Uncertain Evidence in Probabilistic Models and Stochastic Simulators

Andreas Munk [1] **Alexander Mead** [1] **Frank Wood** [1 2 3]

Abstract

We consider the problem of performing Bayesian inference in probabilistic models where observations are accompanied by uncertainty, referred to as “uncertain evidence.” We explore how to interpret uncertain evidence, and by extension the importance of proper interpretation as it pertains to inference about latent variables. We consider a recently-proposed method “distributional evidence” as well as revisit two older methods: Jeffrey’s rule and virtual evidence. We devise guidelines on how to account for uncertain evidence and we provide new insights, particularly regarding consistency. To showcase the impact of different interpretations of the same uncertain evidence, we carry out experiments in which one interpretation is defined as “correct.” We then compare inference results from each different interpretation illustrating the importance of careful consideration of uncertain evidence.

1. Introduction

In classical Bayesian inference, the task is to infer the posterior density $p(x|y) \propto p(y,x)$ over the latent variable x given (observed) y. The joint density (or model), $p(y,x)$, is assumed known, and typically factorizes as $p(y,x) = p(y|x)p(x)$ where $p(y|x)$ and $p(x)$ are the likelihood and prior respectively. This paper deals with the case where y is not observed exactly; rather it is associated with uncertainty[1] which we refer to as “uncertain evidence.” This is a common scenario as these uncertainties may stem from: observational errors; distrust in the source providing y; or when y is derived (stochastically) from some other data.

As an example, consider the experiment of recording the time t it takes for a ball to drop to the ground in order to determine the acceleration due to gravity, g. Taking some prior belief about the value of g, we may solve this problem using Bayesian inference. That is, we infer $p(g|t) \propto p(g)p(t|g)$, where $p(g)$ is the prior density of g and $p(t|g)$ is the likelihood representing the physical model (or simulation) of the time t given g. In this setup, the uncertainty about t given g would be due to neglecting air resistance or ignoring variations in the distance the ball drops as a result of vibrations etc. Assume next that the observations (or data) is given as in Table 1. It is not immediately obvious how the uncertainty relates to t. There are arguably at least two valid interpretations of the information in Table 1: (1) it describes a distribution of the real time t. For example, the real time is normally distributed with mean 0.5s and standard deviation 0.05s. (2) It describes additional uncertainty on the predicted time and the observed value is, indeed, 0.5s. For example, given the predicted time t the observed time $\hat{t}$ is normally distributed with mean t and standard deviation 0.05s. Importantly, in either case the uncertainty can be represented with a given *external*[2] density $q(\cdot|\cdot)$ which describes a stochastic relationship between t and an auxiliary variable ζ. We consider in case (1) and (2) the density $q(t|\zeta)$ and $q(\zeta|t)$. In the former case ζ is left implicit (something gave rise to the uncertainty). In the latter $\zeta = \hat{t}$ and the observation is $\hat{t} = 0.5$s. These two approaches are fundamentally different operations that may lead to profoundly different inference results.

The topic of observations associated with uncertainty has been studied since at least 1965 (Jeffrey, 1965). Of particular relevance are the work of Jeffrey (1965) and Shafer (1981); and Pearl (1988), giving rise to *Jeffrey’s rule* (Jeffrey, 1965; Shafer, 1981) and *virtual evidence* (Pearl, 1988). In the example above, inference using approach (1) or (2) corresponds to Jeffrey’s rule or virtual evidence respectively. Since thenm other approaches, closely related to Jeffrey’s rule and virtual evidence, has been proposed (e.g., Valtorta et al., 2002; Tolpin et al., 2021; Yao, 2022). While each approach has its own merits and is applicable under (almost) the same circumstances, the original literature and most prior work comparing these methods, (e.g., Pearl, 2001; Valtorta et al., 2002; Chan & Darwiche, 2005; Ben Mrad et al.,

[1]Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada [2]Inverted AI Ltd., Vancouver, B.C., Canada [3]Mila, CIFAR AI Chair. Correspondence to: Andreas Munk <amunk@cs.ubc.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Ideally one would remodel the system to account for such uncertainties, but this is rarely easy to do.

[2] In this context, external refers to a distribution provided from some external source.

Table 1. Uncertain observation associated with the time t in the *drop of a ball* example.

	VALUE [s]	±[s]
t	0.5	0.05

2013; Tolpin et al., 2021), are reluctant to take a concrete stand on when each is more appropriate.

This paints an obfuscated picture of what to do, practically, when presented with uncertain evidence. This obfuscation becomes problematic when practitioners outside the field of statistics deal with uncertain evidence and look to the literature for ways to address it. Especially now, considering the increased use of Bayesian inference in high-fidelity simulators and probabilistic models (e.g., Papamakarios et al., 2019; Baydin et al., 2019; Lavin et al., 2021; Liang et al., 2021; van de Schoot et al., 2021; Wood et al., 2022; Mishra-Sharma & Cranmer, 2022; Munk et al., 2022). For example, in physics it is not uncommon that likelihoods are given relatively ad-hoc forms where some notion of "measurement error" is attached to uncertain observations. However, the underlying (stochastic) *physical* model is usually taken to be understood perfectly. This is the case, for instance, when inferring; the Hubble parameter via supernovae brightness (e.g., Riess et al., 2022); pre-merger parameters of black-hole/neutron star binaries via gravitational waves (e.g., Thrane & Talbot, 2019; Dax et al., 2021); neutron star orbital/spin-down/post-Newtonian parameters via pulsar timings (e.g., Vigeland & Vallisneri, 2014; Lentati et al., 2014); planetary orbital parameters via radial velocity/transit-time observations (e.g., Schulze-Hartung et al., 2012; Feroz & Hobson, 2014; Liang et al., 2021). In most cases a Gaussian likelihood is assumed for the data, but exactly how the error relates to the data generation process is not specified. If uncertainties about simulator/model observations arise given external data, then usually Jeffrey's rule would apply, but it appears that virtual evidence is more often employed.

It is the purpose of this paper to provide novel insights, theoretical contributions and guidance as to how to deal with observations with associated uncertainty as it pertains to Bayesian inference. We show, experimentally, how misinterpretations of uncertain evidence can lead to vastly different inference results; emphasizing the importance of carefully accounting for uncertain evidence.

2. Background

Bayesian inference aims to characterize the posterior distribution of the latent random vector $\boldsymbol{x}$ given the observed random vector $\boldsymbol{y}$. When observing $\boldsymbol{y}$ with certainty the inference problem is "straightforward" in the sense that $p(\boldsymbol{x}|\boldsymbol{y}) = p(\boldsymbol{y},\boldsymbol{x})/p(\boldsymbol{y})$. However, exact inference is often infeasible as $p(\boldsymbol{y})$ is usually intractable, but if the joint $p(\boldsymbol{y},\boldsymbol{x})$ is calculable then inference is achievable via approximate methods such as importance sampling (e.g., Hammersley & Handscomb, 1964), Metropolis-Hastings (Metropolis & Ulam, 1949; Metropolis et al., 1953; Hastings, 1970), and Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 1994). Unfortunately, standard Bayesian inference is incompatible with uncertain evidence where exact values of $\boldsymbol{y}$ are unavailable.

Before discussing ways to treat uncertain evidence, we first introduce the highest level abstraction representing uncertain evidence. Specifically, we consider $\epsilon \in \mathcal{E}$, where $\mathcal{E}$ is a set of "statements" specifying the uncertainty about $\boldsymbol{y}$. For example, in the drop of a ball example ϵ would be a statement represented as Table 1. In contrast, ζ is a lower level abstraction which is encoded in ϵ. Dealing with uncertain evidence is a matter of decoding or interpreting ϵ, possibly identifying ζ and relating it to $p(\boldsymbol{y},\boldsymbol{x})$. The canonical example of interpreting uncertain evidence, as introduced by Jeffrey (1965, p. 165), is "observation by candlelight," which motivated *Jeffrey's rule*:

Definition 2.1 (Jeffrey's Rule (Jeffrey, 1965))**.** *Given $p(\mathbf{y},\mathbf{x})$, let the interpretation of a given $\epsilon \in \mathcal{E}$ lead to $\mathbf{y}$ being associated with uncertainty, conditioned on auxiliary evidence ζ—where ζ may be unknown—and denote the decoded uncertainty by $q(\mathbf{y}|\zeta)$. Then the updated (posterior) density $p(\mathbf{x}|\zeta)$ is:*

$$p(\mathbf{x}|\zeta) = \int p(\mathbf{x}|\mathbf{y})q(\mathbf{y}|\zeta)\,\mathrm{d}\mathbf{y}. \tag{1}$$

In particular, one considers the updated joint $p(\mathbf{y},\mathbf{x}|\zeta) = p(\mathbf{x}|\mathbf{y})q(\mathbf{y}|\zeta)$, such that $q(\mathbf{y}|\zeta)$ is a marginal of $p(\mathbf{y},\mathbf{x}|\zeta)$.

Jeffrey envisioned the existence of the auxiliary variable (or vector), ζ; however, Jeffrey's rule is often defined without it (e.g., Chan & Darwiche, 2005). Nonetheless, we argue that reasoning about an auxiliary variable (or vector) ζ is the more intuitive perspective as *some* evidence must have given rise to q. Further, accompanying the introduction of Jeffrey's rule is the preservation of the conditional distribution of $\boldsymbol{x}$ upon applying Jeffrey's rule (e.g., Jeffrey, 1965; Diaconis & Zabell, 1982; Valtorta et al., 2002; Chan & Darwiche, 2005). That is, the evidence ζ giving rise to $q(\boldsymbol{y}|\zeta)$ must not also alter the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ (and ζ). Mathematically, Jeffrey's rule requires that, $p(\boldsymbol{x}|\boldsymbol{y},\zeta) = p(\boldsymbol{x}|\boldsymbol{y})$. This, for instance, relates to the commutativity of Jeffrey's rule—also referred to as the issue of iterated revision (Chan & Darwiche, 2005). This topic is treated in full detail by Diaconis & Zabell (1982) and further discussed by Wagner (2002). For completeness a brief discussion in Appendix A.

