A Model of Prior Ignorance for Inferences in the One-parameter Exponential Family

Alessio Benavoli and Marco Zaffalon

IDSIA, Galleria 2, CH-6928 Manno (Lugano), Switzerland
alessio@idsia.ch, zaffalon@idsia.ch

Abstract

This paper proposes a model of prior ignorance about a scalar variable based on a set of distributions $\mathcal{M}$. In particular, a set of *minimal properties* that a set $\mathcal{M}$ of distributions should satisfy to be a model of prior ignorance without producing vacuous inferences is defined. In the case the likelihood model corresponds to a one-parameter exponential family of distributions, it is shown that the above minimal properties are equivalent to a special choice of the domains for the parameters of the conjugate exponential prior. This makes it possible to define the largest (that is, the least-committal) set of conjugate priors $\mathcal{M}$ that satisfies the above properties. The obtained set $\mathcal{M}$ is a model of prior ignorance with respect to the functions (queries) that are commonly used for statistical inferences; it is easy to elicit and, because of conjugacy, tractable; it encompasses frequentist and the so-called objective Bayesian inferences with improper priors. An application of the model to a problem of inference with count data is presented.

Keywords: Prior ignorance, exponential families, lower and upper expectations.

1. Introduction

Scientific experimental results generally consist of sets of data $\{y_1, \ldots, y_n\}$. Statistical methods are then typically used to derive conclusion about both the nature of the process which has produced those observations, and about the expected behaviour at future instances of the same process. A central element of most statistical analysis is the specification of a likelihood function. This is assumed to describe the mechanism that has generated the

observations as a function of a parameter $w \in \mathbb{R}$ (in the following we assume that w is scalar), the so-called state-of-nature, about which only limited information (if any) is available. In the Bayesian paradigm, this information is modelled by means of a probability distribution (a prior), where the probabilistic model over w is not a description of the variability of w (parameters are typically fixed unknown quantities) but a description of the uncertainty about their values before observing the data.

An important problem in Bayesian analysis is how to define the prior distribution. If any prior information about the parameter w is available, it should be incorporated in the prior distribution. On the other hand, in the case (almost) no prior information is available about w, the prior should be selected so as to reflect such state of ignorance. The search for a prior distribution representing ignorance constitutes a fascinating chapter in the history of Bayesian statistics [1]. There are two main avenues to represent ignorance.

The first assumes that ignorance can be modelled satisfactorily by a single so-called noninformative prior density such as for instance Laplace's prior, Jeffreys' prior, or the reference prior of Bernardo (see [1, Sec. 5.6.2] for a review). This view has been questioned on diverse grounds. Noninformative priors are typically improper and may lead to an improper posterior. Moreover, even if the posterior is proper, it can be inconsistent with the likelihood model (i.e., incoherent in the subjective interpretation of probability [2, Ch. 7], as will be shown with an example in Section 2). In our view, however, the most important criticism of noninformative priors is that they are not expressive enough to represent ignorance (this will become more precise later in this Introduction).

An alternative is to use a set of prior distributions, $\mathcal{M}$, rather than a single distribution, to model prior ignorance about statistical parameters. Each prior distribution in $\mathcal{M}$ is updated by Bayes' rule, producing a set of posterior distributions. In fact there are two distinct approaches of this kind, which have been compared by Walley [2]. The first approach, known as *Bayesian sensitivity analysis* or *Bayesian robustness* [3, 4], assumes that there is an ideal prior distribution π_0 which could, ideally, model prior uncertainty. It is assumed that we are unable to determine π_0 accurately because of limited time or resources. The criterion for including a particular prior distribution π in $\mathcal{M}$ is that π is a plausible candidate to be the ideal distribution π_0. The resulting set of priors is in general a *neighbourhood model*, i.e., the set of all distributions that are close (w.r.t. some criterion) to π_0. Exam-

ples of neighbourhood models are: ϵ-contamination models [5, 6]; restricted ϵ-contamination models [7]; intervals of measures [6, 8]; the density ratio class [2, 8], etc. However, this approach can be unsatisfactory when there is (almost) no prior information or the information is of doubtful relevance. Then there is no ideal prior distribution π_0, because no single prior distribution could adequately model the limited prior information. Therefore, in this case, also a neighbourhood model can be inadequate.

