\infty} p(x \mid s = 0)q(C = 1 \mid x)\,\mathrm{d}x.$$

Next, Hangya et al. (2016) assume $q(C = 1 \mid x) = q(s > 0 \mid x)$. This is true only in the case of nonoverlapping categories, in which $C = 1$ is equivalent to $s > 0$:

$$\begin{aligned}
\mathop{\mathrm{E}}_{x|s=0}[\text{conf}(x)] &= 2\int_{0}^{\infty} p(x \mid s = 0)q(s > 0 \mid x)\,\mathrm{d}x \\
&= 2\int_{0}^{\infty} p(x \mid s = 0)\frac{q(s > 0, x)}{q(x)}\,\mathrm{d}x \\
&= 2\int_{0}^{\infty} p(x \mid s = 0)\left[\frac{\int_{0}^{\infty} q(x \mid \tilde{s})q(\tilde{s})\,\mathrm{d}\tilde{s}}{q(x)}\right]\mathrm{d}x. \qquad (3.1)
\end{aligned}$$

Next, Hangya (2016) assume that for $s > 0$, $q(s) = q(x) = k$, where k is a constant. We will comment on this assumption below. Under this assumption,

$$\mathop{\mathrm{E}}_{x|s=0}[\mathrm{conf}(x)] = 2\int_0^\infty p(x \mid s=0)\left[\int_0^\infty q(x \mid s)\,\mathrm{d}s\right]\mathrm{d}x. \tag{3.2}$$

Then they assume that $q(x \mid s) = p(x \mid s)$—that the observer has accurate knowledge of their measurement distribution—and apply a change of variables $\tilde{x} = x - s$:

$$\begin{aligned}\mathop{\mathrm{E}}_{x|s=0}[\mathrm{conf}(x)] &= 2\int_0^\infty p(x \mid s=0)\left[\int_0^\infty p(x \mid s)\,\mathrm{d}s\right]\mathrm{d}x \\ &= 2\int_0^\infty p(x \mid s=0)\left[\int_{-\infty}^x p(\tilde{x} \mid s=0)\,\mathrm{d}\tilde{x}\right]\mathrm{d}x.\end{aligned}$$

Finally, Hangya et al. (2016) use the following lemma: $\int_0^\infty f(t)F(t)\,\mathrm{d}t = \frac{3}{8}$, where $f(t)$ is a probability density function symmetric about zero, and its cumulative distribution function is $F(t) = \int_{-\infty}^t f(x)\,\mathrm{d}x$. (Incidentally, their proof of this lemma can be dramatically shortened. We present the shortened version in appendix C.) Then,

$$\mathop{\mathrm{E}}_{x|s=0}[\mathrm{conf}(x)] = 0.75,$$

concluding the proof.

The assumption that we want to draw attention to is $q(s) = q(x) = k$. This assumption is never exactly satisfied because such distributions would be improper (i.e., not normalizable on $\mathbb{R}$). However, we can relax the assumption to $q(s)$ being locally constant around $s = 0$ in a neighborhood that is large relative to the measurement noise $p(x \mid s)$. The reasoning is intuitively as follows: In equation 3.1, $p(x \mid s = 0)$ in effect filters out all values of x more than, say, 3σ away from $s = 0$. Thus,

$$\mathop{\mathrm{E}}_{x|s=0}[\mathrm{conf}(x)] \approx 2\int_0^{3\sigma} p(x \mid s=0)\left[\frac{\int_0^\infty q(x \mid \tilde{s})q(\tilde{s})\,\mathrm{d}\tilde{s}}{q(x)}\right]\mathrm{d}x. \tag{3.3}$$

As a consequence, we can assume that inside the $[\ldots]$, $x \in [0, 3\sigma]$. Applying the same 3σ buffer to $\tilde{s}$ around x, we approximate the inner integral as

$$\int_0^\infty q(x \mid \tilde{s})q(\tilde{s})\,\mathrm{d}\tilde{s} \approx \int_0^{6\sigma} q(x \mid \tilde{s})q(\tilde{s})\,\mathrm{d}\tilde{s}.$$

Similarly, the normalization is

$$q(x) = \int_{-\infty}^\infty q(x \mid \tilde{s})q(\tilde{s})\,\mathrm{d}\tilde{s}$$

$$\approx \int_{-3\sigma}^{6\sigma} q(x \mid \tilde{s}) q(\tilde{s})\, \mathrm{d}\tilde{s}.$$

If now $q(s) = k$ for $s \in [-3\sigma, 6\sigma]$, we can approximate the part inside the square brackets in equation 3.3 as

$$\frac{\int_0^\infty q(x \mid \tilde{s}) q(\tilde{s})\, \mathrm{d}\tilde{s}}{q(x)} \approx \frac{k \int_0^{6\sigma} q(x \mid \tilde{s})\, \mathrm{d}\tilde{s}}{k \int_{-3\sigma}^{6\sigma} q(x \mid \tilde{s})\, \mathrm{d}\tilde{s}} \approx \frac{\int_0^\infty q(x \mid \tilde{s})\, \mathrm{d}\tilde{s}}{1},$$

which brings us to equation 3.2. From there, the proof proceeds identically. Of course, the choice of a multiplier of 3 on σ is arbitrary, and $q(s)$ does not have to be exactly constant near 0, but the quality of the approximation relies on σ being small relative to the size of the neighborhood around 0 over which s is believed to be approximately constant. (A more rigorous proof would involve a Taylor expansion of $q(s)$ around x.)

In summary, we have highlighted two assumptions that are required for Hangya et al.'s (2016) proof of the 0.75 signature: first, that the observer believes the CCSDs are nonoverlapping, and second, that measurement noise is negligible relative to the size of the neighborhood around zero over which s is believed by the observer to be constant. If either assumption is violated, the proof does not apply, and the 0.75 signature is not expected under the BCH.

3.2 The 0.75 Signature Is Not a Sufficient Condition for Bayesian Confidence. We have shown that the 0.75 signature is not a necessary condition for Bayesian confidence, but is it a sufficient condition? It is possible to show that a signature is a sufficient condition if it is not possible to observe it under any other model. One could put forward a trivial model that always produces exactly midrange confidence on each trial, regardless of the measurement. Therefore, the 0.75 signature is not a sufficient condition.

