Bayesian Analysis (2014) **9**, Number 4, pp. 909–922

The Evidentiary Credible Region

David Shalloway *

Abstract. Many disparate definitions of Bayesian credible intervals and regions are in use, which can lead to ambiguous presentation of results. It is particularly unsatisfactory when intervals are specified that do not match the one-sided character of the evidence. We suggest that a sensible resolution is to use the parameterization-independent region that maximizes the information gain between the initial prior and posterior distributions, as assessed by their Kullback-Leibler divergence, subject to the constraint on included posterior probability. This turns out to be equivalent to the relative surprise region previously defined by Evans (1997), and thus provides information theoretic support for its use. We also show that this region is the constrained optimizer over the posterior measure of any strictly monotonic function of the likelihood, which explains its many optimal properties, and that it is guaranteed to be consistent with the sidedness of the evidence. Because all of its equivalent derivations depend on the evidence as well as on the posterior distribution, we suggest that it be called the evidentiary credible region.

Keywords: credible region, credible interval, highest posterior density, parameterization invariance, Kullback-Leibler, information gain, relative surprise region

1 Introduction

A Bayesian γ-*credible region (CR)* $\mathbb{C}_{\boldsymbol{t},\gamma}$ with *credibility* $\gamma = 1-\alpha$ is a subregion of the probability space parameterized by vector $\boldsymbol{t} \in \boldsymbol{T}$, where

$$\int_{\mathbb{C}_{\boldsymbol{t},\gamma}} p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t} = \gamma\ ; \tag{1}$$

$\boldsymbol{t}$ may either be a full or marginal parameter vector (see Appendix) and, correspondingly, $p(\boldsymbol{t}|\mathcal{E})$ is the full or marginal posterior distribution given the evidence $\mathcal{E}$. When t is a scalar, the credible region is called a *credible interval (CI)*. [To avoid unimportant complications we assume that $\boldsymbol{T}$ is the intersection of the supports of $p(\boldsymbol{t}|\mathcal{E})$ and of the prior distribution $p(\boldsymbol{t})$ and that the distributions have no singularities.]

Equation (1) does not fix the placement and shape of the CR, which may vary between different definitions. A popular definition (Casella and Berger 1990) is the *highest posterior density* (HPD) CR $\mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma}$, which satisfies (1) using the region where $p(\boldsymbol{t}|\mathcal{E})$ is highest:

$$\mathbb{C}^{\text{HPD}}_{p(\boldsymbol{t},\gamma)} = \{\boldsymbol{t}\ :\ \int_{p(\boldsymbol{t}|\mathcal{E})\geq c} p(\boldsymbol{t}'|\mathcal{E})\, d\boldsymbol{t}' = \gamma\}\ . \tag{2}$$

Therefore,

$$p(\bar{\boldsymbol{t}}|\mathcal{E}) < c \leq p(\boldsymbol{t}|\mathcal{E}) \qquad\qquad (\boldsymbol{t} \in \mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma}\ ;\ \bar{\boldsymbol{t}} \notin \mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma})\ , \tag{3}$$

*Dept. of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853 dis2@cornell.edu

 DOI:10.1214/14-BA883

where c is the largest value that permits (1) to be satisfied. The boundary of $\mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma}$ is a level set $p(\boldsymbol{t}|\mathcal{E}) = c$ unless $\mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma}$ is one-sided. In that case, $\mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma}$ extends to the boundary of $\boldsymbol{t}$ in at least one direction.

However, $\mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma}$ is not *parameterization invariant*. A CR is invariant under the invertible mapping $\boldsymbol{t} \leftrightarrow \boldsymbol{u}$ if $\mathbb{C}_{\boldsymbol{t},\gamma}$ and $\mathbb{C}_{\boldsymbol{u},\gamma}$, the CR's computed in the two parameterizations, are equivalent:

$$\boldsymbol{u}(\mathbb{C}_{\boldsymbol{t},\gamma}) = \mathbb{C}_{\boldsymbol{u},\gamma} \qquad \text{and} \qquad \boldsymbol{t}(\mathbb{C}_{\boldsymbol{u},\gamma}) = \mathbb{C}_{\boldsymbol{t},\gamma}\ , \tag{4}$$

where $\boldsymbol{u}(\boldsymbol{t})$ and $\boldsymbol{t}(\boldsymbol{u})$ are the mappings between the parameter spaces. In contrast, re-expressing the HPD criterion (3) in terms of $\boldsymbol{u}$ may change the specified region, so (4) may not be satisfied. This introduces a subjective element—the choice of parameterization—that renders the HPD CR ambiguous and reduces its utility for summarizing experiments.

Different solutions to this problem have been suggested: The symmetric CI, which excludes $1-\gamma/2$ posterior probability on either side of the interval is parameterization invariant but may contain regions that are excluded by the evidence, cannot be one-sided (even when the evidence is), and cannot be sensibly generalized to a CR. In the special case when a reference prior (Bernardo 1979; Berger et al. 2009) is being used, Bernardo (2005) has suggested using the *intrinsic CR*, which is defined using a symmetrized form of the Kullback-Leibler divergence between the reference prior and posterior distributions. When a non-informative "standard" prior is being used, Box and Tiao (1992) suggest distinguishing the HPD CR computed in the parameterization where the prior is uniform as the *standardized* HPD CR. However, they provide no mathematical justification for this choice. A justification that can be extended to arbitrary priors has been provided in a series of papers by Evans and coworkers (Evans 1997; Evans et al. 2006; Evans and Shakhatreh 2008; Baskurt and Evans 2013) using the concept of *relative surprise*. They follow a motivational trail in which the *surprise* of an observation is defined as some measure of the deviation of the observation from prior expectation (Good 1988, 1989). By choosing a specific form for the surprise, they define the parameterization-invariant *relative surprise region*, which they show is equivalent to the HPD CR in the parameterization where the prior is uniform. They also show that it has the smallest prior measure, maximizes the Bayes factor, and maximizes the relative belief ratio among all γ-CR's (Evans et al. 2006; Evans and Shakhatreh 2008; Baskurt and Evans 2013).