In contrast to Jeffrey's rule is *virtual evidence*, as proposed

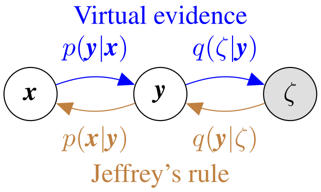
Virtual evidence
$p(\boldsymbol{y}|\boldsymbol{x})$
$q(\zeta|\boldsymbol{y})$
$\boldsymbol{x}$
$\boldsymbol{y}$
ζ
$p(\boldsymbol{x}|\boldsymbol{y})$
$q(\boldsymbol{y}|\zeta)$
Jeffrey's rule

Figure 1. Jeffrey's rule compared to virtual evidence in terms of the auxiliary evidence ζ. Both virtual evidence and Jeffrey's rule are defined in terms of the base model $p(\boldsymbol{y}, \boldsymbol{x})$.

by Pearl (1988). Virtual evidence also includes an auxiliary *virtual* variable (or vector), but does so via the likelihood $q(\zeta|\boldsymbol{y}, \boldsymbol{x}) := q(\zeta|\boldsymbol{y})$, with the only parents of ζ being $\boldsymbol{y}$:

Definition 2.2 (Virtual evidence (Pearl, 1988)**).** *Given* $p(\mathbf{y}, \mathbf{x})$ *and suppose a given* $\epsilon \in \mathcal{E}$ *leads to the interpretation that we extend* $p(\mathbf{y}, \mathbf{x})$ *with an auxiliary virtual variable (or vector)* ζ *such that: (1) in the discrete case,* $\mathbf{y} \in \{\mathbf{y}_k\}_{k=1}^K$*, the uncertain evidence is decoded as likelihood ratios*[3] $\{\lambda_k\}_{k=1}^K$*:*

$$\lambda_1 : \cdots : \lambda_K = q(\zeta|\mathbf{y}_1) : \cdots : q(\zeta|\mathbf{y}_K). \tag{2}$$

The posterior over $\mathbf{x}$ *given uncertain evidence is (Chan & Darwiche 2005; a result we also prove in Appendix B):*

$$p(\mathbf{x}|\zeta) = \frac{\sum_{k=1}^K \lambda_k p(\mathbf{y}_k, \mathbf{x})}{\sum_{j=1}^K \lambda_j p(\mathbf{y}_j)} . \tag{3}$$

(2) If $\mathbf{y}$ *is continuous, decoding* ϵ *leads to the virtual likelihood* $q(\zeta|\mathbf{y})$ *such that the posterior is proportional to the (virtual) joint*

$$p(\mathbf{x}|\zeta) \propto \int p(\zeta, \mathbf{y}, \mathbf{x}) \, \mathrm{d}\mathbf{y} = \int q(\zeta|\mathbf{y}) p(\mathbf{y}, \mathbf{x}) \, \mathrm{d}\mathbf{y}. \tag{4}$$

In practice, in the continuous case one can approximate the posterior using standard approximate inference algorithms requiring only the evaluation of the joint. In the discrete case, Eq. 3, the posterior inference is exact assuming a known $p(\boldsymbol{y}_i)$ for all $i \in \{1, \ldots, K\}$. When comparing Jeffrey's rule and virtual evidence (e.g., Pearl, 1988; Valtorta et al., 2002; Jacobs, 2019) we can do so in terms of ζ and the corresponding graphical model, see Figure 1. This figure is a graphical representation of how Jeffrey's rule and virtual evidence relate ζ to the existing probabilistic model, $p(\boldsymbol{y}, \boldsymbol{x})$. Particularly, Jeffrey's rule and virtual evidence affect the model in *opposite* directions. Jeffrey's rule pertains to uncertainty about $\boldsymbol{y}$ *given* some evidence, while virtual evidence requires reasoning about $q(\zeta|\boldsymbol{y})$.

It is (perhaps) not surprising that one may apply Jeffrey's rule, yet implement it as a special case of virtual evidence, by choosing a particular form of likelihood ratios, Equation (2), and vice versa (Pearl, 1988; Chan & Darwiche, 2005). However, this is of purely algorithmic significance as the two approaches remain fundamentally different.

A third approach to uncertain evidence, recently introduced by Tolpin et al. (2021), treats the uncertain evidence on $\boldsymbol{y}$ as an event. This approach, which here is referred to as *distributional evidence*, defines a likelihood on the event $\{\boldsymbol{Y} \sim \mathbb{Q}\}$ (reads as "the event that the distribution of $\boldsymbol{Y}$ is $\mathbb{Q}$ with density $q(\boldsymbol{y})$") and considers the auxiliary variable $\zeta = \{\boldsymbol{Y} \sim \mathbb{Q}\}$:

Definition 2.3 (Distributional evidence (Tolpin et al., 2021)**).** *Let* $p(\mathbf{y}, \mathbf{x}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ *be the joint density with a known factorization. Assume the interpretation of a given* $\epsilon \in \mathcal{E}$ *yields a density* $q(\mathbf{y})$*, with distribution* $\mathbb{Q}$*. Define the likelihood* $p(\mathbf{Y} \sim \mathbb{Q}|\mathbf{x})$ *as:*

$$p(\mathbf{Y} \sim \mathbb{Q}|\mathbf{x}) = \frac{\exp \int (\ln p(\mathbf{y}|\mathbf{x})) \, q(\mathbf{y}) \, \mathrm{d}\mathbf{y}}{Z(\mathbf{x})} \tag{5}$$

where $Z(\mathbf{x})$ *is a normalization constant that generally depends on* $\mathbf{x}$*. See (Tolpin et al., 2021) for sufficient conditions for which* $Z(\mathbf{x}) < \infty$*.*

3. Which Approach?

The lack of general consensus on how best to approach uncertain evidence means that it is difficult to know what to do, in practical terms, when faced with uncertain evidence. In isolation, each approach discussed in the previous section appears well supported, even when applied to the same model (e.g., Ben Mrad et al., 2013). However, the underlying arguments remain somewhat circumstantial. Prior work tends to create contexts tailored for each approach and it is unclear how relatable or generalizable those contexts are. As such, much prior work is not particularly instructive when deducing which approach to adopt for new applications that do not fit those prior context. We argue that the apparent philosophical discourse fundamentally stem from a disagreement about the model $M \in \mathcal{M}$ in which we seek to do inference given uncertain evidence $\epsilon \in \mathcal{E}$. This can be framed as an inference problem where we seek to find (or directly define) $p(M|\epsilon)$. The significance of this perspective is that reasoning about the triplet $M \in \mathcal{M}$, $\epsilon \in \mathcal{E}$, and $p(M|\epsilon)$ makes for a better foundation that encourages discussions about and makes clear the underlying assumptions.

How then should we define $p(M|\epsilon)$? In the general case, reaching consensus is close to impossible as it requires fully specifying $\mathcal{M}$ and $\mathcal{E}$ (all possible models and conceivable evidences). However, while universal consensus is arguably unattainable; "local" consensus might be. Here locality

[3] The notation for ratios containing several terms, for example A, B, and C, is written as $x : y : z$. This is understood as: "for every x part of A there is y part B and z part C."

refers to defining $p(M|\epsilon)$ on constrained and application dependent subsets $\tilde{\mathcal{E}} \subset \mathcal{E}$ and $\tilde{\mathcal{M}} \subset \mathcal{M}$. This perspective was considered by (Grove & Halpern, 1997), yet does not seem to have resurfaced in this context since. Grove & Halpern (1997) define $\tilde{\mathcal{M}}$ in terms of a prior $p(M)$ and implicitly defines $\tilde{\mathcal{E}}$ as a set of trusted statements pertaining to (conditional) probabilities. They further define the likelihood $p(\epsilon|M)$ which evaluates to one if the model M is consistent with the evidence ϵ and zero otherwise. From this they are able to compute $p(M|\epsilon) \propto p(\epsilon|M)p(M)$.

3.1. Uncertain Evidence Interpretation

We propose to limit the consideration of $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{M}}$ to constrained, but widely applicable (and application dependent) subsets set in the context of inference. To construct $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{M}}$, begin with the assumption that a *base model* $p(\boldsymbol{y}, \boldsymbol{x})$ is always available. Further, assume that $\tilde{\mathcal{E}}$ contains evidence in the form of statements which is interpreted in a literal sense. To ensure inference with exact evidence is possible, it is required that $\tilde{\mathcal{E}}$ contain evidences that encode exact evidence about $\boldsymbol{Y}$. For example $\epsilon =$ "the value of $\boldsymbol{Y}$ is $\boldsymbol{y}$." Finally, constrain the form of *uncertain evidence* by requiring ϵ to encode uncertainty in one of three ways: (I) ϵ encodes a density q over $\boldsymbol{y}$, for example $\epsilon =$ "The density of $\boldsymbol{y}$ is $q(\boldsymbol{y}|\zeta)$". (II) ϵ encodes a *conditional* density about $\boldsymbol{y}$ given $\boldsymbol{X} = \boldsymbol{x}$ (or a subset of the latent variables). For example $\epsilon =$ "iff $\boldsymbol{X} = \boldsymbol{x}$ then the density of $\boldsymbol{y}$ is $q(\boldsymbol{y}|\boldsymbol{x})$." (III) Uncertain evidence is explicitly expressed in terms of a likelihood of $\boldsymbol{y}$, for example let $\boldsymbol{y} \in \{0, 1\}$ and consider $\epsilon =$ "$\boldsymbol{y} = 1$ is twice as likely to explain the evidence compared to $\boldsymbol{y} = 0$." Define $\tilde{\mathcal{M}}$ implicitly by requiring that the random variable ϵ partitions $\tilde{\mathcal{M}}$ such that the posterior $p(\boldsymbol{x}|\epsilon)$ takes a certain form:

Definition 3.1. *Given $\epsilon \in \tilde{\mathcal{E}}$, define $p(\tilde{\mathcal{M}}|\epsilon)$ and $\tilde{\mathcal{M}}$ implicitly through the partitions of $\tilde{\mathcal{M}}$ as generated by ϵ, such that inference given ϵ becomes,*

$$\begin{aligned} p(\mathbf{x}|\epsilon) &= \mathbb{E}_{p(M|\epsilon)}[p(\mathbf{x}|M)] \\ &= \begin{cases} p(\mathbf{x}|\mathbf{y}), & \textit{if } \epsilon \textit{ is exact}, \\ \int p(\mathbf{x}|\mathbf{y})q(\mathbf{y}|\zeta)\,\mathrm{d}\mathbf{y}, & \textit{if } \epsilon \textit{ is type (I)}, \\ \frac{p(\mathbf{x})p(\mathbf{Y}\sim\mathbb{Q}|\mathbf{x})}{p(\mathbf{Y}\sim\mathbb{Q})}, & \textit{if } \epsilon \textit{ is type (II)}, \\ \frac{\int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})q(\zeta|\mathbf{y})\,\mathrm{d}\mathbf{y}}{p(\zeta)}, & \textit{if } \epsilon \textit{ is type (III)}, \end{cases} \end{aligned} \tag{6}$$

where type (I-III) leads to Jeffrey's rule, distributional evidence, and virtual evidence respectively. It should be emphasize, that the definitions of $\tilde{\mathcal{E}}$, $\tilde{\mathcal{M}}$, and $p(\tilde{\mathcal{M}}|\epsilon)$ are *not* fundamental truths. Rather, they are proposed beliefs about how one ought to approach uncertain evidence in a form found in $\tilde{\mathcal{E}}$. In particular, notice that type (I) and (II) evidences are similar in that they describe a distribution of $\boldsymbol{y}$. The crucial difference lies in the conditional relationship giving rise to said probability. In type (I) uncertainty is assumed due to external (unknown) evidence, represented by ζ not found in $\boldsymbol{x}$ or $\boldsymbol{y}$. On the other hand, in type (II) ζ *is* $\boldsymbol{x}$ (or a subset thereof). Even though Jeffrey's rule is proposed to be preferable given type (I) uncertain evidence, it turns out there are cases where Jeffrey's rule is, in fact, inconsistent with $p(\boldsymbol{y}, \boldsymbol{x})$. This is shown shown in the following section. Nonetheless, from a mathematical perspective, Jeffrey's rule can still be applied. This is justified, in part, as Jeffrey's rule leads to a "new" model $p(\boldsymbol{y}, \boldsymbol{x}|\zeta)$ which is closest to $p(\boldsymbol{y}, \boldsymbol{x})$ as measured by the KL divergence $\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y}, \boldsymbol{x}|\zeta)||p(\boldsymbol{y}, \boldsymbol{x}))$ *constrained* such that $\int p(\boldsymbol{y}, \boldsymbol{x}|\zeta)\mathrm{d}\boldsymbol{x} = q(\boldsymbol{y}|\zeta)$ (Peng et al., 2010, and citations therein). Despite this, if Jeffrey's rule is inconsistent with $p(\boldsymbol{y}, \boldsymbol{x})$ it may be preferable to either: (1) update the model $p(\boldsymbol{y}, \boldsymbol{x})$ to be compatible with the given uncertain evidence or (2) acquire compatible data—be it exact observations or "better" uncertain evidence.

3.2. Consistency

We define consistency in terms of whether or not one can extend the joint distribution with auxiliary variables (or vectors) such as to contain the uncertainty encoded in $\epsilon \in \tilde{\mathcal{E}}$,

Definition 3.2 (Consistency)**.** *Consider an auxiliary variable (or vector) ζ and the associated density q derived from ϵ, where q can take the form of either $q(\zeta|\cdot)$ or $q(\cdot|\zeta)$. We then say that Jeffrey's rule, virtual evidence, and distributional evidence are consistent with $p(\mathbf{y}, \mathbf{x})$ if a joint exists, $p(\zeta, \mathbf{y}, \mathbf{x}) = p(\zeta|\mathbf{y}, \mathbf{x})p(\mathbf{y}, \mathbf{x})$, such that either $p(\zeta|\cdot) = q(\zeta|\cdot)$ or $p(\cdot|\zeta) = q(\cdot|\zeta)$ depending on the form of q.*

Both virtual evidence and distributional evidence are, by their definition, always consistent. Virtual evidence is defined as an extension of the graphical model $p(\boldsymbol{y}, \boldsymbol{x})$ through the auxiliary variable (or vector) ζ and its likelihood $q(\zeta|\boldsymbol{y})$. That is we can always consider $p(\zeta|\boldsymbol{y}) := q(\zeta|\boldsymbol{y})$ such that $p(\zeta, \boldsymbol{y}, \boldsymbol{x}) := p(\zeta|\boldsymbol{y})p(\boldsymbol{y}, \boldsymbol{x})$. Similarly, in the case of distributional evidence, we can consider $p(\zeta, \boldsymbol{y}, \boldsymbol{x}) := p(\zeta|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$. However, despite distributional evidence being consistent, notice that it introduces ζ as independent of $\boldsymbol{y}$. As such, distributional evidence introduces an entirely new likelihood with respect to $\boldsymbol{x}$ and we can consider $p(\zeta|\boldsymbol{x})p(\boldsymbol{x})$ as a *new* model. This results in the loss of the physical interpretation of the relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ as defined through $p(\boldsymbol{y}|\boldsymbol{x})$ even though $p(\zeta|\boldsymbol{x})$ is derived from $p(\boldsymbol{y}|\boldsymbol{x})$. On the other hand, in the case of Jeffrey's rule, we cannot guarantee consistency, and so one needs to be mindful of the potential mismatch between the base model and $q(\boldsymbol{y}|\zeta)$. While (Diaconis & Zabell, 1982) provide an extensive and theoretical examination of Jeffrey's rule they leave out important points concerning necessary conditions for Jeffrey's rule to satisfy consistency that we present here and prove in Appendix B.1:

Theorem 3.3 (Consistency of Jeffrey's rule)**.** *Suppose* $\mathbf{x}$ *and* $\mathbf{y}$ *are random vectors and* ζ *is the auxiliary random variable tied to uncertain evidence. Then the necessary and sufficient conditions for Jeffrey's rule to be consistent with respect to* $p(\mathbf{y}, \mathbf{x})$ *given* $q(\mathbf{y}|\zeta)$ *(i.e. there exists a joint* $p(\zeta, \mathbf{y}, \mathbf{x})$ *such that* $p(\mathbf{y}|\zeta) = q(\mathbf{y}|\zeta)$*) are as follows:*

1. *(Necessary and sufficient) There exists* $p(\zeta|\mathbf{y})$ *such that for all* ζ *and* $\mathbf{y}$*,*

$$q(\mathbf{y}|\zeta) = \frac{p(\zeta|\mathbf{y})p(\mathbf{y})}{\int p(\zeta|\mathbf{y})p(\mathbf{y})\,\mathrm{d}\mathbf{y}}$$

2. *(Necessary) If* $q(\mathbf{y}|\zeta) = \prod_{i=1}^{D} q(y_i|\zeta)$ *then it must hold that: (1)* ζ *is a random vector* $\zeta = (\zeta_1, \ldots, \zeta_D)$ *where each* ζ_i *uniquely links to* y_i *such that* $q(y_i|\zeta) = q(y_i|\zeta_i)$*. (2)* $\mathbf{x}$ *is likewise multivariate and each* x_i *uniquely links to* y_i *such that* $p(y_i|\mathbf{x}) = p(y_i|x_i)$*.*

3. *(Necessary) Let* $p(\zeta) = \int p(\zeta|\mathbf{y})\,\mathrm{d}\mathbf{y}$*, then it must hold that: (1)* $\mathrm{Cov}\left[\mathbf{Y}\right] \succeq \mathbb{E}\left[\mathrm{Cov}\left[\mathbf{Y}|\zeta\right]\right]$*, where* $\succeq$ *denotes determinant inequality. (2) For each* Y_i *(the constituent random variables of* $\mathbf{Y}$*) it holds that* $\mathrm{Var}\left[Y_i\right] \geq \mathbb{E}\left[\mathrm{Var}\left[Y_i|\zeta\right]\right]$*. In particular, if the variance* $\mathrm{Var}\left[Y_i|\zeta\right] = \sigma^2$ *is constant and independent of* ζ *if follows that* $\mathrm{Var}\left[Y_i\right] \geq \sigma^2$ *with equality if and only if* $\mathbb{E}\left[Y_i|\zeta\right] = \mu$ *is constant.*

Unfortunately, validating consistency of Jeffrey's rule is in general infeasible as Theorem 3.3 (1) is usually intractable to assess. One can only reliably conclude if Jeffrey's rule is inconsistent in special cases via Theorem 3.3 (2) and (3).

3.3. Distributional Evidence: Exact or Implied Inference?

While we generally prefer Jeffrey's rule over distributional evidence and although Jeffrey's rule is technically applicable given type (II) uncertain evidence, why then do we prefer distributional evidence preferred given type (II)? If Jeffrey's rule were to be used in this case its interpretation becomes unclear if q is of the form $q(\boldsymbol{y}|g(\boldsymbol{x}))$ where $g(\cdot)$ is a selector function which selects a subset of the variables in $\boldsymbol{x}$. As the task is to ultimately infer a posterior over the latent variables given $\zeta = g(\boldsymbol{x})$, it violates the intuition that ζ should be an auxiliary variable (or vector) not found in $\boldsymbol{x}$, which is required by Jeffrey's rule. Specifically, consider two kinds of uncertain evidence of this form: (1) a functional $q(\boldsymbol{y}|g(\boldsymbol{x}))$ specified for all $\boldsymbol{x}$ and $\boldsymbol{y}$ and (2) a conditional form such that q is a density specified for only a specific value of $g(\boldsymbol{x}) = g(\hat{\boldsymbol{x}})$. In case (1) one arguably ought to replace the model $p(\boldsymbol{y}, \boldsymbol{x}) \rightarrow q(\boldsymbol{y}|g(\boldsymbol{x}))p(\boldsymbol{x})$ such that $q(\boldsymbol{y}|g(\boldsymbol{x}))$ becomes the new likelihood. However, in case (2) the model cannot simply be replaced as the form of q is unknown for any other value of $g(\boldsymbol{x})$ than $g(\hat{\boldsymbol{x}})$. In particular, one can think of case (2) as the limiting case of observing $\mathcal{D} = \{\boldsymbol{y}_i\}_{i=1}^{N}$ for $N \rightarrow \infty$, where each $\boldsymbol{y}_i$ is i.i.d. with probability density $p(\boldsymbol{y}|g(\hat{\boldsymbol{x}}))$. In the limit, the empirical distribution of $\mathcal{D}$ will represent $p(\boldsymbol{y}|g(\hat{\boldsymbol{x}}))$. As pointed out also by Tolpin et al. (2021) there is a similarity between observing $\mathcal{D}$ for large N and instead condition on $q(\boldsymbol{y})$ associated with the empirical distribution represented by $\mathcal{D}$. From this perspective, distributional evidence provides for inferring $p(\boldsymbol{x}|\boldsymbol{Y} \sim \mathbb{Q})$ as opposed to $p(\boldsymbol{x}|\mathcal{D})$. This view is useful, particularly when $\mathcal{D}$ is unavailable yet its distributive representation q is. For example, if provided a q in terms of summary statistics or quantiles it seems reasonable to condition on the event $\{\boldsymbol{Y} \sim \mathbb{Q}\}$ via Equation (5); an example that Tolpin et al. (2021) also showcase. Arguably, if $q(\boldsymbol{y}) = p(\boldsymbol{y}|g(\boldsymbol{x}))$ one may be better off by simply sampling the dataset $\mathcal{D}$ from $\mathbb{Q}$ and perform standard inference. However, if $q \approx p(\boldsymbol{y}|g(\boldsymbol{x}))$, then distributional evidence and thereby conditioning on q may be more appropriate.