4 Divergence Signature 1: As Stimulus Magnitude Increases, Mean Confidence Increases on Correct Trials But Decreases on Incorrect Trials

Hangya et al. (2016) propose the following pattern as a signature of Bayesian confidence. On correctly categorized trials, mean confidence is an increasing function of stimulus magnitude (here, $|s|$), but on incorrect trials, it is a decreasing function (see Figure 3a). We refer to this pattern as

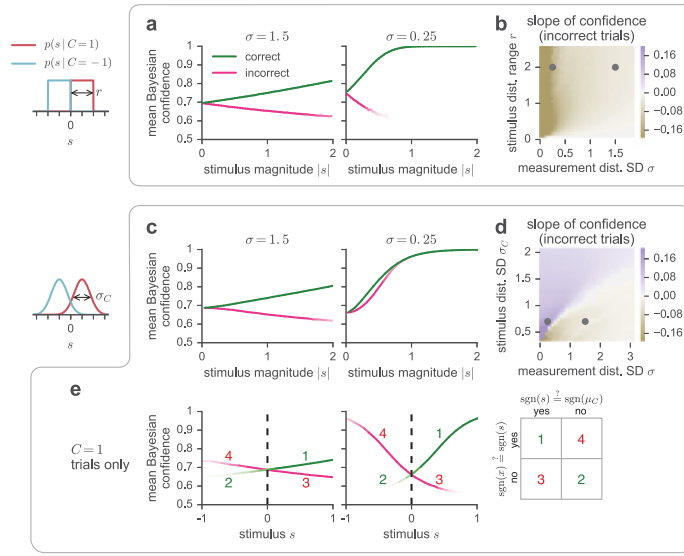

Figure 3: Divergence signature 1 is not a necessary condition for Bayesian confidence. For two stimulus distribution types, we simulated 2 million trials. (a) With uniform stimulus distributions defined by $p(s \mid C = 1) = \mathcal{U}(s; 0, 2)$, the divergence signature is predicted under both high- and low-noise regimes. The fadedness of the line indicates conditions for which there are few trials. (b) The heat map indicates the slope of the pink lines in panel a. At all values of σ and distribution range, the slope is negative. Slopes were obtained by generating binned mean confidence values as in panel a and fitting a line to those values. Black markers indicate the parameters used in panel a, with the left dot corresponding to the right plot and the converse. (c) With gaussian stimulus distributions defined by $p(s \mid C = 1) = \mathcal{N}(s; 1, \sigma_C = 0.7)$, the divergence signature appears only when measurement noise is high (i.e., when $\sigma \gtrsim 0.6$). (d) As in panel b, but for gaussian distributions with means of ± 1. Under some values of σ and σ_C, the slope is positive, indicating that the divergence signature is not a necessary condition for Bayesian confidence. (e) Visual explanation for why, under gaussian stimulus distributions, the divergence signature appears only at relatively high σ values. Plots represent the same data as in panel c, but over s instead of $|s|$. For clarity, we use only trials drawn from category $C = 1$; the argument is mirrored for $C = -1$. Incorrect trials fall into two categories: on trials in which s is positive but x is negative due to noise, confidence goes down as $|s|$ increases (branch 3); on trials in which s and x are both negative, confidence increases with $|s|$ (branch 4). At high levels of noise, branch 3 has more trials than branch 4 and dominates the averaging that occurs when plotting trials from both categories over $|s|$. At low levels of noise, branch 4 instead dominates and the divergence signature disappears. Note that for nonoverlapping distributions (e.g., those in panels a and b), there are no trials in which s has a different sign from the stimulus distribution mean, so branches 2 and 4 do not exist, and the divergence signature is always present.

divergence signature 1.[3] For the rest of the letter, we use *divergence* to refer to the pattern of confidence as an increasing function of some variable on correct trials and a decreasing function on incorrect trials.[4]

Divergence signature 1 has been observed in some behavioral experiments (Kepecs et al., 2008; Komura et al., 2013; Lak et al., 2014; Sanders et al., 2016). However, we demonstrate that as with the 0.75 signature (see section 3), the signature is not always expected under the BCH.[5] Therefore, the appearance of the signature in these papers should not be taken to mean that it should be generally expected.

4.1 Divergence Signature 1 Is Not a Necessary Condition for Bayesian Confidence. In this section, we argue that divergence signature 1 is expected only under specific conditions on the stimulus distribution $p(s \mid C = -1)$ and the noise distribution $p(x \mid s, \sigma)$.

4.1.1 Stimulus Distribution Type. To determine the conditions under which the divergence signature is expected under the Bayesian model, we used Monte Carlo simulation with the following procedure. We generated stimuli s, drawn with equal probability from stimulus distributions $p(s \mid C = -1)$ and $p(s \mid C = 1)$. We generated noisy measurements x from these stimuli, using measurement noise levels σ. We generated observer choices from these measurements, using the decision rule of choosing $\hat{C} = 1$ when $x > 0$. We computed Bayesian confidence for every trial, assuming that the observer has accurate knowledge of their measurement distributions and of the CCSDs: $q(\ldots) = p(\ldots)$.

Nonoverlapping uniform CCSDs. We first consider the case of CCSDs that are uniform on an interval and do not overlap. This is an example covered by Hangya et al.'s (2016) proof. Indeed, we find in simulations that divergence signature 1 is expected under the Bayesian model in both high- and low-noise regimes (see Figures 3a and 3b). The intuition for why this pattern occurs is as follows. On correct trials, as stimulus magnitude increases, the mean magnitude of the measurement x increases. Because measurement

[3] Kepecs and Mainen (2012), Insabato, Pannunzi, and Deco (2016), and Fleming and Daw (2017) call it the (folded) X-pattern.

[4] The term *divergence* does not normally imply opposite trends. For example, the lower function could be flat or even increasing. However, we could not think of a better one-word alternative.

[5] Our finding is distinct from that of Insabato et al. (2016), who show that the signature would not be predicted under a non-Bayesian model in which the observer uses two measurements on each trial. Our analyses concern only Bayesian models in which the observer has a single measurement on each trial.