We find the generality of the solution of Evans et al. attractive and show here that it is not necessary to invoke a definition of surprise to motivate it: it can be derived directly from information theory as the γ-CR that maximizes the posterior expectation of the Kullback-Leibler divergence. Furthermore, we show that optimizing the posterior expectation of any strictly monotonic function of the likelihood over the CR gives the same result; this simply explains the many optimal properties of this CR. To emphasize that it is a hybrid between CRs that depend only on the posterior distribution and frequentist confidence intervals, which depend only on the evidence as represented by the likelihood, we suggest that it be called the *evidentiary* CR.

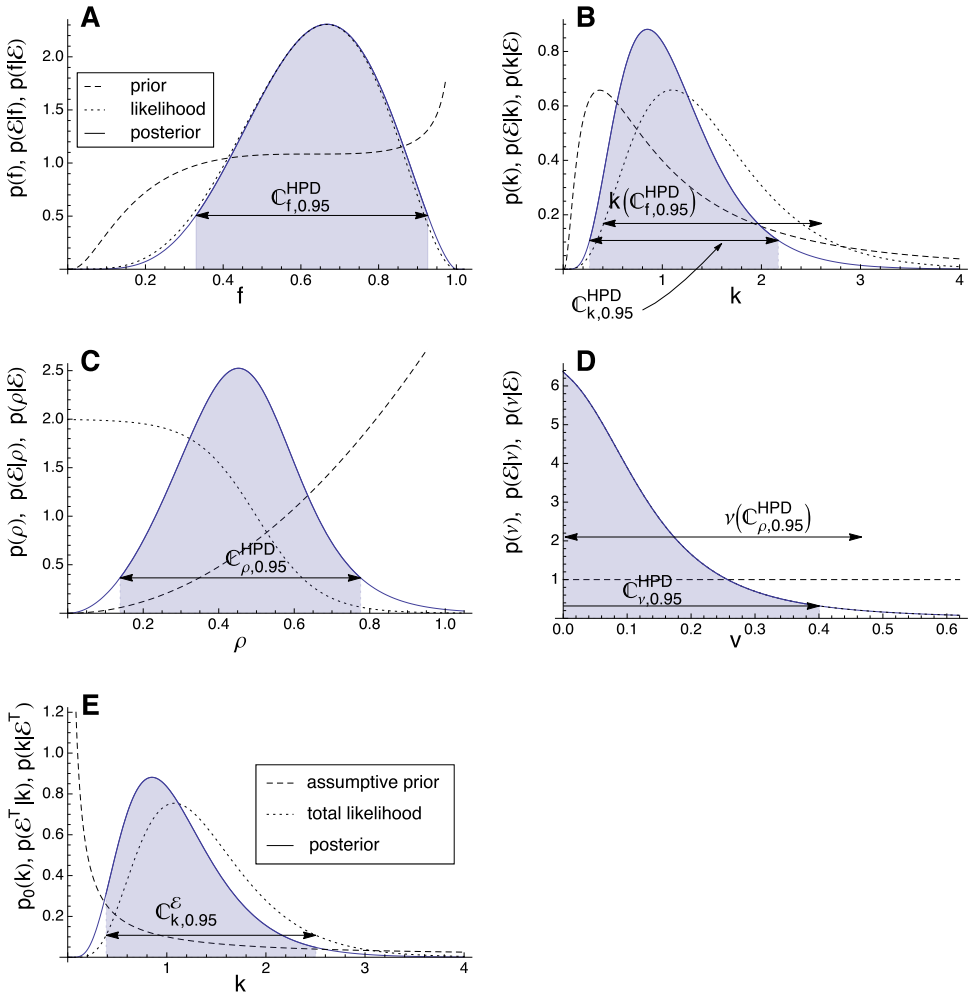

Figure 1: Examples of ambiguities of HPD CIs. **A**: The prior, likelihood, and posterior distributions as functions of f for the experiment measuring the fraction of cells transformed from state A to B. The arrowed line indicates $\mathbb{C}^{\text{HPD}}_{f,0.95}$. The shaded area contains 95% of the posterior density. **B**: The same functions as in **A** but as functions of k. The lower arrowed line is $\mathbb{C}^{\text{HPD}}_{k,0.95}$ and corresponds to the range of the shaded area. The upper arrowed line is $k(\mathbb{C}^{\text{HPD}}_{f,0.95})$, the map of $\mathbb{C}^{\text{HPD}}_{f,0.95}$ into the k-space. **C**: The prior, likelihood and posterior as functions of ρ for the experiment measuring the subvolume of a cellular compartment. The arrowed line indicates $\mathbb{C}^{\text{HPD}}_{\rho,0.95}$, which is double-sided. **D**: The same functions as in **C** as functions of ν. The likelihood and posterior overlap completely. The lower arrowed line is $\mathbb{C}^{\text{HPD}}_{\nu,0.95}$ and corresponds to the range of the shaded area; it is one-sided since it has its left boundary at $\nu = 0$, the boundary of the parameter space. The upper arrowed line is $\nu(\mathbb{C}^{\text{HPD}}_{\rho,0.95})$, the map of the $\mathbb{C}^{\text{HPD}}_{\rho,0.95}$ into the ν-space. Like $\mathbb{C}^{\text{HPD}}_{\rho,0.95}$ it is double-sided; its left boundary is at $\nu = 0.004$. **E**: The evidentiary CI, $\mathbb{C}^{\mathcal{E}}_{0.95}$, for the cell transformation example. The boundaries are a level set of the total likelihood function $p(\mathcal{E}|k)$ (which is unnormalized in this case) and the shaded area within the CI contains 95% of the posterior density. (The posterior distribution is the same as that in panel B.)