One caveat to distributional evidence, that Tolpin et al. (2021) do not discuss, is whether or not $Z(\boldsymbol{x})$ in Equation (5) is calculable. In particular, Tolpin et al. (2021) appears to leave it as a normalization constant that is never calculated. That is, they compute the function $f(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x}) = p(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})Z(\boldsymbol{x})$ when performing inference, where f is the numerator in Equation (5)—a "pseudo-likelihood." The difference between computing $p(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})$ and $f(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})$ in the context of inference is:

$$p(\boldsymbol{x}|\boldsymbol{Y} \sim \mathbb{Q}) \propto \begin{cases} p(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})p(\boldsymbol{x}) & \text{if known } Z(\boldsymbol{x}), \\ f(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})p(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

While the first expression above leads to posterior inference as expected, the second expression leads to an implied posterior via the implied joint:

$$\begin{aligned} f(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})p(\boldsymbol{x}) &= p(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})p(\boldsymbol{x})Z(\boldsymbol{x}) \\ &= p(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})\hat{p}_{\mathrm{a}}(\boldsymbol{x}), \end{aligned} \tag{7}$$

where $\hat{p}_{\mathrm{a}}(\boldsymbol{x}) = p(\boldsymbol{x})Z(\boldsymbol{x})$ is a *distributional evidence adjusted* un-normalized prior on $\boldsymbol{x}$. As such, regardless of whether or not a known $Z(\boldsymbol{x})$ is available, it leads to the same likelihood on the event $\{\boldsymbol{Y} \sim \mathbb{Q}\}$. However, the knowledge of $Z(\boldsymbol{x})$ leads to different priors on $\boldsymbol{x}$. To ensure that the use of Equation (7) leads to a valid posterior, it is enough to show that the adjusted prior $p_{\mathrm{a}}(\boldsymbol{x}) \propto \hat{p}_{\mathrm{a}}(\boldsymbol{x})$ normalizes in $\boldsymbol{x}$:

Theorem 3.4 (Distributional evidence normalizes)**.** *Under the same assumptions as in Theorem 1 in the paper of (Tolpin et al., 2021), the adjusted prior* $p_{\mathrm{a}}(\mathbf{x}) = p(\mathbf{x})Z(\mathbf{x})/C$ *normalizes. That is* $C < \infty$*.*

Proof. Assume, as done by (Tolpin et al., 2021), that the set of distributions $\mathcal{Q}$ is implicitly defined through the set of parameters Θ where $\theta \in \Theta$ parameterizes q_θ such that $\mathcal{Q} = \{q_\theta | \theta \in \Theta\}$. Assume further that $\sup_{\boldsymbol{y}} \int_\Theta q_\theta(\boldsymbol{y}) \, \mathrm{d}\theta < \infty$. Then the bound on $Z(\boldsymbol{x})$, as derived by (Tolpin et al., 2021), is independent of $\boldsymbol{x}$. It then follows that $Z(\boldsymbol{x}) \leq \tilde{Z} < \infty$ for all $\boldsymbol{x}$ such that:

$$\begin{aligned} C &= \int \hat{p}_{\mathrm{a}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{x}) Z(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ &\leq \int p(\boldsymbol{x}) \tilde{Z} \, \mathrm{d}\boldsymbol{x} = \tilde{Z} < \infty. \end{aligned}$$

This implies that $p_{\mathrm{a}}(\boldsymbol{x}) = p(\boldsymbol{x}) Z(\boldsymbol{x}) / C$ is a valid marginal as it normalizes in $\boldsymbol{x}$, which concludes the proof. □

3.4. Complexity

The primary consideration when comparing Jeffrey's rule, virtual evidence, and distributional evidence, is their applicability given a certain type of uncertain evidence. In a practical setting, it is unclear by how much each approach differs in their posteriors over $\boldsymbol{x}$ given the same uncertain evidence. As we illustrate in Section 4.1, this difference may range from significant to negligible and is a function of the base model as well as the uncertain evidence. Therefore, when inference is time-sensitive, it may be beneficial to initially perform inference using an approach of low computational complexity and then subsequently follow up with the appropriate approach to verify inference results. Our complexity analysis assumes no analytical solution is feasible, and that approximate inference is employed; that is, sampling-based inference methods as well as Monte Carlo estimations of expectations is used. The computational complexity for achieving adequate approximate posterior inference is denoted c_{i}. Likewise, n_{e} is used to denote the number of required samples for adequate Monte Carlo estimations of expectations. Note that this relies on the additional assumption that inferring $p(\boldsymbol{x}|\boldsymbol{y})$ has the same complexity as inferring $p(\boldsymbol{x}|\zeta) = \int p(\boldsymbol{x}, \boldsymbol{y}|\zeta) \, \mathrm{d}\boldsymbol{y}$. From this it follows that the complexity of Jeffrey's rule, Equation (1), requires estimating the expected posterior leading to a complexity of $c_{\mathrm{i}} n_{\mathrm{e}}$, while virtual evidence is c_{i} as it only involves inferring the posterior under the joint given by Equation (4). As for distributional evidence, Equation (5), if the new likelihood is analytically tractable the complexity is c_{i}, since it requires only inferring a posterior distribution. If the likelihood is approximated using Monte Carlo estimation the complexity increases to $c_{\mathrm{i}} n_{\mathrm{e}}$.[4] Therefore, virtual evidence is, in general, more efficient than both Jeffrey's rule and distributional evidence.

Finally, we note that a reduction of the complexity gap between Jeffrey's rule and virtual evidence is achievable using amortized inference (Gershman & Goodman, 2014). Amortized inference reduces the cost of inference in exchange for an upfront computational cost. Therefore, estimating an expected posterior, which is the case for Jeffrey's rule, can be significantly sped up.

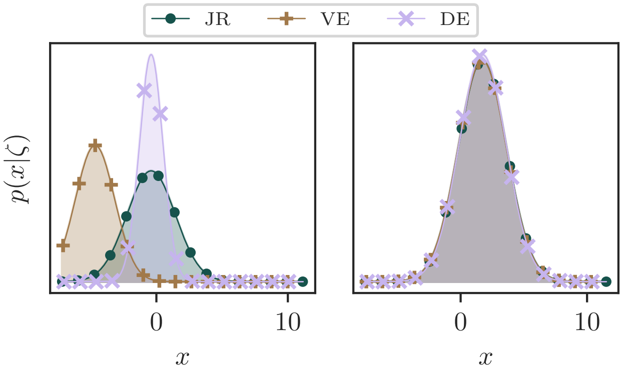

Figure 2. Analytical posterior results given uncertain evidence $q(y|\zeta)$ using Jeffrey's rule (JR), virtual evidence (VE), and distributional evidence (DE). (Left) $\mu_x = -10$, $\sigma_x = 2$, $\sigma_{y|x} = 1$, $\sigma_q = 2$, and $\zeta = 2$ from which the remaining means and (conditional) variances are derived as described in Section 4.1. (Right) same as (left) except with $\mu_x = 0$, $\sigma_x = 5$, $\sigma_{y|x} = 2$, $\sigma_q = 0.5$, and $\zeta = 2$.

4. Experiments

In this section the importance of making the appropriate interpretation and treatment of uncertain evidence is illustrated. It contains three experiments constructed such that in experiment one and three, the appropriate treatment of the given uncertain evidence is to use Jeffrey's rule. In the second experiment the appropriate treatment is virtual evidence. Comparisons are made against making a *misinterpretation*. It is demonstrated how such misinterpretations can lead to inference results that range from being significantly different to almost indistinguishable. Most prior work (e.g., Chan & Darwiche, 2005; Ben Mrad et al., 2013; Mrad et al., 2015; Jacobs, 2019) compares only Jeffrey's rule and virtual evidence for discrete problems, whereas the focus here is on the continuous case.

4.1. Uncertain Evidence and the Multivariate Gaussian

Consider a multivariate Gaussian model where the base model factorizes as $p(x, y) = p(x)p(y|x)$ with $p(x) = \mathcal{N}(\mu_x, \sigma_x^2)$ and $p(y|x) = \mathcal{N}(x, \sigma_{y|x}^2)$. The aim is to infer the posterior distribution of x, ideally given an exact observation of y. However, assume this is unavailable and what is instead given is uncertain evidence ϵ of type (I); that is, the density $q(y|\zeta) = \mathcal{N}(\zeta, \sigma_q^2)$. The aim then is to infer

[4] Although, due to recent advances for inference in models with tall data the effective number required to estimate the likelihood might be much smaller, $k_{\mathrm{e}} < n_{\mathrm{e}}$ (Tolpin et al., 2021, Section 4.). In this case the complexity is $c_{\mathrm{i}} k_{\mathrm{e}} < c_{\mathrm{i}} n_{\mathrm{e}}$.

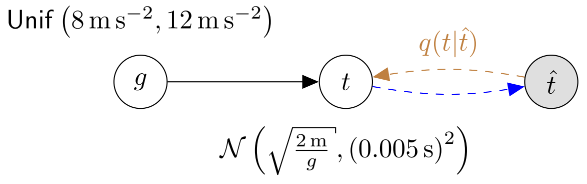
Unif $(8\,\mathrm{m\,s^{-2}}, 12\,\mathrm{m\,s^{-2}})$
g
t
$\hat{t}$
$q(t|\hat{t})$
$\mathcal{N}\left(\sqrt{\frac{2\,\mathrm{m}}{g}}, (0.005\,\mathrm{s})^2\right)$

Figure 3. Graphical model of the *drop of a ball* experiment in Section 4.2. A brown dashed edge is used to specify the unknown true conditional density for which Jeffrey's rule is inappropriately used. The blue dashed edge emphasizes that the true $p(\hat{t}|t)$, which is also recovered using virtual evidence. That is, the brown edge represent *interpreting* uncertain evidence of type (III) as type (I) leading to Jeffrey's rule.