Our finding is also distinct from that of Fleming and Daw (2017), who show that the divergence signature would not be predicted if the experimenter could plot confidence as a function of the internal measurement x. Our analyses concern confidence only as a function of the stimulus s, which, unlike x, is known by the experimenter.

magnitude is monotonically related to Bayesian confidence, this increases mean confidence. However, on incorrect trials (in which x and s have opposite signs), the mean magnitude of the measurement decreases (see Figure 5a), which in turn decreases mean confidence (see Figures 5b and 5c). The proof by Hangya et al. (2016) and the intuition are not limited to uniform CCSDs (truncated gaussians will also work, for example), but do require the CCSDs to be nonoverlapping. When the stimulus distributions are nonoverlapping, divergence is expected under any level of measurement noise (see Figures 3a and 3b).

Gaussian CCSDs. We now consider gaussian CCSDs. In this case, when measurement noise is high relative to stimulus distribution width (see Figure 3c, left), the signature is still expected. However, when measurement noise is low relative to stimulus distribution width, the divergence signature is not expected (see Figures 3c and 3d). To gain intuition for why this is, imagine an optimal observer with zero measurement noise. In tasks with overlapping categories, even this observer cannot achieve perfect performance; some trials from category $C = 1$ will have negative s and x values, resulting in an incorrect choice. For such stimuli, confidence increases with stimulus magnitude. At relatively low noise levels, these stimuli represent the majority of all incorrect trials for category $C = 1$ (see Figure 3e, right). This effect causes the divergence signature to disappear when plotting over $|s|$, that is, averaging over errors with positive and negative s. In this particular case, an experimenter could "rescue" the signature by plotting confidence as a function of signed stimulus value s for a given true category. This would produce plots such as Figure 3e (right), which have a characteristic crossing pattern. Researchers using more unusual categories than the ones presented here might consider running simulations to see if the signature is expected and, if not, whether this method could "rescue" the signature in their case.

4.1.2 Relevant Assumption in Hangya et al. (2016). The gaussian CCSD example shows that divergence signature 1 is not a necessary condition for Bayesian confidence. By contrast, the proof in Hangya et al. (2016) seems quite general. We can resolve this paradox by making explicit the assumptions hidden in the proof. The authors assume that "for incorrect choices . . . with increasing evidence discriminability, the relative frequency of low-confidence percepts increases while the relative frequency of high-confidence percepts decreases" (p. 1847).[6] This assumption is violated in the case of overlapping gaussian stimulus distributions. For some incorrect

[6] Earlier in their paper, Hangya et al. (2016) phrase this assumption as, "For any given confidence c, the relative frequency of percepts mapping to c by ξ changes monotonically with evidence discriminability for any fixed choice" (p. 1847). In our terminology, this is equivalent to saying that as $|s|$ increases, the frequency of reporting any particular level of confidence changes monotonically. This is not correct even in the case of nonoverlapping

choices (see branch 4 of Figure 3e), as s becomes more discriminable (i.e., very negative), the frequency of high-confidence reports increases. At low levels of measurement noise, this causes the divergence signature to disappear when plotting over $|s|$.

4.2 Divergence Signature 1 is Not a Sufficient Condition for Bayesian Confidence. It has been previously noted that the signature is expected under a number of non-Bayesian models (Fleming & Daw, 2017; Insabato et al., 2016; Kepecs & Mainen, 2012). Here, we describe an additional non-Bayesian model—one in which confidence is a function only of $|x|$, the magnitude of the measurement (Kepecs et al., 2008). Previous studies have referred to similar models as Fixed (Adler & Ma, 2018; Denison et al., 2018; Qamar et al., 2013) or Difference (Aitchison et al., 2015). In the general family of binary categorization tasks described in section 2, the confidence of this model is monotonically related to the confidence of the Bayesian model $\text{conf}(x, \sigma)$. Thus, when divergence signature 1 is predicted by the Bayesian model, it is also predicted by this measurement model, underscoring that the divergence signature is not a sufficient condition for Bayesian confidence.

5 Divergence Signature 2: As Measurement Noise Decreases, Mean Confidence Increases on Correct Trials but Decreases on Incorrect Trials

Navajas et al. (2017) conduct an experiment in which they present, on each trial, a sequence of oriented Gabors with orientations pseudo-randomly drawn from a uniform distribution on an interval, with the range of the interval chosen randomly from four possible values. They then ask subjects to judge whether the mean orientation is left or right of vertical and to provide a confidence report. They plot mean confidence (conditioned on correctness) as a function of stimulus range. Data from some of their subjects show strongly divergent confidence (i.e., oppositely signed slopes for confidence on correct and incorrect trials), but their averaged data (see Figure 4a) do not.

Navajas et al. (2017) write that normative arguments would lead one to expect a diverging pattern, citing Hangya et al. (2016). However, Hangya et al. (2016) show that divergence is expected only when the x-axis is stimulus magnitude, not stimulus distribution range. Because of this difference, we treat a divergence in this kind of plot as a new possible signature, which we call divergence signature 2. For this to be a signature of Bayesian

uniform stimulus distributions. For example, at low noise, as discriminability increases, the frequency of medium-confidence reports will increase and then decrease. Therefore, we will use the formulation of the assumption further down on p. 1847, which correctly narrows it down to incorrect choices.

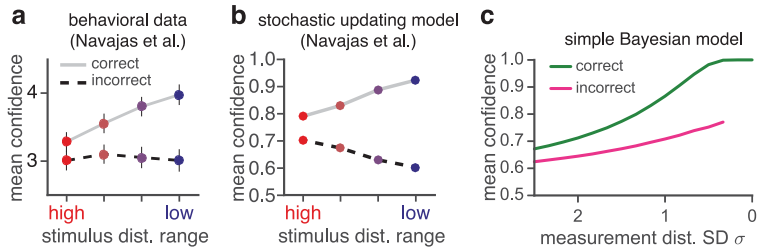

Figure 4: Divergence signature 2 is predicted by Navajas et al.'s (2017) stochastic updating model but is not present in either their data or the prediction of a simple Bayesian model. (a) Averaged confidence data in their perceptual task do not show the signature. (b) Navajas et al. (2017) build a stochastic updating model that does predict divergence signature 2. (c) Mean Bayesian confidence as a function of measurement noise is not expected to show opposite trends when conditioned on correctness, suggesting that divergence signature 2 might not be generally expected. At each value of σ, 50,000 stimuli were stimulated, with $s = \pm 1$. (Panels a and b adapted by permission from Macmillan Publishers Ltd: *Nature Human Behaviour*, Navajas et al., 2017.)

confidence, we would have to show that a Bayesian model would predict this pattern. We show that this pattern is not necessarily expected under the BCH.