2 Results

We first present two examples showing that a subjective choice of "natural" parameterization can be ambiguous, particularly when the one suggested by the probabilistic model differs from that provided by the data. The second example illustrates the inconsistency between the evidence and an HPD CI that can arise when the data is one-sided.

2.1 Examples of HPD CI ambiguities

Parameterization dependence

Consider a stochastic process that transforms normal cells (state A) to cancer cells (state B) with rate k so that

$$\frac{dN_A(t)}{dt} = -kN_A(t)\ ; \qquad \frac{dN_B(t)}{dt} = kN_A(t)\ ,$$

where $N_A(t)$ and $N_B(t)$ are the numbers of cells at time t. If $N_A(0) = N_0$ and $N_B(0) = 0$, the solution for an experiment conducted over time interval τ is

$$N_A(\tau) = e^{-k\tau}N_0\ ; \qquad N_B(\tau) = (1 - e^{-k\tau})N_0\ .$$

The experimentally measurable parameter is f, the fraction of cells transformed to cancer cells. This is related to k by the invertible relations

$$\frac{N_B(\tau)}{N_0} \equiv f(k) = (1 - e^{-k\tau})\ ; \qquad k(f) = -(1/\tau)\log(1 - f)\ .$$

Our goal is to estimate k from multiple experimental measurements of f.

We assume that measurements of f have yielded the likelihood $p(\mathcal{E}|f)$ shown in panel A of Figure 1 and that, based on the original uninformative prior distribution and earlier experiments, the prior distribution at the beginning of these experiments, $p(k)$, is lognormal (panel B). Bayes Theorem gives the posterior distributions $p(f|\mathcal{E})$ and $p(k|\mathcal{E})$, and (1) and (3) give the corresponding HPD CIs, $\mathbb{C}^{\text{HPD}}_{f,0.95}$ and $\mathbb{C}^{\text{HPD}}_{k,0.95}$ (panels A and B, respectively). Note that $k(\mathbb{C}^{\text{HPD}}_{f,0.95})$, the map of the CI computed in f into the k-space, is different from the CI computed in k. It is troubling that the experimentalist must make a subjective decision in presenting his results. Should he or she state that the credible interval is $0.27 \leq k \leq 2.2$, the interval computed in k-space, or $0.40 \leq k \leq 2.6$, the interval computed using the experimental variable f?

One- and two-sided HPD CIs

An HPD CI that is one-sided when computed in one parameter may be two-sided when computed in another. Thus, even this qualitative feature can be ambiguous. For example, consider the hypothetical experiment illustrated in Figure 1, C and D: A roughly spherical type of cell has a roughly spherical subcompartment that occupies a fraction ν of its volume. The uninformative prior distribution is $p(\nu) = 1$ $(0 < \nu < 1)$

(panel D) and the experimental goal is to determine ν by microscopic measurements of the ratio, ρ, between the subcompartment and cell radii. However, because of limited resolution, it can only be determined that $\rho \lesssim 0.5$, as quantitatively described by the experimental device-dependent likelihood function (panel C). If we compute the HPD CI using ρ, the experimental parameter, we conclude that $\mathbb{C}^{\text{HPD}}_{\rho,0.95} = \{\rho : 0.14 \le \rho \le 0.78\}$ (panel C), which maps to $\{\nu : 0.003 < \nu < 0.47\}$ (panel D): the HPD CI is two-sided. However, if we compute the HPD CI using ν, the parameter of scientific interest, we conclude that $\mathbb{C}^{\text{HPD}}_{\nu,0.95} = \{\nu : 0 < \nu < 0.40\}$ (panel D): the HPD CI is one-sided. Intuitively, the latter choice makes more sense since the evidence is one-sided. How is this to be objectively justified?

2.2 The relative surprise region

To get a parameterization invariant CR, Evans (1997) introduced the idea of *least relative surprise* to order parameters: $\boldsymbol{t}'$ is preferred to $\boldsymbol{t}$ as a posterior estimate if

$$\frac{p(\boldsymbol{t}'|\mathcal{E})}{p(\boldsymbol{t}')} > \frac{p(\boldsymbol{t}|\mathcal{E})}{p(\boldsymbol{t})} \,. \tag{5}$$

That is, the preferred parameter is that for which the evidence causes the greatest increase in relative belief. Since the ratios are proportional to the likelihoods, $p(\mathcal{E}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\mathcal{E})/p(\boldsymbol{t})$, (5) implies that $\boldsymbol{t}$ is more surprising than $\boldsymbol{t}'$ when it has lower likelihood. [For convenience we use the notations $p(\boldsymbol{t})$ and $p(\mathcal{E}|\boldsymbol{t})$ even when the prior and likelihood are not normalizable. The results do not depend on their normalizability.] Evans defines the *observed relative surprise* at $\boldsymbol{t}$ as the posterior probability that $\boldsymbol{t}$ is less preferred than other values $\boldsymbol{t}'$:

$$\Pi(\boldsymbol{t}, \mathcal{E}) = \int_{\boldsymbol{T}} H\left[\frac{p(\boldsymbol{t}'|\mathcal{E})}{p(\boldsymbol{t}')} - \frac{p(\boldsymbol{t}|\mathcal{E})}{p(\boldsymbol{t})}\right] p(\boldsymbol{t}'|\mathcal{E})\, d\boldsymbol{t}' \,,$$

where $H(\cdot)$ is the Heaviside step function, $H(x) = 0$ $(x < 0)$; $H(x) = 1$ $(x \ge 0)$. A γ-relative surprise region $\mathbb{C}^{\text{RS}}_\gamma$ is then defined as the the set of $\boldsymbol{t}$ having observed relative surprise no greater than γ:

$$\mathbb{C}^{\text{RS}}_\gamma = \{\boldsymbol{t} \in \boldsymbol{T} \ : \Pi(\boldsymbol{t}, \mathcal{E}) \le \gamma\} \,. \tag{6}$$

The parameterization invariance of $\mathbb{C}^{\text{RS}}_\gamma$ follows from the invariance of $\Pi(\boldsymbol{t}|\mathcal{E})$.