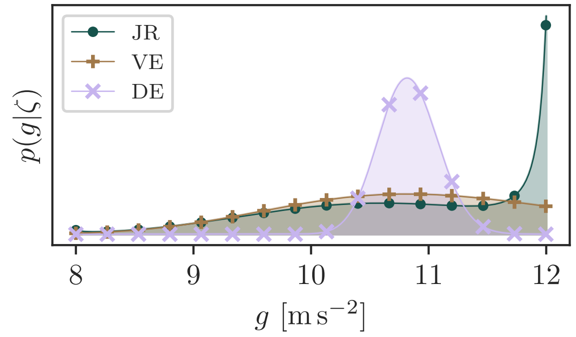

Figure 4. Posterior distributions over the gravitational acceleration, g, on the surface of Earth inferred by an experiment in which the time taken, $\hat{t} = 0.43\,\mathrm{s}$, for a ball to fall $1\,\mathrm{m}$ is measured. Given the uncertain evidence $q(\hat{t}|t) = \mathcal{N}(t, (0.025\,\mathrm{s})^2)$, notice that $g \simeq 9.81\,\mathrm{m\,s^{-2}}$ is well covered by the posteriors of Jeffrey's rule and virtual evidence but is excluded by distributional evidence. Recall, however, that both Jeffrey's rule and distributional evidence are *inappropriate* in this case and serve to illustrate what happens when the uncertain evidence is misinterpreted.

$p(x|\zeta)$. Using Equation (6) implies performing inference using Jeffrey's rule. To ensure Jeffrey's rule is consistent, Theorem 3.3, take on the perspective of an "oracle" and impose the restriction that all marginal and conditionals are Gaussians. This leads to the joint also being Gaussian (e.g., Bishop, 2006, ch. 2.3). Theorem 3.3 (II) is trivially satisfied as $\boldsymbol{x} = x, \boldsymbol{y} = y$, and ζ are one-dimensional. Further, one finds a $p(\zeta|y)$ that satisfies Theorem 3.3 (I) by choosing $p(\zeta|y) = \mathcal{N}(\mu_{\zeta|y}, \sigma^2_{\zeta|y})$ such that $\mu_{\zeta|y} = (y\sigma^2_\zeta + \mu_x\sigma^2_q)/(\sigma^2_\zeta + \sigma^2_q)$ and $\sigma^2_{\zeta|y} = \sigma^2_\zeta\sigma^2_q/(\sigma^2_\zeta + \sigma^2_q)$ where $\sigma^2_\zeta = \sigma^2_x + \sigma^2_{y|x} - \sigma^2_q$. Specifically, the variance constraint $\sigma^2_\zeta \geq 0$ ensures that Theorem 3.3 (III) is satisfied since $\sigma^2_\zeta \geq 0 \Rightarrow \sigma^2_y = \sigma^2_x + \sigma^2_{y|x} \geq \sigma^2_q = \mathbb{E}\left[\mathrm{Var}\left[y|\zeta\right]\right]$. See Figure 2 for the values used in the experiment.

When comparing Jeffrey's rule to virtual and distributional evidence, fix the base model $p(x, y)$ and the density $q(y|\zeta)$ but vary the interpretation of the distributional evidence. In particular, since $q(y|\zeta)$ is symmetric in y and ζ, take for virtual evidence $q_{\mathrm{V}}(\zeta|y) = \mathcal{N}(y, \sigma^2_{q_\zeta})$. In the case of distributional evidence, analytically solve for $p(Y \sim \mathbb{Q}|x)$ as well as the adjusted prior $p_{\mathrm{a}}(x)$. This is achieved by assuming that the density $p(Y \sim \mathbb{Q}|x)$ normalizes w.r.t. the mean of $q(y)$ being the density of a normal distribution with constant standard deviation σ_{q_ζ}, see Appendix C.1. In all cases, the posteriors $p(x|\zeta)$ are Gaussian, with the different posteriors shown in Figure 2. Note how in Figure 2, in the left panel the three methods result in vastly different posteriors, whereas those in the right panel are indistinguishable. This emphasizes the importance of carefully choosing the approach in dealing with uncertain evidence.

4.2. The Drop of a Ball

Consider the classic "high school" experiment in which a student attempts to measure gravitational acceleration, g, by timing, t, how long it takes for a ball to fall a distance, x. Armed with the formula $x = gt^2/2$ the student can convert measurements of t into estimates of g if x is known. In this setup $x = 1\,\mathrm{m}$ and is measured a priori, a "model" error of $0.005\,\mathrm{s}$ is assumed to account for physics ignored by the formula (e.g., air resistance). The student then attempts to infer g from a single experiment, during which they observe a time on the stopwatch[5] of $0.43\,\mathrm{s}$. Suppose it is asserted that given the true time it took the ball to hit the ground, the student is equally likely to be too eager or too slow in regards to hitting the stop bottom, but that the probability that the student is increasingly slow/eager vanishes. This uncertain evidence is of type (III), virtual evidence, leading to[6] $p(\hat{t}|t) = \mathcal{N}(t, (0.025\,\mathrm{s})^2)$ as a measurement error of $0.025\,\mathrm{s}$ is assumed. The graphical model is shown in Figure 3.

For Jeffrey's rule, the density is "flipped" as it is symmetric in its mean and random variable, such that $q(t|\hat{t}) = \mathcal{N}(0.43\,\mathrm{s}, (0.025\,\mathrm{s})^2)$. For distributional evidence notice that the form of $q(Y \sim \mathbb{Q}|g)$ (Y being the time variable) is the same as in Section 4.1 which allows for an analytical likelihood. In the case of Jeffrey's rule, Theorem 3.3 (II) is trivially satisfied as g, t, and $\hat{t}$ are one-dimensional. Theorem 3.3 (III) is also satisfied since $\mathrm{Var}[t] = \mathbb{E}\left[\mathrm{Var}[t|g]\right] + \mathrm{Var}[\mathbb{E}[t|g]] = 0.005^2 + \mathbb{E}[2/g] - \left(\mathbb{E}[\sqrt{2/g}]\right)^2 = 0.0007 \geq 0.025^2 = \mathbb{E}[\mathrm{Var}[t|\hat{t}]]$. Also note that different to the experiment in Section 4.1 posteriors here are not exactly calculated. Rather, they are

[5] A perfect experiment would record $\simeq 0.45\,\mathrm{s}$ for the terrestrial $g \simeq 9.81\,\mathrm{m\,s^{-2}}$.

[6] Since it is assumed that virtual evidence is correct, p rather than q is used to denote this density.

approximated via numerical integration, as the latent space is low dimensional. Figure 4 shows the posteriors using the three different interpretations of the given uncertain evidence. Each posterior is different with particularly distributional evidence exhibiting a (comparatively) small variance. This results in near zero probability on the true value of the gravitational acceleration at $g \simeq 9.81\,\mathrm{m\,s^{-2}}$.

This again exemplifies the potential error one might make when a certain type of uncertain evidence is misinterpreted. Particularly, distributional evidence should not be expected to produce reasonable results in this case. Recall that both Jeffrey's rule and distributional evidence are inappropriate by construction. To make, for example, distributional evidence the correct interpretation, this example can be modified as follows: suppose the student's setup is somewhat shaky, leading to x varying slightly. The student instead uses a very accurate time measurement device; the measurements of the time are *exact*. The student may then carry out repeated measurements in this *single* experiment and conclude that the measured time t is distributed as $q(t|g) = \mathcal{N}(\mu_t, \sigma_t^2)$. In this case the uncertain evidence is of type (II) since the uncertainty is conditioned on the latent variable g.

4.3. Planet Orbiting Kepler 90

The Kepler satellite (Borucki et al., 2010) measured the flux from over half a million stars over 5 years. Dips in the observed flux can occur when a planet transits in front of the stellar disk, and accurate measurements of the exact transit times allow one to infer the orbital properties of the planets. However, the received flux from distant stars varies for other reasons (e.g., stellar pulsations, telescope temperature) and in principle one should fit a joint orbit/stellar/telescope model to the observed flux to infer orbital parameters. However, it is common (e.g. Liang et al., 2021) to extract this information in two phases: first to fit a model of the star and to extract from this the *transit times* and second to extract orbital parameters from these transit times. Thus, the measurements of transit times constitute uncertain evidence, in that they are provided as estimated times with an associated error (type (I) uncertain evidence). In the case of a single planet, it is only possible to infer the orbital period, P, and the anomaly angle $\omega + M$, while the other (planar) orbital parameters, eccentricity, e, and periapsis argument, ω, remain marginally unconstrained (but not in correlation with P and $\omega + M$). Data is simulated based on Kepler-90g, with $P = 210$ days, $e = 0.05$, $\omega = 100$ deg and $\omega + M = 198$ deg using TTVFAST (Deck et al., 2014). Approximate posterior inference is carried out with PYPROB (Baydin & Le, 2018). The prior over P is taken to be normal with 210 ± 1 days. The prior over eccentricity is taken to be uniform between 0 and 0.15. The angular variables have uniform priors between 0 and 360 deg. In Figure 5 additional experimental details is provided and the figure shows the 1D and 2D marginal posterior distributions over orbital parameters given the three different approaches to uncertain evidence. Note that while the marginal posterior distributions of e, ω and $\omega + M$ are all in agreement, the posterior over P is significantly different when extracted using Jeffrey's rule compared to when using the other two methods.