5.1 Navajas et al.'s (2017) Stochastic Updating Model. Instead of a Bayesian model, Navajas et al. (2017) use a model that on each trial, updates a variable μ that is meant to be the estimate of the mean orientation,

$$\mu_i \sim \mathcal{N}\left((1-\lambda)\mu_{i-1} + \lambda\theta_i, \sigma_i\right), \tag{5.1}$$

where μ_i is the estimate after i samples ($\mu_0 = 0$), θ_i is the ith orientation stimulus in the sequence, and λ is a constant between 0 and 1. Navajas et al. (2017) incorporate into their model an assumption of orientation-dependent noise (Girshick et al., 2011) by setting $\sigma_i = \gamma|\theta_i|$, where γ is a free parameter indicating the strength of the noise. Because the SD of each update is proportional to $|\theta_i|$, more tilted orientations are measured with greater noise, and trials drawn from distributions with greater range therefore have lower performance. (Their subjects performed worse on trials with greater stimulus distribution range but, without orientation-dependent noise, their model would perform equally well on each condition.) Because of this relationship, we will also use "divergence signature 2" to refer to confidence divergence (conditioned on correctness) as a function of measurement noise.

Navajas et al. (2017) then derive their measure of confidence from this decision variable. After fitting, this model produces a diverging pattern (see

Figure 4b). Because this pattern is not present in their averaged data (see Figure 4a), they conclude that the stochastic updating model is inadequate. To account for the discrepancy, they then incorporate Fisher information into their model, which produces a better fit; the authors' main result relies on analysis of the parameters of this "hybrid" model.

Critically, however, the stochastic updating model is not a Bayesian model. Under a Bayesian model, each θ_i would contribute equally to the final estimate of the mean. For that to follow from equation 5.1, λ would have to equal $\frac{1}{i}$. However, their λ is not i-dependent. Therefore, μ_i is not the decision variable that a Bayesian observer would base either choice or confidence on. The fact that the stochastic updating model is not Bayesian has two implications. First, the stochastic updating model producing divergence signature 2 does not imply that it is expected under the BCH. Second, the deviation of their model predictions from the data does not provide any evidence against the BCH.

5.2 Simple Bayesian Model. We constructed a simple Bayesian model to test whether divergence signature 2 is generally expected under the BCH. Our model does not include an updating component because the temporal dynamics in this task are irrelevant for optimal choice and confidence.

In Navajas et al. (2017), the mean of all the stimuli presented on each trial is forced to be either 3° or -3°. Accordingly, we generated stimuli with $s = \pm 1$, corresponding to $C = \pm 1$. In our model, we drew noisy measurements x from $p(x \mid s, \sigma) = \mathcal{N}(x; s, \sigma)$. Under Navajas et al.'s (2017) assumption of orientation-dependent noise, draws from distributions with greater range are measured with higher levels of noise. We build this assumption into our simple model by using σ as a proxy for stimulus distribution range. Higher values of σ correspond to trials drawn from distributions with greater range. As described in section 4.1.1, we generated observer choices and computed Bayesian confidence assuming that the observer has accurate knowledge of their measurement distributions and of the CCSDs.

We find that as measurement noise decreases, mean confidence increases for both correct and incorrect trials (see Figure 4c). This pattern also holds when the category-conditioned stimulus distributions are uniform or gaussian and if one plots a measure of stimulus distribution variance on the x-axis (either uniform distribution range r or gaussian distribution SD σ_C). This indicates that divergence signature 2 is not necessarily expected under the BCH.

We emphasize that we are not claiming that Navajas et al.'s (2017) data are best explained by a Bayesian model. In fact, just as they use Fisher information to bend the predictions of their stochastic updating model (see Figure 4b) upward to fit their data (see Figure 4a), our simulation (see Figure 4c) suggests that a post hoc addition to our Bayesian model would have to bend the predictions downward. However, our goal is not to fit their data

but merely to show that divergence signature 2 is not necessarily expected under a Bayesian model. There are several ways in which we can imagine constructing a more complete Bayesian model of their task. For example, the observer might marginalize over the nuisance parameter of stimulus range when computing confidence. Determining whether confidence in Navajas et al.'s (2017) data is Bayesian would thus require careful quantitative model comparison.

We also note that in our Bayesian model, the observer has accurate knowledge of their own measurement noise, which may not be the case for the observers in Navajas et al. (2017) However, even when observers have incorrect beliefs about their measurement noise, the pattern of mean confidence still does not show divergence as in Figure 4b (see appendix D).

5.3 Why the Intuition for Divergence Signature 1 Does Not Predict Divergence Signature 2. We have shown that although divergence signature 1 is not completely general, it is expected under the BCH in some cases (see Figure 3a). By contrast, we have no indication of whether divergence signature 2 is ever expected from simple Bayesian models, such as the one described in section 5.2, when plotting measurement noise on the x-axis. This may be surprising, because the intuition for divergence signature 1 might seem to apply equally to this case. However, the effect of measurement noise on mean confidence is different from the effect of stimulus magnitude because measurement noise, unlike stimulus magnitude, affects the mapping from measurement to confidence on a single trial.

Mean Bayesian confidence is a function of two factors: confidence on a single trial and the probability of the corresponding measurement:

$$\mathop{\mathrm{E}}_{x}[\mathrm{conf}(x,\sigma)] = \int p(x \mid s,\sigma)\,\mathrm{conf}(x,\sigma)\,\mathrm{d}x.$$

The intuition for divergence signature 1 is as follows. As stimulus magnitude $|s|$ increases, the measurement distribution $p(x \mid s,\sigma)$ shifts, and the mean measurement magnitude on incorrect trials decreases (see Figure 5a). One might expect this intuition to also result in divergence signature 2, since the effect of decreased measurement noise σ on $p(x \mid s,\sigma)$ also results in a decreased measurement magnitude on incorrect trials (see Figure 5d). However, σ additionally affects $\mathrm{conf}(x,\sigma)$, the per trial deterministic mapping from measurement and noise level to Bayesian confidence (see Figure 5e), whereas stimulus magnitude does not (see Figure 5b). Therefore, when σ is variable, the resulting effect on the measurement distribution is insufficient for describing the pattern of mean confidence on incorrect trials, requiring simulation. We simulated experiments as described in section 4.1 and demonstrate why stimulus magnitude and measurement noise have different effects on mean confidence on incorrect trials (see Figure 5).