To show that $\mathbb{C}^{\text{RS}}_\gamma$ contains posterior probability γ, we re-express (6) as

$$\mathbb{C}^{\text{RS}}_\gamma = \{\boldsymbol{t} \in \boldsymbol{T} \ : \Phi_\varepsilon[p(\mathcal{E}|\boldsymbol{t})] \le \gamma\} \,,$$

where

$$\Phi_\varepsilon(p) = \int_{\boldsymbol{T}} H[p(\mathcal{E}|\boldsymbol{t}') - p]\, p(\boldsymbol{t}'|\mathcal{E})\, d\boldsymbol{t}' \,. \tag{7}$$

$\Pi(\boldsymbol{t}, \mathcal{E})$ is a non-decreasing function of $p(\boldsymbol{t}|\mathcal{E})/p(\boldsymbol{t})$ and, postponing until Section 2.5 discussion of cases where $p(\mathcal{E}|\boldsymbol{t})$ has a finite-measure level set, it is also continuous. Therefore, there will be a solution c to

$$\Phi_\varepsilon(c) = \gamma \,. \tag{8}$$

[In the extremely unlikely case that $p(\mathcal{E}|\boldsymbol{t})$ is discontinuous on the entire boundary of $\mathbb{C}_\lambda^{\text{RS}}$, there will be a range of solutions. This does not significantly affect the argument; any solution can be used since they all yield the same $\mathbb{C}_\lambda^{\text{RS}}$.] Therefore, $\mathbb{C}_\gamma^{\text{RS}}$ consists of all points with

$$p(\mathcal{E}|\bar{\boldsymbol{t}}) < c \leq p(\mathcal{E}|\boldsymbol{t}) \qquad (\boldsymbol{t} \in \mathbb{C}_\gamma^{\text{RS}}\,;\ \bar{\boldsymbol{t}} \notin \mathbb{C}_\gamma^{\text{RS}})\,, \tag{9}$$

and it follows that

$$\int_{\mathbb{C}_{\boldsymbol{t},\gamma}^{\text{RS}}} p(\boldsymbol{t}|\mathcal{E})\,d\boldsymbol{t} = \int_{\boldsymbol{T}} H\left[p(\mathcal{E}|\boldsymbol{t}) - c\right]\, p(\boldsymbol{t}|\mathcal{E})\,d\boldsymbol{t} = \Phi_\varepsilon(c) = \gamma\,.$$

[An alternative approach is presented by Baskurt and Evans (2013); there the relative surprise region is defined by an equivalent of (9) with c determined by (7) and (8).]

Comparing (9) with (3) shows that the relative surprise region is the same as the HPD CR computed in the parameterization where prior is constant; i.e., it is equal to the standardized HPD CR when a uniform prior has been used.

2.3 The Kullback-Leibler CR

Information theory provides a different starting point for deriving a parameterization-independent CR. Although the resulting *Kullback-Leibler CR*, $\mathbb{C}_\gamma^{KL}$, will turn out to be the same as the relative surprise region, we derive it independently to demonstrate its straightforward motivation.

The example presented in Section 2.1 provides an important clue: Intuitively we expect that one-sided evidence should lead to a one-sided CR, which implies that our analysis should focus on the information gain provided by the evidence to determine the placement and shape of the HPD CR. The most commonly used measure of information gain is the relative entropy or Kullback-Leibler divergence; therefore, we use it to quantify the information gain provided by the evidence. Over subregion $\mathbb{C}$ this is

$$\int_{\mathbb{C}} p(\boldsymbol{t}|\mathcal{E}) \log \frac{p(\boldsymbol{t}|\mathcal{E})}{p(\boldsymbol{t})}\,d\boldsymbol{t}\,. \tag{10}$$

Therefore, the subregion containing posterior probability γ that maximizes the Kullback-Leibler divergence is

$$\mathbb{C}_\gamma^{KL} = \underset{\mathbb{C}}{\text{argmax}} \int_{\mathbb{C}} p(\boldsymbol{t}|\mathcal{E})\, \log \frac{p(\boldsymbol{t}|\mathcal{E})}{p(\boldsymbol{t})}\,d\boldsymbol{t} \tag{11}$$

$$\text{with} \qquad \gamma = \int_{\mathbb{C}_\gamma^{KL}} p(\boldsymbol{t}|\mathcal{E})\,d\boldsymbol{t}\,. \tag{12}$$

Since both (11) and (12) are parameterization invariant, $\mathbb{C}_\gamma^{KL}$ is also invariant.

Consider the change in (11) if $\mathbb{C}_\gamma^{KL}$ is modified by simultaneously adding and subtracting small regions with hypervolumes surrounding $\bar{\boldsymbol{t}} \notin \mathbb{C}_\gamma^{KL}$ and $\boldsymbol{t} \in \mathbb{C}_\gamma^{KL}$, where

neither point is at a discontinuity of the prior or posterior distributions so that they are well-defined. (Formally, this is an interior variation of a domain functional.) To maintain the probability constraint to first order in an arbitrarily small constant Δ, the variational region hypervolumes are set to $\Delta\, p(\boldsymbol{t}|\mathcal{E})$ and $\Delta\, p(\bar{\boldsymbol{t}}|\mathcal{E})$, respectively. The change induced in the Kullback-Leibler divergence is

$$\delta \int_{\mathbb{C}_\gamma^{KL}} p(\boldsymbol{t}|\mathcal{E})\, \log \frac{p(\boldsymbol{t}|\mathcal{E})}{p(\boldsymbol{t})}\, d\boldsymbol{t} = \Delta\, p(\bar{\boldsymbol{t}}|\mathcal{E})\, p(\boldsymbol{t}|\mathcal{E}) \left[\log \frac{p(\bar{\boldsymbol{t}}|\mathcal{E})}{p(\bar{\boldsymbol{t}})} - \log \frac{p(\boldsymbol{t}|\mathcal{E})}{p(\boldsymbol{t})}\right] + O(\Delta^2)\ ,$$