The second approach, known as the theory of imprecise probabilities or coherent lower (and upper) previsions, was developed by Walley [2] from earlier ideas [9, 10, 11]. This approach revises Bayesian sensitivity analysis by directly emphasizing the upper and lower expectations (also called previsions) that are generated by $\mathcal{M}$. The upper and lower expectations of a bounded real-valued function g (we call it a gamble) on a possibility space $\mathcal{W}$, denoted by $\overline{E}(g)$ and $\underline{E}(g)$, are respectively the supremum and infimum of the expectations $E_P(g)$ over the probability measures P in $\mathcal{M}$ (if $\mathcal{M}$ is assumed to be closed[1] and convex, it is fully determined by the upper and lower expectations for all gambles). The upper and lower expectations have a behavioural interpretation (explained in Section 2), and, contrary to the robust Bayesian approach, there is no special commitment to the individual probability distributions in $\mathcal{M}$.

In choosing a set $\mathcal{M}$ to model prior ignorance, the main aim is to generate upper and lower expectations with the property that $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$ for a specific class of gambles g of interest. This means that the only information about $E(g)$ is that it belongs to $[\inf g, \sup g]$, which is equivalent to state a condition of complete prior ignorance about the value of g (this is the reason why we said that a single, however noninformative, prior cannot model prior ignorance).

For instance, assume that the variable W is a location parameter and that $\mathcal{W} = \mathbb{R}$; then in case there is no prior information about the value w of W, one expects that for[2] $g = I_{\{(-\infty,a]\}}$ for any finite $a \in \mathbb{R}$, the lower and upper expected values of g be $\underline{E}(I_{\{(-\infty,a]\}}) = 0$ and $\overline{E}(I_{\{(-\infty,a]\}}) = 1$. Here $I_{\{(-\infty,a]\}}$ denotes the indicator function over the set $(-\infty, a]$, i.e., $I_{\{(-\infty,a]\}}(w) = 1$ if w belongs to $(-\infty, a]$, and $I_{\{(-\infty,a]\}}(w) = 0$ otherwise. Since $E(I_{\{(-\infty,a]\}})$ is equal

[1] In the weak* topology; see [2, Sec. 3.6] for more details.

[2] In the following, we use the notation g without argument to denote the gamble, while $g(w)$ is used to denote the value of the gamble g in w.

to the cumulative distribution of W, $\underline{E}(I_{\{(-\infty,a]\}}) = 0$ and $\overline{E}(I_{\{(-\infty,a]\}}) = 1$ state that the only knowledge about the cumulative distribution of w is that it is between 0 and 1, which is a state of complete ignorance.

Modeling a state of prior ignorance about the value w of the random variable W is not the only requirement for $\mathcal{M}$, it should also lead to non-vacuous posterior inferences. Posterior inferences are vacuous if the posterior lower and upper expectations of all gambles g coincide with the infimum and, respectively, the supremum of g. Notice that, in case $\mathcal{M}$ includes all the possible prior distributions, any inference about W is vacuous, i.e., the set of posterior distributions obtained by applying Bayes' rule to a given likelihood and to each distribution in the set of priors includes again all the possible distributions. This means that our beliefs do not change with experience (i.e., there is no learning from data). This point is clearly stated in [12], where the authors define some properties that a general set $\mathcal{M}$ of distributions should fulfil to model a state of prior ignorance about W. In particular, $\mathcal{M}$ should produce self-consistent probabilistic models, represent the state of prior ignorance, give non-vacuous inferences, satisfy certain invariance properties and be easy to elicit and tractable [12].

In this paper, we follow the second approach to modelling prior ignorance about statistical parameters by using Walley's theory of coherent lower previsions [2]. Let us summarize the main contributions of this paper.

In Section 2, inspired by the work in [12], we define some *minimal properties* that a set $\mathcal{M}$ of distributions should satisfy to be a model of prior ignorance that does not lead to vacuous inferences. These minimal properties are obtained by relaxing and generalizing the properties in [12]. The new set of properties now also captures the case in which the likelihood is not perfectly known.

Next, we focus on exponential families. We give some preliminaries about these families in Section 3. In Section 4, we consider the case that the likelihood model is a one-parameter exponential family and $\mathcal{M}$ includes a subset of the corresponding conjugate exponential priors. We show that there exists a parametrization of $\mathcal{M}$ which equivalently satisfies the properties defined in Section 2, and which is unique up to the choice of its size which determines the degree of robustness (or caution) of the inferences. Stated differently, we prove that the set of priors $\mathcal{M}$ satisfying the properties from Section 2 can be uniquely obtained by letting the parameters of the conjugate exponential prior vary in suitable sets. The obtained set $\mathcal{M}$ has the following characteristics, which could make of it an appealing alternative to noninformative

priors to model prior ignorance:

- $\mathcal{M}$ produces self-consistent (or *coherent*) probabilistic models;
- $\mathcal{M}$ is a model of prior ignorance w.r.t. the functions (queries) that are commonly used for statistical inferences (i.e., lower/upper expectations of such functions are vacuous a-priori);
- it is easy to elicit and, because of conjugacy, tractable;
- the inferences drawn with the model $\mathcal{M}$ encompass, with a suitable choice of the size of $\mathcal{M}$, the frequentist inferences and objective Bayesian inferences with improper priors (the inferences are naturally robust).