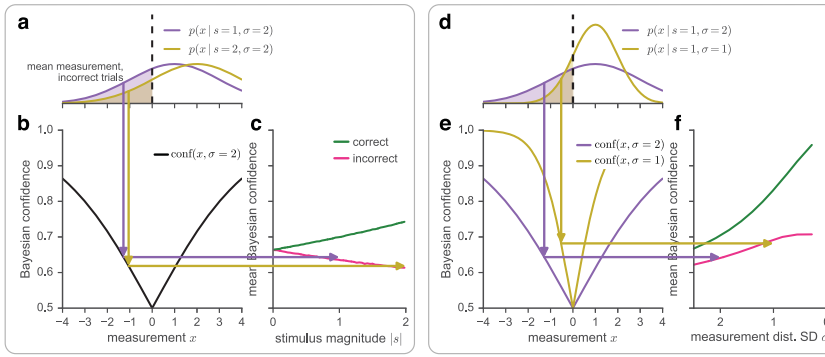

Figure 5: Explanation for why divergence signature 1 is sometimes expected but why divergence as a function of measurement noise might never be expected. Although both increased stimulus magnitude and decreased measurement noise cause the mean measurement magnitude to decrease on incorrect trials, they have different effects on mean confidence. At each value of σ, 2 million stimuli were simulated, using uniform stimulus distributions defined by $p(s \mid C = 1) = \mathcal{U}(s; 0, 2)$ (the case of Figure 3a). (a) As described previously (Drugowitsch, 2016; Hangya et al., 2016; Kepecs et al., 2008), an increase in stimulus magnitude causes the mean measurement magnitude to decrease on incorrect trials. (b) Measurements are mapped onto confidence values using the deterministic function $\mathrm{conf}(x, \sigma)$, which is equivalent to the posterior probability that the choice is correct (see section 2). (c) This mapping results in divergence signature 1, a decrease in mean confidence on incorrect trials. Arrows do not align precisely with the simulated mean because the confidence of the mean measurement is not exactly equal to the mean confidence. (d) A decrease in measurement noise also causes the mean measurement magnitude to decrease on incorrect trials. (e) Because the mapping from measurement to confidence $\mathrm{conf}(x, \sigma)$ is dependent on σ, measurements from the less noisy distribution have higher confidence. (f) Because the confidence mapping is dependent on σ, divergence as a function of measurement noise is not necessarily expected under Bayesian confidence.

6 Other Signatures

A third signature in Hangya et al. (2016) that we do not discuss here (that confidence equals accuracy) is like the 0.75 signature in that it requires either explicit reports of perceived probability of being correct or the experimenter to choose a mapping between rating and perceived probability of being correct (see section 3.1). For any monotonic relationship between accuracy and confidence, it is likely that there is some mapping that equates the two, in which case the signature would not be a sufficient condition for the BCH.

A fourth signature (that confidence allows a better prediction of accuracy than stimulus magnitude alone) is, like divergence signature 1, also predicted by the measurement model (see section 4.2) and is therefore also not a sufficient condition for the BCH.

7 Discussion

We have demonstrated that even in the relatively restricted class of binary categorization tasks that we consider here (see section 2), some signatures are neither necessary nor sufficient conditions for the BCH. Specifically, the 0.75 signature is expected only when observers have very low measurement noise and believe that the CCSDs are nonoverlapping. Additionally, despite claims that divergence signature 1 is "robust to different stimulus distributions" (Kepecs & Mainen, 2012) it is only expected under nonoverlapping stimulus distributions or overlapping (e.g., gaussian) stimulus distributions with high measurement noise. (However, a researcher using overlapping stimulus distributions may still be able to "rescue" the signature by plotting a slightly modified version, as we describe in section 4.1.1.) Because of their nongenerality, these signatures are therefore not necessary conditions of Bayesian confidence. Furthermore, they may be observed under non-Bayesian models, indicating that they are also not sufficient conditions (Fleming & Daw, 2017; Insabato et al., 2016).

A discrepancy in the literature (Navajas et al., 2017) has emerged through the confusion of divergence signature 1 with a second form, in which stimulus magnitude is replaced with another variable that is related to accuracy.[7] We have shown that while divergence signature 1 holds in some cases, there is no evidence that the second form is ever expected under the BCH, which resolves this discrepancy.

The appearance of confidence signatures may depend on the observer's belief about the CCSDs, $q(s \mid C)$. For instance, we showed that the 0.75 signature is not expected if the observer believes that the CCSDs are overlapping, regardless of the true distribution $p(s \mid C)$. In our simulations of divergence signature 1, we assumed that $q(s \mid C) = p(s \mid C)$, but it may be that there are erroneous beliefs $q(s \mid C)$ that eliminate this signature as well. This may be an important consideration for some experimenters due to the difficulty of communicating the CCSDs to observers, especially nonhuman observers. One might assume that with enough training, observers would learn the CCSDs, but critically, the observer has access only to x and not to s. At high levels of measurement noise, for instance, this could lead to a belief that the categories are overlapping, which would eliminate the 0.75 signature.

[7] Kiani, Corthell, and Shadlen (2014) also note the lack of the divergence signature in their data, but because their stimuli have variable duration, optimality is more complicated to characterize (Drugowitsch, DeAngelis, Klier, Angelaki, & Pouget, 2014), and the explanation we offer here may not apply.

For human observers, experimenters may be able to ameliorate this issue by training observers on the categories at low noise, informing the subject that the CCSD will be the same at higher noise levels. However, even this might not ensure that $q(s \mid C) = p(s \mid C)$. Additionally, we are not aware of a good strategy for nonhuman observers. Because the signatures might not be present in data from an otherwise Bayesian observer with erroneous beliefs about the CCSDs, an experimenter expecting the signatures might incorrectly rule out that the observer is Bayesian.