while the change in the included probabilty is $O(\Delta^2)$. The change in the divergence must be non-positive in $O(\Delta)$ if $\mathbb{C}_\gamma^{KL}$ is to be a maximizer, which implies that $p(\mathcal{E}|\bar{\boldsymbol{t}}) \le p(\mathcal{E}|\boldsymbol{t})$. The equality can hold for a region of finite measure only if $p(\mathcal{E}|\boldsymbol{t})$ has a finite-measure level set; a special case that we again defer to Section 2.5. We therefore must have

$$p(\mathcal{E}|\bar{\boldsymbol{t}}) < c \le p(\mathcal{E}|\boldsymbol{t}) \qquad\qquad (\boldsymbol{t} \in \mathbb{C}_\gamma^{\mathcal{E}}\,;\ \bar{\boldsymbol{t}} \notin \mathbb{C}_\gamma^{\mathcal{E}})\ , \tag{13}$$

where c is the smallest value of $p(\mathcal{E}|\boldsymbol{t})$ in $\mathbb{C}_\gamma^{KL}$. This is the same as (9); we conclude that the Kullback-Leibler CR and the relative surprise region are identical.

2.4 Other optimization properties; the common role of the evidence

From the proof above, it is evident that $\mathbb{C}_\gamma^{KL}$ is also the probability-constrained maximizer of the posterior expectation of any strictly increasing function $f(\cdot)$ of $p(\mathcal{E}|\boldsymbol{t})$. For example, if $f(x) = x$,

$$\mathbb{C}_\gamma^{KL} = \underset{\mathbb{C}}{\text{argmax}} \int_{\mathbb{C}} p(\mathcal{E}|\boldsymbol{t})\, p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t} \tag{14}$$

along with the posterior probability constraint (12). That is, $\mathbb{C}_\gamma^{KL}$ is the constrained maximizer of the posterior expectation value of the likelihood. Since the Bayes factor between γ-CRs $\mathbb{C}_{a,\gamma}$ and $\mathbb{C}_{b,\gamma}$ is

$$\frac{\int_{\mathbb{C}_{a,\gamma}} p(\mathcal{E}|\boldsymbol{t})\, p(\boldsymbol{t}|\mathcal{E}) d\boldsymbol{t}}{\int_{\mathbb{C}_{b,\gamma}} p(\mathcal{E}|\boldsymbol{t})\, p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}}\ ,$$

(14) implies that $\mathbb{C}_\gamma^{KL}$ maximizes the Bayes factor amongst all γ-CRs. This property has already been proven for the relative surprise region by a different means (Evans et al. 2006; Evans and Shakhatreh 2008; Baskurt and Evans 2013).

Conversely, $\mathbb{C}_\gamma^{KL}$ is the probability-constrained minimizer of the posterior expectation of any strictly decreasing function of $p(\mathcal{E}|\boldsymbol{t})$. For example,

$$\mathbb{C}_\gamma^{KL} = \underset{\mathbb{C}}{\text{argmin}} \int_{\mathbb{C}} \frac{1}{p(\mathcal{E}|\boldsymbol{t})}\, p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t} = \underset{\mathbb{C}}{\text{argmin}} \int_{\mathbb{C}} p(\boldsymbol{t})\, d\boldsymbol{t}\ . \tag{15}$$

This shows that $\mathbb{C}_\gamma^{KL}$ is the constrained minimizer, subject to (12), of the prior measure, a property that has also been proven for the relative surprise region (Evans et al. 2006; Evans and Shakhatreh 2008).

These alternative definitions do not have the an information theoretic interpretation as does (11), but emphasize that this CR can be defined in an infinite number of ways. Since the common factor is that all involve a focus on the evidence via the likelihood function, we suggest that the standardized HPD CR, the relative surprise region, and the information-gain CR all be called the evidentiary CR, $\mathbb{C}^{\varepsilon}_{\gamma}$.

Collecting results, we have proven

Theorem. $\mathbb{C}^{\varepsilon}_{\gamma} = \underset{\mathbb{C}}{\text{argmax}} \int_{\mathbb{C}} f[p(\mathcal{E}|\boldsymbol{t})]\, p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$ *with* $\gamma = \int_{\mathbb{C}^{\varepsilon}_{\gamma}} p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$, *where* $f(\cdot)$ *is any strictly increasing function, is a parameterization-invariant* γ*-CR that is equivalent to the relative surprise region.*

Corollary 9. *An equivalent definition is* $\mathbb{C}^{\varepsilon}_{\gamma} = \underset{\mathbb{C}}{\text{argmin}} \int_{\mathbb{C}} g[p(\mathcal{E}|\boldsymbol{t})]\, p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$ *with* $\gamma = \int_{\mathbb{C}^{\varepsilon}_{\gamma}} p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$, *where* $g(\cdot)$ *is any strictly decreasing function.*

Some specific instances are:

Corollary 10. $\mathbb{C}^{\varepsilon}_{\gamma}$ *is the probability-constrained maximizer of the Kullback-Leibler divergence:* $\mathbb{C}^{\varepsilon}_{\gamma} = \underset{\mathbb{C}}{\text{argmax}} \int_{\mathbb{C}} p(\boldsymbol{t}|\mathcal{E})\, \log \frac{p(\boldsymbol{t}|\mathcal{E})}{p(\boldsymbol{t})}\, d\boldsymbol{t}$ *with* $\gamma = \int_{\mathbb{C}^{\varepsilon}_{\gamma}} p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$.

Corollary 11. $\mathbb{C}^{\varepsilon}_{\gamma}$ *is the probability-constrained maximizer of the likelihood:* $\mathbb{C}^{\varepsilon}_{\gamma} = \underset{\mathbb{C}}{\text{argmax}} \int_{\mathbb{C}} p(\mathcal{E}|\boldsymbol{t})\, p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$ *with* $\gamma = \int_{\mathbb{C}^{\varepsilon}_{\gamma}} p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$.