We compare the obtained set of priors $\mathcal{M}$ with other models of prior ignorance expressed through set of distributions in Section 4.1. In Section 5, we discuss the choice of the parameters governing the robustness of the inferences that are obtained through $\mathcal{M}$, while in Section 6 we apply these ideas to a problem of inference with count data.

2. Properties for Prior Near-Ignorance

The aim of this section is to define which minimal properties the set of priors $\mathcal{M}$ should satisfy in the case where there is (almost) no prior information about a random variable W taking values in $\mathcal{W} \subseteq \mathbb{R}$. Before listing these properties, we discuss the interpretation of $\mathcal{M}$ by briefly introducing the behavioural interpretation of upper and lower expectations.

Stemming from de Finetti's [13] work on subjective probability, Walley [2] gives a behavioural interpretation of $\mathcal{M}$ in terms of buying and selling prices. Let us briefly sketch how this is done.

By regarding a gamble $g : \mathcal{W} \to \mathbb{R}$ as a random reward, which depends on the a priori unknown value w of W, the expectation (also called prevision) of g w.r.t. w, i.e., $E(g)$, represents a subject's fair price for the function g. This means that he should be disposed to accept the uncertain rewards $g - E(g) + \epsilon$ (i.e., to *buy* g at the price $E(g) - \epsilon$) and $E(g) - g + \epsilon$ (i.e., to *sell* g at the price $E(g) + \epsilon$) for every $\epsilon > 0$. In de Finetti's theory, the acceptable buying and selling prices must coincide. This fair price is what we normally call the expectation of g.

More generally speaking, the supremum acceptable buying price and the infimum acceptable selling prices for g need not coincide, meaning that there

may be a range of prices $[a, b]$ for which our subject is neither disposed to buy nor to sell g at a price $k \in [a, b]$. His supremum acceptable buying price for g is then his lower expectation $\underline{E}(g)$, and it holds that the subject is disposed to accept the uncertain reward $g - \underline{E}(g) + \epsilon$ for every $\epsilon > 0$; and his infimum acceptable selling price for g is his upper prevision $\overline{E}(g)$, meaning that he is disposed to accept the reward $\overline{E}(g) - g + \epsilon$ for every $\epsilon > 0$. A consequence of this interpretation is that $\underline{E}(g) = -\overline{E}(-g)$ for every function g.[3]

Under this behavioural interpretation, a state of ignorance about a gamble g is modelled by setting $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$. This means that our subject is neither disposed to buy nor to sell g at any price $k \in (\inf g, \sup g)$. In other words, our subject is disposed to buy (sell) g only at a price strictly less (greater) than the minimum (maximum) expected reward that he would gain from g. This means that the information available about W does not allow our subject to set any meaningful buying or selling price for g. It corresponds to stating that our subject is in a state of ignorance.

Walley [2] proves that a closed and convex set of (finitely additive) probabilities can be equivalently characterized by the lower (or upper) expectation functional that it generates as the lower (upper) envelope of the expectations obtained from the probabilities in such a set. Vice versa, given a functional $\underline{E}$ that satisfies some regularity properties [2, Ch. 2], it is possible to define a family $\mathcal{M}$ of probabilities that generates the lower expectation $\underline{E}(g)$ for any g. This establishes a one-to-one correspondence between closed convex sets of probabilities and lower expectations.

In case the available prior information is scarce, it seems more natural to define $\mathcal{M}$ according to the behavioural interpretation, i.e., in terms of the upper and lower expectations it generates [12]. For instance, in problems where there is (almost) no prior information one would expect the set $\mathcal{M}$ to be large in the sense that it generates upper and lower expectations that are relatively far apart (vacuous or almost vacuous).

Modelling a state of prior ignorance about W is not the only requirement for $\mathcal{M}$, it must also produce non-vacuous posterior inferences (otherwise it is useless in practice). Hereafter, inspired by the work in [12], we define a set of minimal properties that $\mathcal{M}$ or, equivalently, the lower and upper expectations it generates, should satisfy to be a model of prior ignorance and to produce consistent and meaningful posterior inferences. Then, in the next

[3]Because of the relationship $\underline{E}(g) = -\overline{E}(-g)$, only $\underline{E}(g)$ or $\overline{E}(g)$ needs to be specified.

sections, we specialize these requirements to the case of the one-parameter exponential families. The first requirement for $\mathcal{M}$ is coherence.