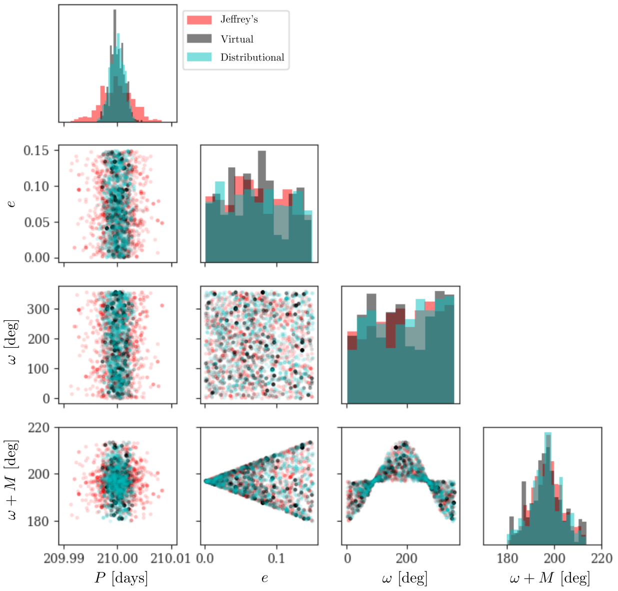

Figure 5. Inferred orbital parameters of an exoplanet around a Kepler star. 7 transits, $\boldsymbol{t}$, of the system is simulated, and an error (standard deviation) on the measured transits of 20 mins ($q(\boldsymbol{t}|\zeta)$) is assumed. Additionally, distrust in the model leads to the likelihood error (standard deviation) on the transits of 10 mins. While the marginal posterior distributions of e, ω and $\omega + M$ are all in agreement, the posterior over P is significantly different when extracted using Jeffrey's rule compared to the other two methods.

5. Related Work

Of important related work is that by Valtorta et al. (2002) who propose an approach for dealing with uncertain evidence which is in some way an extension to Jeffrey's rule. Their algorithm, the *soft evidential update method*, is tailored for Bayesian networks (BN) and they incorporate uncertain evidence by extending the BN with evidence nodes for each new piece of uncertain evidence. Their approach updates the prior BN (prior to receiving uncertain evidence), denoted M_P, by solving for a new "updated" BN, M_U. The resulting M_U minimizes the Kullback-Leibler divergence between M_P and M_U under the constraint that the marginal distribution of M_U of each uncertain evidence variable must equal the given distributions. Given a single piece of uncertain evidence their update method reduces to Jeffrey's rule. Another approach is that of Yao (2022), which is similar to, and discussed by Tolpin et al. (2021). The difference of this approach compared to distributional evidence lies in the definition of the likelihood $p(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x})$, for which Yao (2022) proposes $p(\boldsymbol{Y} \sim \mathbb{Q}|\boldsymbol{x}) \propto \int p(\boldsymbol{y}|\boldsymbol{x})q(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$. However, as discussed by Tolpin et al. (2021), this definition lacks many (what they deem) desired properties associated with distributional evidence, Equation (5).

Finally, this work should be compared to another popular perspective taken when dealing with uncertain evidence. This perspective considers the dichotomy between the concepts of focusing and revision (e.g., Smets, 1993; Chan & Darwiche, 2005). Informally, focusing can be thought of as standard conditioning, while revision, as the name implies, refers to revising the model altogether. In relation to this work, one can discuss revision and focusing in regards to the notion of consistency and the auxiliary variable ζ. In the case of virtual evidence and Jeffrey's Rule, consistency leads to focusing, as inference conditioned on ζ is consistent with the base model. On the other hand, inconsistency implies either (1) the uncertainty about the observation is erroneous and arguably should be re-evaluated. (2) The base model is mis-specified, in which case one may update the model from first principles. (3) Treat Jeffrey's Rule as the model revision mechanism. Distributional evidence is arguably a revision by definition despite it being consistent.

6. Conclusions

We have considered the problem of Bayesian inference when given uncertain evidence and the importance of its proper interpretation. This involved discussing and provided new insights into three different approaches in dealing with uncertain evidence: Jeffrey's rule, virtual evidence, and distributional evidence. Particularly, this lead to the definition of four types of commonly encountered uncertain evidence. We have discussed compatibility between a given probabilistic model and uncertain evidence as defined in terms of consistency. We have demonstrated in three different experiments how misinterpretations of the type of uncertain evidence may lead to different inference results. This illustrates the importance of carefully making the proper interpretation of uncertain evidence on a case-by-case basis.

Acknowledgements

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and the Intel Parallel Computing Centers program. Additional support was provided by UBC's Composites Research Network (CRN), Data Science Institute (DSI), and Lawrence Berkley National Lab (under subcontract 7623401). This research was enabled in part by technical support and computational resources provided by WestGrid (www.westgrid.ca), Compute Canada (www.computecanada.ca), and Advanced Research Computing at the University of British Columbia (arc.ubc.ca).

References

Baydin, A. G. and Le, T. A. *pyprob*, 2018. URL https://github.com/probprog/pyprob.

Baydin, A. G., Shao, L., Bhimji, W., Heinrich, L., Naderiparizi, S., Munk, A., Liu, J., Gram-Hansen, B., Louppe, G., Meadows, L., Torr, P., Lee, V., Cranmer, K., Prabhat, Mr., and Wood, F. Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6d19c113404cee55b4036fce1a37c058-Abstract.html.

Ben Mrad, A., Delcroix, V., Piechowiak, S., Maalej, M. A., and Abid, M. Understanding soft evidence as probabilistic evidence: Illustration with several use cases. In *2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO)*, pp. 1–6, April 2013. doi: 10.1109/ICMSAO.2013.6552583.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer,, 2006. ISBN 0-387-31073-8 978-0-387-31073-2.

Borucki, W. J., Koch, D., Basri, G., Batalha, N., Brown, T., Caldwell, D., Caldwell, J., Christensen-Dalsgaard, J., Cochran, W. D., DeVore, E., Dunham, E. W., Dupree, A. K., Gautier, T. N., Geary, J. C., Gilliland, R., Gould, A., Howell, S. B., Jenkins, J. M., Kondo, Y., Latham, D. W., Marcy, G. W., Meibom, S., Kjeldsen, H., Lissauer, J. J., Monet, D. G., Morrison, D., Sasselov, D., Tarter, J., Boss, A., Brownlee, D., Owen, T., Buzasi, D., Charbonneau, D., Doyle, L., Fortney, J., Ford, E. B., Holman, M. J., Seager, S., Steffen, J. H., Welsh, W. F., Rowe, J., Anderson, H., Buchhave, L., Ciardi, D., Walkowicz, L., Sherry, W., Horch, E., Isaacson, H., Everett, M. E., Fischer, D., Torres, G., Johnson, J. A., Endl, M., MacQueen, P., Bryson, S. T., Dotson, J., Haas, M., Kolodziejczak, J., Van Cleve, J., Chandrasekaran, H., Twicken, J. D., Quintana, E. V., Clarke, B. D., Allen, C., Li, J., Wu, H., Tenenbaum, P., Verner, E., Bruhweiler, F., Barnes, J., and Prsa, A. Kepler planet-detection mission: Introduction and first results. *Science (New York, N.Y.)*, 327(5968): 977, February 2010. doi: 10.1126/science.1185402.

Chan, H. and Darwiche, A. On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence*, 163(1):67–90, 2005.

Dax, M., Green, S. R., Gair, J., Macke, J. H., Buonanno, A., and Schölkopf, B. Real-time gravitational wave science with neural posterior estimation. 127(24):241103, December 2021. doi: 10.1103/PhysRevLett.127.241103.

Deck, K. M., Agol, E., Holman, M. J., and Nesvorný, D. TTVFast: An efficient and accurate code for transit timing inversion problems. 787(2):132, June 2014. doi: 10.1088/0004-637X/787/2/132.

Diaconis, P. and Zabell, S. L. Updating subjective probability. *Journal of the American Statistical Association*, 77 (380):822–830, 1982.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, September 1987. ISSN 0370-2693. doi: 10.1016/0370-2693(87)91197-X. URL https://www.sciencedirect.com/science/article/pii/037026938791197X.

Feroz, F. and Hobson, M. P. Bayesian analysis of radial velocity data of GJ667C with correlated noise: Evidence for only two planets. 437(4):3540–3549, February 2014. doi: 10.1093/mnras/stt2148.

Gershman, S. and Goodman, N. Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

Grove, A. J. and Halpern, J. Y. Probability update: Conditioning vs. cross-entropy. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI'97, pp. 208–214, San Francisco, CA, USA, August 1997. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-485-8.

Hammersley, J. M. and Handscomb, D. C. *Monte Carlo Methods*. Springer Netherlands, Dordrecht, 1964. ISBN 978-94-009-5821-0 978-94-009-5819-7. doi: 10.1007/978-94-009-5819-7. URL http://link.springer.com/10.1007/978-94-009-5819-7.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109, April 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. URL https://doi.org/10.1093/biomet/57.1.97.

Jacobs, B. The Mathematics of Changing One's Mind, via Jeffrey's or via Pearl's Update Rule. *Journal of Artificial Intelligence Research*, 65:783–806, August 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11349. URL https://www.jair.org/index.php/jair/article/view/11349.

Jeffrey, R. C. *The Logic of Decision*. University of Chicago press, 2nd (1983) edition, 1965.

Lavin, A., Zenil, H., Paige, B., Krakauer, D., Gottschlich, J., Mattson, T., Anandkumar, A., Choudry, S., Rocki, K., Baydin, A. G., et al. Simulation intelligence: Towards a new generation of scientific methods. *arXiv preprint arXiv:2112.03235*, 2021.

Lentati, L., Hobson, M. P., and Alexander, P. Bayesian estimation of non-Gaussianity in pulsar timing analysis. 444(4):3863–3878, November 2014. doi: 10.1093/mnras/stu1721.

Liang, Y., Robnik, J., and Seljak, U. Kepler-90: Giant Transit-timing Variations Reveal a Super-puff. *The Astronomical Journal*, 161(4):202, March 2021. ISSN 1538-3881. doi: 10.3847/1538-3881/abe6a7. URL https://doi.org/10.3847/1538-3881/abe6a7.

Metropolis, N. and Ulam, S. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335–341, September 1949. ISSN 0162-1459. doi: 10.1080/01621459.1949.10483310. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483310.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 0021-9606. doi: 10.1063/1.1699114. URL https://aip.scitation.org/doi/abs/10.1063/1.1699114.

Mishra-Sharma, S. and Cranmer, K. Neural simulation-based inference approach for characterizing the Galactic Center γ-ray excess. *Physical Review D: Particles and Fields*, 105(6):063017, March 2022. doi: 10.1103/PhysRevD.105.063017. URL https://link.aps.org/doi/10.1103/PhysRevD.105.063017.