Some of our critique of the signatures has focused on the implicit assumption that experiments use nonoverlapping stimulus distributions. One could object to our critique by questioning the relevance of overlapping stimulus distributions, given that nonoverlapping stimulus distributions are the norm in the confidence literature (Aitchison et al., 2015; Kepecs & Mainen, 2012; Kepecs et al., 2008; Sanders et al., 2016). But although overlapping categories are only just beginning to be used to study confidence (Adler & Ma, 2018; Denison et al., 2018), such categories have a long history in the perceptual categorization literature (Ashby & Gott, 1988; Green & Swets, 1966; Healy & Kubovy, 1981; Lee & Janke, 1964; Liu, Knill, & Kersten, 1995; Qamar et al., 2013; Sanborn, Griffiths, & Shiffrin, 2010). It has been argued that overlapping gaussian stimulus distributions have several properties that make them more naturalistic than nonoverlapping distributions (Maddox, 2002). The property most relevant here is that with overlapping categories, perfect performance is impossible, even with zero measurement noise. With overlapping categories, as in real life, identical stimuli may belong to multiple categories. Imagine a coffee drinker pouring salt rather than sugar into her drink, a child reaching for his parent's glass of whiskey instead of his glass of apple juice, or a doctor classifying a malignant tumor as benign (Augsburger, Corrêa, Trichopoulos, & Shaikh, 2008). In all three examples, stimuli from opposing categories may be visually identical, even under zero measurement noise. For more naturalistic experiments with overlapping categories, qualitative signatures will be unusable if their derivations assume nonoverlapping categories.

Given our demonstration that proposed qualitative signatures of confidence have limited applicability, what is the way forward? One option available to confidence researchers is to discover more signatures, being careful to find the specific conditions under which they are expected. Confidence experimentalists should then make sure to look for such signatures only when their tasks satisfy the specified conditions (e.g., stimulus distribution type, noise level). However, for researchers interested in testing the BCH, we do not necessarily advocate for this course of action because even when applied to relevant experiments, the presence or absence of qualitative signatures provides an uncertain amount of evidence for or against the BCH. Testing for the presence of qualitative signatures is a weak substitute for accumulating probabilistic evidence, something that careful (Palminteri,

Wyart, & Koechlin, 2017) quantitative model comparison does more objectively. Testing for signatures requires the experimenter to make two subjective judgments. First, the experimenter must determine whether the signature is present, a task potentially made difficult by the fact that real data are noisy. Second, the experimenter must determine how much evidence provides in favor of the BCH and whether further investigation is warranted. By contrast, model comparison provides a principled quantity (namely, a log likelihood) in favor of the BCH over some other model (Adler & Ma, 2018; Aitchison et al., 2015; Denison et al., 2018). Given the caveats associated with qualitative signatures, it may be that as a field, we have no choice but to rely on formal model comparison.

Appendix A: Sufficient Conditions for the MAP Decision Rule to Be $x \overset{?}{>} 0$

We wish to specify conditions under which, for all $x > 0$, the maximum a posteriori (MAP) decision rule is $\hat{C} = 1$, that is, $q(C = 1 \mid x) > q(C = -1 \mid x)$, in which $q(C = 1 \mid x)$ is the posterior probability that the category is $C = 1$, given a measurement x. For clarity, we remove σ from $q(x \mid s, \sigma)$ and $q(C \mid x, \sigma)$, as it is not necessary for the proof.

Condition 1. The observer believes that each category is equally probable:

$$q(C = 1) = q(C = -1).$$

Condition 2. The observer believes that the category-conditioned stimulus distributions are mirrored across $s = 0$:

$$q(s \mid C = 1) = q(-s \mid C = -1).$$

Condition 3. The observer believes that a nonnegative stimulus is at least as probable under category $C = 1$ as under category $C = -1$. For $s \geq 0$,

$$q(s \mid C = 1) \geq q(s \mid C = -1).$$

Condition 4. The observer believes that the measurement distribution is a symmetric, monotonically decreasing function of the stimulus:

$$q(x \mid s) = F(|x - s|),$$

where F is a monotonically decreasing function. Gaussian measurement noise satisfies this assumption.

We will use $\Delta_{\text{posterior}}(x) \equiv q(C = 1 \mid x) - q(C = -1 \mid x)$.
Under the above conditions, for all $x > 0$, $\Delta_{\text{posterior}}(x) \geq 0$.

Proof. By Bayes' rule,

$$\Delta_{\text{posterior}}(x) = \frac{q(x \mid C=1)q(C=1)}{q(x)} - \frac{q(x \mid C=-1)q(C=-1)}{q(x)}.$$

By condition 1,

$$\begin{aligned}\Delta_{\text{posterior}}(x) &\propto q(x \mid C=1) - q(x \mid C=-1) \\ &= \int_{-\infty}^{\infty} q(x \mid s)q(s \mid C=1)\,\mathrm{d}s - \int_{-\infty}^{\infty} q(x \mid s)q(s \mid C=-1)\,\mathrm{d}s.\end{aligned}$$

Using $\Delta_{\text{s}}(s) \equiv q(s \mid C=1) - q(s \mid C=-1)$,

$$\begin{aligned}\Delta_{\text{posterior}}(x) &\propto \int_{-\infty}^{\infty} q(x \mid s)\Delta_{\text{s}}(s)\,\mathrm{d}s \\ &= \int_{-\infty}^{0} q(x \mid s)\Delta_{\text{s}}(s)\,\mathrm{d}s + \int_{0}^{\infty} q(x \mid s)\Delta_{\text{s}}(s)\,\mathrm{d}s.\end{aligned}$$

We perform a change of variables $\tilde{s} = -s$:

$$\Delta_{\text{posterior}}(x) \propto \int_{0}^{\infty} q(x|-\tilde{s})\Delta_{\text{s}}(-\tilde{s})\,\mathrm{d}\tilde{s} + \int_{0}^{\infty} q(x|s)\Delta_{\text{s}}(s)\,\mathrm{d}s.$$

Using condition 2, some rearrangement, and then condition 4:

$$\begin{aligned}\Delta_{\text{posterior}}(x) &\propto -\int_{0}^{\infty} q(x \mid -s)\Delta_{\text{s}}(s)\,\mathrm{d}s + \int_{0}^{\infty} q(x \mid s)\Delta_{\text{s}}(s)\,\mathrm{d}s \\ &= \int_{0}^{\infty} \left[q(x \mid s) - q(x \mid -s)\right]\Delta_{\text{s}}(s)\,\mathrm{d}s \\ &= \int_{0}^{\infty} \left[F(|x-s|) - F(|x+s|)\right]\Delta_{\text{s}}(s)\,\mathrm{d}s. \qquad \text{(A.1)}\end{aligned}$$