Corollary 12. $\mathbb{C}^{\varepsilon}_{\gamma}$ *maximizes the Bayes factor amongst all* γ*-CRs.*

Corollary 13. $\mathbb{C}^{\varepsilon}_{\gamma}$ *is the probability-constrained minimizer of the prior measure:* $\mathbb{C}^{\varepsilon}_{\gamma} = \underset{\mathbb{C}}{\text{argmin}} \int_{\mathbb{C}} p(\boldsymbol{t})\, d\boldsymbol{t}$ *with* $\gamma = \int_{\mathbb{C}^{\varepsilon}_{\gamma}} p(\boldsymbol{t}|\mathcal{E})\, d\boldsymbol{t}$.

Corollary 14. $\mathbb{C}^{\varepsilon}_{\gamma} = \mathbb{C}^{\text{HPD}}_{\boldsymbol{t},\gamma}$ *if* $p(\boldsymbol{t})$ *is a constant.*

2.5 Fine points

Evidentiary level sets and credible regions

The likelihood function in the cell measurement example of Section 2.1 has an approximate level set at $p(\mathcal{E}|\rho) \approx 2$ because of the limited resolution of the measurement device (Figure 1C), and it is possible that other experimental devices could have exact finite-measure level sets of $p(\mathcal{E}|\boldsymbol{t})$. These can potentially prevent the unique definition of $\mathbb{C}^{\varepsilon}_{\gamma}$ for specific values of γ. Referring to (7), we see that a level set of posterior measure ΔP at $p(\mathcal{E}|\boldsymbol{t}) = p_d$ will induce a discontinuity in $\Phi_{\varepsilon}(p_d)$ of magnitude ΔP. If this discontinuity spans γ, then (8) will not have a solution and there will not be a relative surprise region of posterior probability γ. There is a corresponding problem in the Kullback-Leibler approach: In this case (11) will have a continuous range of solutions, each including just enough of the level set to satisfy (12); the solution will not be unique. In this case, the smallest value of γ for which (8) and (11) have a unique solution can

be specified. When analyzing device measurements, special treatment may be necessary even for an approximate level set if its variation is less than the experimental accuracy. Such situations must be handled on a case-by-case basis.

The initial prior and total evidence

A Bayesian analysis begins with an initial *assumptive prior* p_0, which is usually uninformative. This is then updated by one or more experiments, each contributing its own evidence. When there are multiple sequential experiments, the final posterior distribution will be the same whether the analysis is performed sequentially, with each posterior being used as the prior for the next step, or in parallel, with the combined *total evidence*, $\mathcal{E}^T$, used to update p_0. The CR must also be the same for both analyses. However, if the evidentiary CR were to be defined using only the evidence from the final experiment, it would depend on the order in which the experiments were conducted, which is not appropriate. To avoid this, the evidentiary CR must be computed using p_0 and $\mathcal{E}^T$ so that the *total* information gain is maximized. For example, the lognormal prior in the cancer cell example is not the assumptive prior because it is based on evidence from related cell types. In analyzing this experiment we must follow the chain of Bayesian reasoning back to the assumptive prior, which we posit in this case to be a uniform prior in $\log k$. Correspondingly, in this case $\mathcal{E}^T$ is the product of a lognormal likelihood function over k from previous experiments and the likelihood function from the current experiment, $p(k|\mathcal{E})$ shown in Figure 1B. In the cell subcompartment example, the prior distribution, which is an uninformative prior over v, is the assumptive prior and the evidence from this experiment alone is the total evidence.

2.6 Resolution of ambiguities by the evidentiary CR

Computing the evidentiary CR for the examples presented above illustrates how it resolves the ambiguities. The result for the first example is illustrated in Figure 1E. The assumptive prior (dashed line) is the improper distribution $p_0(k) \propto 1/k$, corresponding to a flat prior in $\log k$. The total likelihood function (dotted line) is the product of the (lognormal) likelihood function for the evidence collected before this experiment and the likelihood function from this experiment (the dotted line in panel B). [The lognormal prior shown in panels A and B is the product of $c(k)$ and the earlier likelihood function.] In this case the evidentiary CI, $\mathbb{C}^{\mathcal{E}}_{0.95}$, is intermediate between the HPD CIs computed in k and t:

$$\begin{aligned} \mathbb{C}^{\text{HPD}}_{k,0.95}: &\quad 0.26 \le k \le 2.17 \\ \mathbb{C}^{\mathcal{E}}_{0.95}: &\quad 0.39 \le k \le 2.50 \\ k\left(\mathbb{C}^{\text{HPD}}_{t,0.95}\right): &\quad 0.40 \le k \le 2.61\ . \end{aligned}$$

In the second example, the assumptive prior is uniform in v. Thus, $\mathbb{C}^{\mathcal{E}}_{0.95} = \mathbb{C}^{\text{HPD}}_{v,0.95}$; the evidentiary CI is just the one-sided interval that has already been computed and illustrated in Figure 1D. This removes the ambiguity and gives, in agreement with

the one-sided evidence, a one-sided CI. This agreement between the sidedness of the evidence and $\mathbb{C}_\gamma^\varepsilon$ is guaranteed in general: (13) implies that the CR must extend to the boundary if there is a direction in which $p(\mathcal{E}|\boldsymbol{t})$ is an increasing function everywhere in $\boldsymbol{T}$.

3 Discussion

A variety of definitions of the credible region are available and, in some cases, any one may be adequate as long as it is clearly specified. However, in other cases the CRs differ so much that the need for an objective choice is compelling. This is illustrated by the cell transformation example. The need for an objective choice is particularly strong when the evidence is one-sided because different HPD CRs may be either one-sided or not. This is illustrated by the cellular subcompartment example: the experimental data is one-sided so a one-sided CR is appropriate, but there is no objective basis for this choice in the HPD formulation. As illustrated, the HPD CI is one-sided when parameterized in ν, the variable of scientific interest, but two-sided when parameterized in ρ, the experimental variable. Other common definitions of the CI do not resolve the problem: the symmetric CI *ipso facto* cannot represent a one-sided interval and the mean-centered CI, like the HPD CI, can be one- or two-sided depending on the parameterization.