Mrad, A. B., Delcroix, V., Piechowiak, S., Leicester, P., and Abid, M. An explication of uncertain evidence in Bayesian networks: Likelihood evidence and probabilistic evidence. *Applied Intelligence*, 43(4): 802–824, December 2015. ISSN 1573-7497. doi: 10.1007/s10489-015-0678-6. URL https://doi.org/10.1007/s10489-015-0678-6.

Munk, A., Zwartsenberg, B., Scibior, A., Baydin, A. G., Stewart, A. L., Fernlund, G., Poursartip, A., and Wood, F. Probabilistic surrogate networks for simulators with unbounded randomness. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Neal, R. M. An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics*, 111(1):194–203, 1994. ISSN 0021-9991. doi: 10.1006/jcph.1994.1054. URL https://www.sciencedirect.com/science/article/pii/S0021999184710540.

Paksoy, V., Turkmen, R., and Zhang, F. Inequalities of generalized matrix functions via tensor products. *The Electronic Journal of Linear Algebra*, 27:332–341, 2014.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 837–848. PMLR, April 2019. URL https://proceedings.mlr.press/v89/papamakarios19a.html.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan kaufmann, 1988.

Pearl, J. On Two Pseudo-Paradoxes in Bayesian Analysis. *Annals of Mathematics and Artificial Intelligence*, 32 (1):171–177, August 2001. ISSN 1573-7470. doi: 10.1023/A:1016709416174. URL https://doi.org/10.1023/A:1016709416174.

Peng, Y., Zhang, S., and Pan, R. Bayesian Network Reasoning with Uncertain Evidences. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 18:539–564, October 2010. doi: 10.1142/S0218488510006696.

Riess, A. G., Yuan, W., Macri, L. M., Scolnic, D., Brout, D., Casertano, S., Jones, D. O., Murakami, Y., Anand, G. S., Breuval, L., Brink, T. G., Filippenko, A. V., Hoffmann, S., Jha, S. W., D'arcy Kenworthy, W., Mackenty, J., Stahl, B. E., and Zheng, W. A comprehensive measurement of the local value of the hubble constant with 1 km s^{-1} mpc^{-1} uncertainty from the hubble space telescope and the SH0ES team. 934(1):L7, July 2022. doi: 10.3847/2041-8213/ac5c5b.

Schulze-Hartung, T., Launhardt, R., and Henning, T. Bayesian analysis of exoplanet and binary orbits. Demonstrated using astrometric and radial-velocity data of ¡ASTROBJ¿Mizar A¡/ASTROBJ¿. 545:A79, September 2012. doi: 10.1051/0004-6361/201219074.

Shafer, G. Jeffrey's rule of conditioning. *Philosophy of Science*, 48(3):337–362, 1981.

Smets, P. Jeffrey's rule of conditioning generalized to belief functions. In *Uncertainty in Artificial Intelligence*, pp. 500–505, 1993.

Thrane, E. and Talbot, C. An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models. 36: e010, March 2019. doi: 10.1017/pasa.2019.2.

Tolpin, D., Zhou, Y., Rainforth, T., and Yang, H. Probabilistic Programs with Stochastic Conditioning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10312–10323. PMLR, July 2021. URL https://proceedings.mlr.press/v139/tolpin21a.html.

Valtorta, M., Kim, Y.-G., and Vomlel, J. Soft evidential update for probabilistic multiagent systems. *International Journal of Approximate Reasoning*, 29(1):71–106, January 2002. ISSN 0888-613X. doi: 10.1016/S0888-613X(01)00056-1. URL https://www.sciencedirect.com/science/article/pii/S0888613X01000561.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26, January 2021. ISSN 2662-8449. doi: 10.1038/s43586-020-00001-2. URL https://www.nature.com/articles/s43586-020-00001-2.

Vigeland, S. J. and Vallisneri, M. Bayesian inference for pulsar-timing models. 440(2):1446–1457, May 2014. doi: 10.1093/mnras/stu312.

Wagner, C. G. Probability Kinematics and Commutativity. *Philosophy of Science*, 69(2):266–278, June 2002. ISSN 0031-8248, 1539-767X. doi: 10.1086/341053. URL https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/probability-kinematics-and-commutativity/0F4936CD60FE722EC010F4001AEE75D2.

Wood, F., Warrington, A., Naderiparizi, S., Weilbach, C., Masrani, V., Harvey, W., Ścibior, A., Beronov, B., Grefenstette, J., Campbell, D., and Nasseri, S. A. Planning as Inference in Epidemiological Dynamics Models. *Frontiers in Artificial Intelligence*, 4, 2022. ISSN 2624-8212. URL https://www.frontiersin.org/articles/10.3389/frai.2021.550603.

Yao, K. Bayesian inference with uncertain data of imprecise observations. *Communications in Statistics - Theory and Methods*, 51(15):5330–5341, 2022. doi: 10.1080/03610926.2020.1838545. URL https://doi.org/10.1080/03610926.2020.1838545.

A. Commutativity of Jeffrey's Rule

It is well known (Diaconis & Zabell, 1982) that Jeffrey's rule does not *generally* commute with respect to different pieces of uncertain evidence, ϵ_A, ϵ_B. That is, applying Jeffrey's rule first with respect to ϵ_A, and then subsequently with respect to ϵ_B, is *not* necessarily equal to applying Jeffrey's rule in the reverse order. This is easily seen with the following example: Let ϵ_A and ϵ_B carry contradictory information about the same variable $\boldsymbol{y}$. For each piece of uncertain evidence, consider the associated auxiliary variable ζ_A and ζ_B and the densities $q(\boldsymbol{y}|\zeta_A)$ and $q(\boldsymbol{y}|\zeta_B)$. Then from Jeffrey's rule the updated density of the latent variable $\boldsymbol{x}$ is (depending on the the order of applied uncertain evidence):

$$
\begin{aligned}
p(\boldsymbol{x}|\zeta_A, \zeta_B) &= \int p(\boldsymbol{x}|\boldsymbol{y}, \zeta_A) q(\boldsymbol{y}|\zeta_B) \, \mathrm{d}\boldsymbol{y} \overset{\overset{\text{By definition of applying Jeffrey's rule}}{\curvearrowright}}{=} \int p(\boldsymbol{x}|\boldsymbol{y}) q(\boldsymbol{y}|\zeta_B) \, \mathrm{d}\boldsymbol{y} = p(\boldsymbol{x}|\zeta_B) \\
p(\boldsymbol{x}|\zeta_B, \zeta_A) &= \int p(\boldsymbol{x}|\boldsymbol{y}, \zeta_B) q(\boldsymbol{y}|\zeta_A) \, \mathrm{d}\boldsymbol{y} = \int p(\boldsymbol{x}|\boldsymbol{y}) q(\boldsymbol{y}|\zeta_A) \, \mathrm{d}\boldsymbol{y} = p(\boldsymbol{x}|\zeta_A),
\end{aligned}
$$

where we use $p(\cdot|\zeta_1, \zeta_2)$ as an overloaded denotation for applying Jeffrey's rule first with respect to ζ_1 and subsequently with respect to ζ_2. In this example, we see that the second piece of uncertain evidence dominates and "overwrites" or "forgets" the first. This illustrates that if two pieces of "incompatible" uncertain evidence are given, care must be taken when using Jeffrey's rule. We leave the topic of addressing commutativity of Jeffrey's rule for future discussion, but we briefly mention that a potential remedy could be to define a mixture of $q(\boldsymbol{y}|\zeta_A)$ and $q(\boldsymbol{y}|\zeta_B)$, which would require incorporating ϵ_A and ϵ_B jointly rather than sequentially.

As a final note, we point out the likelihood-bases approaches to uncertain evidence, such as virtual evidence, does commute with respect to multiple pieces of uncertain evidence. In particular, given two incompatible pieces of uncertain evidence and associated auxiliary variables ζ_A and ζ_B, the joint density would assign zero probability on that event, $p(\zeta_A, \zeta_B, \boldsymbol{y}, \boldsymbol{x}) = 0$, which in turn may indicate a mis-specification of the model.

B. Proofs

Proof of Equation (3). Consider the assumptions in Definition 2.2 and let $\boldsymbol{y} \in \{\boldsymbol{y}_k\}_{k=1}^K$ be discrete. From Equation (2) it follows that $p(\zeta|\boldsymbol{y}_k) = c\lambda_k$ with $k = 1, \ldots, K$ for some $c \in \mathbb{R}_+$. This leads to,

$$
\begin{aligned}
p(\boldsymbol{x}|\zeta) &= \frac{\sum_{k=1}^K p(\boldsymbol{x}, \boldsymbol{y}_k, \zeta)}{p(\zeta)} = \frac{\sum_{k=1}^K p(\zeta|\boldsymbol{y}_k) p(\boldsymbol{x}, \boldsymbol{y}_k)}{\sum_{j=1}^K p(\zeta, \boldsymbol{y}_j)} \\
&= \frac{\sum_{k=1}^K p(\zeta|\boldsymbol{y}_k) p(\boldsymbol{x}, \boldsymbol{y}_k)}{\sum_{j=1}^K p(\zeta|\boldsymbol{y}_j) p(\boldsymbol{y}_j)} = \frac{\sum_{k=1}^K c\lambda_k p(\boldsymbol{x}, \boldsymbol{y}_k)}{\sum_{j=1}^K c\lambda_j p(\boldsymbol{y}_j)} \\
&= \frac{c}{c} \frac{\sum_{k=1}^K \lambda_k p(\boldsymbol{x}, \boldsymbol{y}_k)}{\sum_{j=1}^K \lambda_j p(\boldsymbol{y}_j)} = \frac{\sum_{k=1}^K \lambda_k p(\boldsymbol{x}, \boldsymbol{y}_k)}{\sum_{j=1}^K \lambda_j p(\boldsymbol{y}_j)},
\end{aligned}
$$

which concludes the proof. □

B.1. Proofs for Theorem 3.3

Proof of Theorem 3.3 (1). (Necessary) Given $p(\boldsymbol{y}, \boldsymbol{x})$ and uncertain evidence $q(\boldsymbol{y}|\zeta)$ it needs to be to be shown that if the approach of Jeffrey's rule is consistent, then Theorem 3.3 (1) is true. Consistency requires that there exists a joint $p(\zeta, \boldsymbol{y}, \boldsymbol{x}) = p(\zeta|\boldsymbol{y}, \boldsymbol{x}) p(\boldsymbol{y}, \boldsymbol{x})$ "containing" $q(\boldsymbol{y}|\zeta)$. This implies finding a $p(\zeta|\boldsymbol{y}, \boldsymbol{x})$ such that for all ζ and $\boldsymbol{y}$ it holds that:

$$
\begin{aligned}
q(\boldsymbol{y}|\zeta) &= \frac{\int p(\zeta|\boldsymbol{y}, \boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}{\int p(\zeta|\boldsymbol{y}, \boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}} \\
&= \frac{p(\zeta|\boldsymbol{y}) p(\boldsymbol{y})}{\int p(\zeta|\boldsymbol{y}) p(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}},
\end{aligned}
$$

where $p(\zeta|\boldsymbol{y}) = \int p(\zeta|\boldsymbol{y}, \boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}) / p(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{x}$. That is, if no such $p(\zeta|\boldsymbol{y})$ exists then the approach of Jeffrey's rule cannot be consistent.