Our integral spans only the nonnegative domain, $s \geq 0$. Additionally, we only consider $x > 0$. For $s \geq 0$, $|x+s| \geq |x-s|$ and thus, from condition 4, $F(|x-s|) - F(|x+s|) \geq 0$. It also follows from condition 3 that $\Delta_{\text{s}} \geq 0$. Because both factors in equation A.1 are nonnegative, $\Delta_{\text{posterior}}(x) \geq 0$ for all $x > 0$. When $\Delta_{\text{posterior}}(x) > 0$, the category with the higher posterior probability is $C = 1$; when $\Delta_{\text{posterior}}(x) = 0$, both categories have equal posterior probability. □

Appendix B: Derivation of Bayesian Confidence

As described in section 2, if an observer's confidence behavior is Bayesian, it is a function of the posterior probability of the most probable category. By Bayes' rule,

$$\begin{aligned}\mathrm{conf}(x,\sigma) &= \max_C q(C \mid x,\sigma) \\ &= \max_C \frac{q(x \mid C,\sigma)q(C)}{\sum_C q(x \mid C,\sigma)q(C)} \\ &= \max_C \frac{q(x \mid C,\sigma)}{\sum_C q(x \mid C,\sigma)}. \end{aligned} \tag{B.1}$$

In the last step, we eliminated the prior because each category is equally likely and we assume that the observer knows this (i.e., $q(C=1) = q(C=-1)$). We now derive the task-specific likelihood functions $q(x \mid C,\sigma)$ used in our simulations. The observer does not know the true stimulus value s, but does know that the measurement is drawn from a gaussian distribution with a mean of s and SD σ. Using this knowledge, the Bayesian observer marginalizes over s by convolving the stimulus distributions with their noise distribution:

$$\begin{aligned} q(x \mid C,\sigma) &= \int q(x \mid s,\sigma)q(s \mid C)\,\mathrm{d}s \\ &= \int \mathcal{N}(x;s,\sigma)q(s \mid C)\,\mathrm{d}s. \end{aligned} \tag{B.2}$$

For uniform category distributions, we plug $q(s \mid C) = \mathcal{U}(s;a,b)$ into equation B.2 and simplify:

$$\begin{aligned} q_{\mathrm{U}}(x \mid C,\sigma) &= \int \mathcal{N}(x;s,\sigma)\mathcal{U}(s;a,b)\,\mathrm{d}s \\ &= \frac{1}{b-a}\int_a^b \mathcal{N}(x;s,\sigma)\,\mathrm{d}s \\ &= \frac{1}{b-a}\left(\Phi\left(\frac{b-x}{\sigma}\right) - \Phi\left(\frac{a-x}{\sigma}\right)\right), \end{aligned} \tag{B.3}$$

where Φ is the cumulative distribution function of the standard normal distribution. For gaussian category distributions, we plug $q(s \mid C) = \mathcal{N}(s;\mu_C,\sigma_C)$ into equation B.2 and simplify,

$$q_G(x \mid C, \sigma) = \int \mathcal{N}(x; s, \sigma)\mathcal{N}(s; \mu_C, \sigma_C)\, \mathrm{d}s$$
$$= \mathcal{N}\left(x; \mu_C, \sqrt{\sigma^2 + \sigma_C^2}\right), \tag{B.4}$$

using $\sigma_C = 0$ if stimuli from a given category always take on the same value μ_C.

Finally, we plug the task-appropriate likelihood function (equation B.3 or B.4) into equation B.1.

Appendix C: Simpler Proof of Hangya et al. (2016) Lemma

The last step of the proof of the 0.75 signature (see section 3.1.1) uses a lemma proved by Hangya et al. (2016):

Lemma. *Integrating the product of the probability density function $f(t)$ and the distribution function $F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x$ of any probability distribution symmetric to zero over the positive half-line results in 3/8:*

$$\int_0^\infty f(t)F(t)dt = \frac{3}{8}.$$

There is a simpler proof of the lemma than the one by Hangya et al. (2016):

Proof. Using integration by parts and that $f(t) = F'(t)$ by definition,

$$\int_0^\infty f(t)F(t)\,\mathrm{d}t = F(\infty)F(\infty) - F(0)F(0) - \int_0^\infty f(t)F(t)\,\mathrm{d}t$$
$$2\int_0^\infty f(t)F(t)\,\mathrm{d}t = F(\infty)F(\infty) - F(0)F(0).$$

Because F is a cumulative distribution function of a probability distribution symmetric across zero, $F(\infty) = 1$ and $F(0) = \frac{1}{2}$:

$$2\int_0^\infty f(t)F(t)\,\mathrm{d}t = 1 - \frac{1}{4}$$
$$\int_0^\infty f(t)F(t)\,\mathrm{d}t = \frac{3}{8}.$$

□

Appendix D: False Beliefs about Measurement Noise

So far, this letter, as in Hangya et al. (2016), has assumed that observers have accurate knowledge of their own measurement noise. Because it may be of

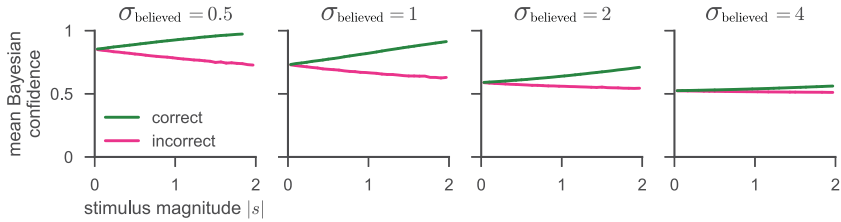

Figure 6: As in Figure 3a. True measurement noise is $\sigma = 1$. The divergence signature is present at all levels of σ_{believed}.