An attractive resolution has been described in a series of papers by Evans and coworkers (Evans 1997; Evans et al. 2006; Evans and Shakhatreh 2008; Baskurt and Evans 2013), who introduce the concept of relative surprise and use it to derive a parameterization-independent "relative surprise region" that maximizes the "relative belief ratio," maximizes the Bayes factor, and minimizes the prior measure among all γ-CRs. It is equivalent to the HPD CR computed in the parameterization where the prior is constant and, when a non-informative prior is used, is equivalent to the "standardized HPD" of Box and Tiao (1992). In support of the use of this CR, we showed here that it can be directly derived without the introduction of surprise from basic information theory: it is the CR that maximizes the information gain provided by the evidence as quantified by the Kullback-Leibler divergence. We also showed that it is the constrained maximizer of any strictly increasing function of the likelihood (equivalently, the constrained minimizer of any strictly decreasing function of the likelihood), which explains its many properties and potential definitions. The common feature of all of these is a focus on the evidence, so we suggest that it be called the "evidentiary CR." This emphasizes that it stands midway between CRs that depend only on the posterior distribution and frequentist confidence intervals, which depend only on the evidence as represented by the likelihood.

The dependence of the evidentiary CR on the likelihood ensures satisfaction of the natural requirement that it be one-sided when the total evidence is one-sided. In that case the CR boundary will extend to the boundary of the parameter space itself. When the evidentiary CR is in the interior of the parameter space, its boundary will be a level set of the likelihood.

The "maximizing missing information property" used to define a reference prior by Bernardo and Berger (Berger et al. 2009) also uses the Kullback-Leibler information gain criterion. But it is the prior that is varied when determining a reference prior, while it is the CR boundary that is varied when determining the evidentiary CR. Extending the reference analysis approach, Bernardo (2005) has also suggested augmenting the use of a reference prior with a parameterization-independent "intrinsic CR" that is defined using a symmetrized modification of the Kullback-Leibler divergence between the reference prior and posterior distributions. This has some similarities to the evidentiary CR, at least in cases where a reference prior is being used, but the information theoretic interpretation of using a symmetrized divergence is unclear because the information gain is an intrinsically asymmetric concept. We have not investigated whether the symmetrization has a significant effect in practice or whether it might be advantageous to generalize the intrinsic CR to cases where a reference prior is not being used.

We have provided one-dimensional examples for simplicity, but the evidentiary CR can also be defined over multi-dimensional spaces. When t is one-dimensional and the full parameter space is multidimensional, Monte Carlo sampling methods that have been used for computing HPD CIs [e.g., Chen and Shao (1999)] can be adapted (see Appendix). A sampling method of this type has been developed by Evans et al. (2006) in the context of the relative surprise region. When $\boldsymbol{t}$ is many-dimensional, the computation can be challenging. In the bivariate case it may be possible to adapt approaches that have been used for HPD CR's [e.g., Wei and Tanner (1990); Turkkan and Pham-Gia (1997)]. Alternatively, a natural approximation would be to restrict the CR to a convenient N-dimensional shape—e.g., an orthotope or ellipsoid—that is described by a few parameters, and then to solve the information gain maximization condition in terms of these parameters using Monte Carlo sampling over the full parameter space and a standard constrained optimization method.

Acknowledgments

The author wishes to thank Martin T. Wells and Brian S. White for critiques of the manuscript and to an anonymous reviewer for drawing his attention to the papers of Evans et al.

References

Baskurt, Z. and Evans, M. (2013). "Hypothesis assessment and inequalities for Bayes factors and relative belief ratios." *Bayesian Analysis*, 8: 569–590. 910, 914, 915, 918

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The formal definition of reference priors." *Annals of Statistics*, 37: 905–938. 910, 919

Bernardo, J. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society. Series B (Methodological)*, 41: 113–147. 910

— (2005). "Intrinsic credible regions: An objective Bayesian approach to interval estimation." *Test*, 14: 317–384. 910, 919

Box, G. E. and Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. Hoboken, NJ: Wiley-Interscience. 910, 918

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Pacific Grove, CA: Wadsworth & Brooks. 909

Chen, M.-H. and Shao, Q.-M. (1999). "Monte Carlo estimation of Bayesian crediable and HPD intervals." *Journal of Computational and Graphical Statistics*, 8: 69–92. 919

Evans, M. (1997). "Bayesian inference procedures derived via the concept of relative surprise." *Communications in Statistics - Theory and Methods*, 26: 1125–1143. 909, 910, 913, 918

Evans, M., Guttman, I., and Swartz, T. (2006). "Optimality and computations for relative surprise inferences." *Canadian Journal of Statistics*, 34: 113–29. 910, 915, 918, 919

Evans, M. and Shakhatreh, M. (2008). "Optimal properties of some Bayesian inferences." *Electronic Journal of Statistics*, 2: 1268–280. 910, 915, 918

Good, I. (1988). "Surprise index." In Kotz, S., Johnson, N., and Reed, C. (eds.), *Encyclopedia of Statistical Sciences*, volume 7. New York: John Wiley & Sons. 910

— (1989). "Surprise indices and p-values." *Journal of Statistical and Computational Simulation*, 32: 90–92. 910

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. New York: Springer, second edition. 921

Turkkan, N. and Pham-Gia, T. (1997). "Highest posterior density credible region and minmum area confidence region: the bivariate case." *Applied Statistics*, 46: 131–182. 919

Wei, G. and Tanner, M. (1990). "Calculating the content and boundary of the highest posterior density region via data augmentation." *Biometrika*, 77: 649–652. 919