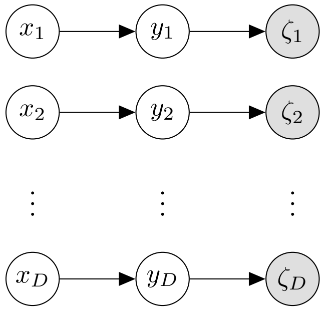

Figure 6.

(Sufficient) Assume there exists $p(\zeta|\boldsymbol{y})$ satisfying Theorem 3.3 (1). Define $p(\zeta, \boldsymbol{y}, \boldsymbol{x}) = p(\zeta|\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$ from which it immediately follows that $p(\boldsymbol{y}|\zeta) = q(\boldsymbol{y}|\zeta)$. Further, using d-separation (Pearl, 1988), it follows that defining $p(\zeta, \boldsymbol{y}, \boldsymbol{x})$ in this way ensures that $p(\boldsymbol{x}|\boldsymbol{y}, \zeta) = p(\boldsymbol{x}|\boldsymbol{y})$. This shows that Jeffrey's rule is consistent and that:

$$\begin{aligned} p(\boldsymbol{x}|\zeta) &= \int p(\boldsymbol{x}|\boldsymbol{y}, \zeta)p(\boldsymbol{y}|\zeta)\,\mathrm{d}\boldsymbol{y} \\ &= \int p(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{y}|\zeta)\,\mathrm{d}\boldsymbol{y}, \end{aligned}$$

thereby concluding the proof. □

Proof of Theorem 3.3 (2). Let each $\{y_i\}_{i=1}^d$ be conditionally independent given ζ such that $q(\boldsymbol{y}|\zeta) = \prod_{i=1}^d q(y_i|\zeta)$. Further, assume Jeffrey's rule is consistent such that there exists a joint model $p(\zeta, \boldsymbol{y}, \boldsymbol{x})$ where $p(\boldsymbol{y}|\zeta) = q(\boldsymbol{y}|\zeta)$. Then it follows, via d-separation, that if and only if all paths between each $\{y_i\}_{i=1}^d$ are conditionally blocked can they be conditionally independent given ζ. This implies that no two or more y_i can share the same auxiliary variable, latent variable, or depend on each other. Figure 6 shows the only possible graphical model satisfying this constraint, which leads to:

$$p(\boldsymbol{y}|\zeta) = \prod_{i=1}^d p(y_i|\zeta_i) \Rightarrow p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^d p(y_i|x_i).$$

This concludes the proof. □

Proof of Theorem 3.3 (3). Assume Jeffrey's rule is consistent such that $q(\boldsymbol{y}|\zeta) = p(\boldsymbol{y}|\zeta)$ which implies $p(\zeta) = \int p(\zeta|\boldsymbol{y})p(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$. From the law of total variance it follows that:

$$\mathrm{Cov}\left[\boldsymbol{Y}\right] = \mathbb{E}\left[\mathrm{Cov}\left[\boldsymbol{Y}|\zeta\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\boldsymbol{Y}|\zeta\right]\right], \tag{8}$$

with the right-hand side being a sum of two positive semi-definite matrices. Since for two positive semi-definite matrices A and B it holds that $\det(A + B) \geq \det(A) + \det(B)$ (Paksoy et al., 2014), and as $\det(A), \det(B) \geq 0$ this leads to:

$$\mathrm{Cov}\left[\boldsymbol{Y}\right] \succeq \mathbb{E}\left[\mathrm{Cov}\left[\boldsymbol{Y}|\zeta\right]\right].$$

Further, the elements in the diagonal of the left-hand side of Equation (8) are the variances of the constituent parts of $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$. Therefore:

$$\begin{aligned} \mathrm{Var}\left[Y_i\right] &= \mathbb{E}\left[\mathrm{Var}\left[Y_i|\zeta\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[Y_i|\zeta\right]\right] \\ &\geq \mathbb{E}\left[\mathrm{Var}\left[Y_i|\zeta\right]\right], \text{ since } \mathrm{Var}\left[\mathbb{E}\left[Y_i|\zeta\right]\right] \geq 0. \end{aligned} \tag{9}$$

Finally it is proven that:

$$\mathrm{Var}\left[Y_i\right] = \mathbb{E}\left[\mathrm{Var}\left[Y_i|\zeta\right]\right] \Leftrightarrow \mathbb{E}\left[Y_i|\zeta\right] = c,$$

with $c \in \mathbb{R}$ being a constant. First prove "⇒":

$$\mathrm{Var}\left[Y_i\right] = \mathbb{E}\left[\mathrm{Var}\left[Y_i|\zeta\right]\right] \Rightarrow \mathrm{Var}\left[\mathbb{E}\left[Y_i|\zeta\right]\right] = 0.$$

As $\mathrm{Var}\left[Y\right] = \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\right]\right)^2\right]$ is an expectation of a non-negative variable, it follows that $\mathrm{Var}\left[Y\right] = 0$ if and only if Y is constant. Therefore, it follows that:

$$\mathrm{Var}\left[\mathbb{E}\left[Y_i|\zeta\right]\right] = 0 \Leftrightarrow \mathbb{E}\left[Y_i|\zeta\right] = c, \tag{10}$$

where $c \in \mathbb{R}$ is some constant.

Next "⇐" is proven. This follows trivially by combining Equation (10) with Equation (9):

$$\mathrm{Var}\left[\mathbb{E}\left[Y_i|\zeta\right]\right] = 0 \Rightarrow \mathrm{Var}\left[Y_i\right] = \mathbb{E}\left[\mathrm{Var}\left[Y_i|\zeta\right]\right].$$

From this it can be concluded that:

$$\mathrm{Var}\left[Y_i\right] = \mathbb{E}\left[\mathrm{Var}\left[Y_i|\zeta\right]\right] \Leftrightarrow \mathbb{E}\left[Y_i|\zeta\right] = c,$$

as desired, thereby concluding the proof. □

C. Other Derivations

C.1. Distributional Evidence and Normal Distributions

Consider the densities $p(y|x) = \mathcal{N}(\mu_{y|x}|\sigma^2_{y|x})$ and $q(y) = \mathcal{N}(\mu_q|\sigma^2_q)$ (with μ_q being the parameter in which q normalizes) and the distributional evidence likelihood, Equation (5):

$$\begin{aligned}
\ln p(Y \sim \mathbb{Q}_{\mu_q}|x) &\propto \int \ln p(y|x) q(y)\,\mathrm{d}y \\
&= -\frac{1}{2\sigma^2_{y|x}} \int \left[\left(y - \mu_{y|x}\right)^2 q(y)\,\mathrm{d}y\right] - \ln\left(\sqrt{2\pi}\,\sigma_{y|x}\right) \\
&= -\frac{1}{2\sigma^2_{y|x}} \left[\int y^2 q(y)\,\mathrm{d}y - 2\mu_q\mu^2_{y|x} + \mu^2_{y|x}\right] - \ln\left(\sqrt{2\pi}\,\sigma_{y|x}\right) \\
&= -\frac{1}{2\sigma^2_{y|x}} \left[\mu_q^2 - 2\mu_q\mu^2_{y|x} + \mu^2_{y|x}\right] - \left(\ln\left(\sqrt{2\pi}\,\sigma_{y|x}\right) + \frac{\sigma^2_q}{2\sigma_{y|x}}\right) \\
&= -\frac{1}{2\sigma^2_{y|x}} \left(\mu_q - \mu_{y|x}\right)^2 - \left(\ln\left(\sqrt{2\pi}\,\sigma_{y|x}\right) + \frac{\sigma^2_q}{2\sigma_{y|x}}\right).
\end{aligned}$$

Assuming the distribution $\mathbb{Q}_{\mu_q}$ is implicitly defined via its mean μ_q, such that $p(Y \sim \mathbb{Q}_{\mu_q})$ normalizes with respect to μ_q it is identified to be a Gaussian with mean $\mu_{y|x}$ and variance $\sigma^2_{y|x}$. Therefore, the normalization constant $Z(x)$, generally a function of x, is found to be:

$$Z(x) = \int p(Y \sim \mathbb{Q}_{\mu_q}|x)(\mu_q)\,\mathrm{d}\mu_q = e^{-\frac{\sigma^2_q}{2\sigma_{y|x}}}.$$

Thus, when $\sigma_{y|x}$ is a function of x, this leads to the following distributional evidence adjusted prior $p_{\mathrm{a}}(x) \propto p(x)Z(x)$:

$$p_{\mathrm{a}}(x) = \frac{p(x)Z(x)}{\int p(x)Z(x)\mathrm{d}x}.$$

Further, and importantly, this reveals that if the variance $\sigma_{y|x}$ of the model likelihood is independent of x, then so too is $Z(x)$. In this case $Z(x) = Z$, which is also the case in all experiments in Section 4, and it can be concluded that distributional evidence adjusted prior normalizes trivially:

$$p_{\mathrm{a}}(x) = \frac{p(x)Z}{\int p(x)Z\mathrm{d}x} = p(x).$$

This leads to distributional evidence in this special case reducing to the usual exact evidence with respect to the given base model $p(x, y)$ with the observation $y = \mu_q$.

These results generalize to the multivariate case where the likelihood in the base model and the distributional evidence density are defined on an observable vector $\boldsymbol{y} = (y_1, \ldots, y_K)$. That is, each y_k is i.i.d. such that $p(\boldsymbol{y}|x) = \prod_{k=1}^{K} p(y_k|x)$ and $q(\boldsymbol{y}|x) = \prod_{k=1}^{K} q(y_k|x)$.