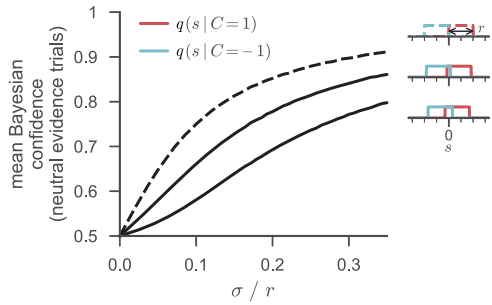

Figure 7: As in Figure 2a, except that simulations were conducted under $\frac{\sigma_{\text{believed}}}{r} = 0.1$. The 0.75 signature is not present, even for nonoverlapping distributions (dashed line) at zero measurement noise.

interest to readers to know what happens when this assumption is violated, we ran our simulations under the condition that the observer has incorrect beliefs about her measurement noise. Specifically, we ran simulations using $p(x \mid s, \sigma) = \mathcal{N}(x; s, \sigma)$ and $q(x \mid s, \sigma_{\text{believed}}) = \mathcal{N}(x; s, \sigma_{\text{believed}})$, where σ may or may not be equal to σ_{believed}.

D.1 Divergence Signature 1. First, we find that under nonoverlapping categories, divergence signature 1 (see section 4) holds regardless of the observer's knowledge of their measurement noise (see Figure 6).

D.2 0.75 Signature. We find that in addition to the conditions described in section 3, the 0.75 signature holds only when the observer has accurate knowledge of her own measurement noise (see Figure 7).

D.3 Divergence Signature 2. We observe that for no value of σ_{believed} that we tested does divergence signature 2 (see section 5) appear (see Figure 8). Our conclusion that divergence signature 2 is not expected under

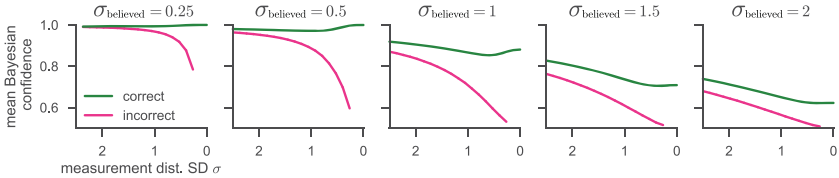

Figure 8: As in Figure 4c. True measurement noise is $\sigma = 1$. Divergence signature 2 is not present or is present only weakly.

the BCH is therefore robust to the observer having incorrect beliefs about her measurement noise.

To understand why, as σ decreases, mean confidence for correct and incorrect trials slope decreases rather than increases, as in Figure 4c, consider the didactic presented in Figure 5. In each individual case shown in Figure 8, we vary σ in $p(x \mid s, \sigma)$ but fix σ_{believed} to a single value in $q(x \mid s, \sigma_{\text{believed}})$. This means that the generating distributions $p(x \mid s, \sigma)$ will vary with σ as depicted in Figure 5d, but that in Figure 5e, there will be only one confidence mapping function $\text{conf}(x, \sigma_{\text{believed}})$ for all σ. This will change the sign of the slope of mean confidence.

Appendix E: Terminology and Notation

Because some of our terminology and notation relate to that used in Hangya et al. (2016), we provide Table 1 to enable easier comparison between the two papers. In some cases, the variables are not exactly identical: the terms in Hangya et al. may be more general. This does not affect the validity of our claims. For consistency, we always describe their work using our terminology and notation.

Table 1: Comparison of Terminology and Notation.

This Letter	Hangya et al. (2016)
True category C	Not used
Stimulus s	Evidence d
Stimulus magnitude $\lvert s \rvert$	Discriminability Δ
Measurement x	Percept $\hat{d}$
Measurement noise σ	Not used
Choice $\hat{C}$	Choice ϑ
Confidence $q(C = \hat{C} \mid x, \sigma) = \text{conf}(x, \sigma)$	Confidence $c = \xi(\hat{d}, \vartheta)$

Acknowledgments

We thank Luigi Acerbi, Rachel N. Denison, Andra Mihali, and Joaquín Navajas for helpful conversations and comments on the manuscript; Bas van Opheusden for the simple proof of the lemma; and Rachel Adler for some clever real-life examples of overlapping categories. This material is based on work supported by the National Science Foundation Graduate Research Fellowship under grant DGE-1342536.

References

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, *10*(6), e1003661.

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Computational Biology*, doi:10.1371/journal .pcbi.1006572.

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, *11*(10), e1004519.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Augsburger, J. J., Corrêa, Z. M., Trichopoulos, N., & Shaikh, A. (2008). Size overlap between benign melanocytic choroidal nevi and choroidal malignant melanomas. *Investigative Ophthalmology and Visual Science*, *49*(7), 2823–2826.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, *115*(43), 11090–11095.

Drugowitsch, J. (2016). Becoming confident in the statistical nature of human confidence judgments. *Neuron*, *90*(3), 425–427.

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., & Pouget, A. (2014). Optimal multisensory decision-making in a reaction-time task. *eLife*, *3*, e03005.

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between belief and performance in perceptual decision making. *PLoS ONE*, *9*(5), e96511.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, & estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 344–354.

Insabato, A., Pannunzi, M., & Deco, G. (2016). Neural correlates of metacognition: A critical perspective on current tasks. *Neuroscience and Biobehavioral Reviews*, *71*, 167–175.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–755.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., & Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron*, *84*(1), 1–12.

Lee, W., & Janke, M. (1964). Categorizing externally distributed stimulus samples for three continua. *Journal of Experimental Psychology*, *68*(1), 376–382.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, *35*(4), 549–568.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*(1), 205–220.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, *78*(3), 567–595.

Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*(325), 1455.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, *11*(11), 1–12.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*(6237), 52–54.

Norton, E. H., Fleming, S. M., Daw, N. D., & Landy, M. S. (2017). Suboptimal criterion learning in static and dynamic environments. *PLoS Computational Biology*, *13*(1), e1005304.

Orhan, A. E., & Jacobs, R. A. (2014). *Are performance limitations in visual short-term memory tasks due to capacity limitations of model mismatch?* arXiv:1407.0644.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433.

Peirce, C. S., & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*, 73–83.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.

Qamar, A. T., Cotton, R. J., George, R. G., Beck, J. M., Prezhdo, E., Laudano, A., . . . Ma, W. J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences*, *110*(50), 20332–20337.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*(2), 63–106.

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, *90*(3), 499–506.

Received February 11, 2018; accepted August 18, 2018.