Appendix: Bayes' Theorem and the evidentiary CR for marginal parameters

The components of a marginal parameter vector of interest, $\boldsymbol{t}$, are often a subset of the components of the full parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In this case, $\boldsymbol{\theta} = \boldsymbol{t} \otimes \boldsymbol{\psi}$ is the outer product of $\boldsymbol{t} \in \boldsymbol{T}$ and a vector of nuisance parameters $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, and the prior, likelihood

and posterior distributions over $\boldsymbol{t}$ are

$$
\begin{aligned}
p(\boldsymbol{t}) &= \int_{\boldsymbol{\Psi}} p(\boldsymbol{t},\boldsymbol{\psi})\, d\psi \\
p(\mathcal{E}|\boldsymbol{t}) &= \frac{p(\mathcal{E},\boldsymbol{t})}{p(\boldsymbol{t})} = \frac{\int_{\boldsymbol{\Psi}} p(\mathcal{E}|\boldsymbol{t},\boldsymbol{\psi})\, p(\boldsymbol{t},\psi)\, d\boldsymbol{\psi}}{p(\boldsymbol{t})} \\
p(\boldsymbol{t}|\mathcal{E}) &= \int_{\boldsymbol{\Psi}} p(\boldsymbol{t},\boldsymbol{\psi}|\mathcal{E})\, d\boldsymbol{\psi}\ .
\end{aligned}
$$

These respect Bayes' Theorem in $\boldsymbol{t}$. More generally, if $\boldsymbol{t}$ is a projection $\boldsymbol{t}(\boldsymbol{\theta})$ (e.g., projecting Cartesian coordinates to the radius) the marginal distributions are

$$
\begin{aligned}
p(\boldsymbol{t}) &= \int_{\boldsymbol{\Theta}} \delta[\boldsymbol{t}-\boldsymbol{t}(\boldsymbol{\theta})]\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
p(\mathcal{E}|\boldsymbol{t}) &= \frac{\int_{\boldsymbol{\Theta}} \delta[\boldsymbol{t}-\boldsymbol{t}(\boldsymbol{\theta})]\, p(\mathcal{E}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{p(\boldsymbol{t})} \\
p(\boldsymbol{t}|\mathcal{E}) &= \int_{\boldsymbol{\Theta}} \delta[\boldsymbol{t}-\boldsymbol{t}(\boldsymbol{\theta})]\, p(\boldsymbol{\theta}|\mathcal{E})\, d\boldsymbol{\theta}\ ,
\end{aligned}
$$

which also respect Bayes' Theorem. Therefore, as long as the complete prior, $p(\boldsymbol{t},\boldsymbol{\psi})$ or $p(\boldsymbol{\theta})$, and the complete likelihood, $p(\mathcal{E}|\boldsymbol{t},\boldsymbol{\psi})$ or $p(\mathcal{E}|\boldsymbol{\theta})$, is known, we can, at least in principle, compute the marginal distributions needed to compute the evidentiary CR.

Computing the evidentiary CI (i.e., when $\boldsymbol{t} \to t$ is a scalar) will usually not be too difficult, even when $t(\boldsymbol{\theta})$ is nonlinear. If analytic forms of $p(\mathcal{E}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are available, any of the equivalent conditions that specify $\mathbb{C}^{\varepsilon}_{\gamma}$ can be solved using standard constrained minimization methods (Nocedal and Wright 2006). If they are not available, we can determine them by numerically differentiating the prior and posterior cumulative distributions

$$
\begin{aligned}
P(t) &= \int_{\boldsymbol{\Theta}} H[t-t(\boldsymbol{\theta})]\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
P(t|\mathcal{E}) &= \int_{\boldsymbol{\Theta}} H[t-t(\boldsymbol{\theta})]\, p(\boldsymbol{\theta}|\mathcal{E})\, d\boldsymbol{\theta}\ ,
\end{aligned}
$$

which can often be computed using Monte Carlo methods. [$H(\cdot)$ is the Heaviside step function.] Rather than computing $p(\mathcal{E}|t)$ from the ratio $p(t|\mathcal{E})/p(t)$, it can more stably be computed as

$$
p(\mathcal{E}|t) \propto \left.\frac{dP[t^{\text{prior}}(x)|\mathcal{E}]}{dx}\right|_{x=P(t)}
$$

where $t^{\text{prior}}(\cdot)$ is the inverse of $P(\cdot)$. [Applying the chain rule for differentiation shows that the right-hand side equals $p(t|\mathcal{E})/p(t)$.] $t^{\text{prior}}(\cdot)$ can be simply computed from the numerical specification of $P(t)$, and $\mathbb{C}^{\varepsilon}_{\gamma}$ can then be solved, as before, by constrained minimization.

In the most common case, where $p(\mathcal{E}|t)$ is unimodal, $\mathbb{C}_\gamma^\varepsilon$ will be a connected interval determined by the lower and upper bounds t_ℓ and t_u,

$$\mathbb{C}_\gamma^\varepsilon = \{t : t_\ell \leq t \leq t_u\} \ ,$$

which can be determined directly from the cumulative distributions without differentiation. In this case, (12) and (15) imply that the pair (t_ℓ, t_u) satisfies

$$\begin{aligned} (t_\ell, t_u) &= \operatorname*{argmin}_{(t_1,t_2)} P(t)|_{t_1}^{t_2} \\ \text{with} \quad \gamma &= P(t|\mathcal{E})|_{t_\ell}^{t_u} \ . \end{aligned} \tag{16}$$

The constraint is satisfied by setting

$$t_u(t_\ell) = t_\ell + t^{\text{post}}[P(t_\ell|\mathcal{E}) + \gamma] \ , \tag{17}$$

where $t^{\text{post}}(\cdot)$ is the numerical inverse of $P(\cdot|\mathcal{E})$. t_ℓ can then be determined by solving the one-dimensional constrained minimization problem

$$\begin{aligned} t_\ell &= \operatorname*{argmin}_{t} P(t)|_{t}^{t_u(t)} \\ t_\ell &\leq t^{\text{post}}(1-\gamma) \ , \end{aligned}$$

where the inequality is required for (17) to have a solution.