(A.1) Coherence. Prior and posterior inferences based on $\mathcal{M}$ should be strongly coherent [2, Sec. 7.1.4(b)]. Under the behavioural interpretation, this means that we should not be able to raise the lower expectation (supremum acceptable buying price) of a given gamble g taking into account the acceptable transactions implicit in the lower expectations for other gambles.

In practice, strong coherence imposes joint constraints on the prior, likelihood and posterior lower expectation models, in the sense that, when considered jointly, they should not imply inconsistent assessments. Walley [2, Sec. 7.8.1] proves that, in case the prior and likelihood lower expectation models are obtained as lower envelopes of standard expectations w.r.t. sets of proper density functions and the posterior set of densities is obtained from these sets by element-wise application of Bayes' rule for density functions, then strong coherence of the respective lower expectation models is satisfied.[4] The following example shows that, when this is not the case, the inference model can be incoherent [2, Sec. 7.4.4].

Example 1. Consider the following probabilistic model: prior $p(w) = 1$, *Normal likelihood* $\prod_{i=1}^{n} \mathcal{N}(y_i; w, \sigma^2)$ *with variance* σ^2 *and Normal posterior* $\mathcal{N}(w; \hat{y}_n, \sigma^2/n)$ *with mean* $\hat{y}_n = \frac{1}{n}\sum_i y_i$ *(sample mean) obtained by combining, via Bayes' rule, likelihood and prior. To see that the inferences from the likelihood and the posterior model are incoherent, consider the event* $A = \{(w, \hat{y}_n) : |w| \leq |\hat{y}_n|\}$ *and, thus, the gamble* $g = I_A$. *Then, from the likelihood model it follows that* $P(A|w) = E[I_A|w] = \frac{1}{2} + \Phi(-2|w|\sqrt{n}/\sigma) > \frac{1}{2}$, *for any* $w \in \mathbb{R}$, *where* Φ *denotes the cumulative distribution of the standard Normal distribution. By considering the posterior model, one has* $P(A|y_1, \ldots, y_n) = E[I_A|y_1, \ldots, y_n] = \frac{1}{2} - \Phi(-2|\hat{y}_n|\sqrt{n}/\sigma) < \frac{1}{2}$, *for any* $y_1, \ldots, y_n$. *Thus the likelihood and posterior models are inconsistent. In fact, since* $E(I_A|y_1, \ldots, y_n) < 0.5$ *for each* $y_1, \ldots, y_n$ *and since* $E[I_A|y_1, \ldots, y_n]$ *is equal to the probability* $P(A|y_1, \ldots, y_n)$ *that the value* w *of* W *belongs to* A, *the above equality im-*

[4] This holds under standard assumptions about the existence of density functions and the applicability of Bayes' rule [2, Sec. 6.10.4].

plies that we are prepared to bet against A at "even money" according to the posterior model (no matter the value $\hat{y}_n$), but this strategy is inconsistent with the likelihood model which assesses that $P(A|w) > 0.5$ (no matter the value w of W). Observe that the posterior $\mathcal{N}(w; \hat{y}_n, \sigma^2/n)$ has been obtained, via Bayes' rule, from the likelihood and the improper uniform prior on W (i.e., $p(w) = 1$). A criticism of the improper prior is, in fact, that the posterior distribution it generates is often not coherent with the likelihood model for finite n. Finally, observe that, for $n \to \infty$, it follows that $P(A|w), P(A|y_1, \ldots, y_n) \to 0.5$ and, thus, the incoherence vanishes in the limit. ■

Besides coherence, other requirements for the set $\mathcal{M}$ are that it should represent the state of prior ignorance about W, but without producing vacuous posterior inferences (posterior inferences are vacuous if the lower and upper expectations of all gambles g coincide with the infimum and, respectively, the supremum of g). In the case $\mathcal{M}$ includes all the possible prior distributions, any inference about W is vacuous, i.e., the set of posterior distributions obtained by applying Bayes' rule to a given likelihood and to each distribution in the set of priors includes again all the possible distributions. This means that our prior beliefs do not change with experience (there is no learning from data). Thus, $\mathcal{M}$ should be large enough to model a state of prior ignorance w.r.t. a set of suitable gambles (i.e., a set of gambles of interest $\mathcal{G}_0$ w.r.t. which we assess our state of prior ignorance), but no too large to prevent learning from taking place. These two opposite requirements are captured by the following two properties for $\mathcal{M}$.

(A.2) $\mathcal{G}_0$-prior ignorance. The prior upper and lower expectations of some suitable set of gambles $\mathcal{G}_0$ under $\mathcal{M}$ are vacuous, i.e., $\underline{E}[g] = \inf g$ and $\overline{E}[g] = \sup g$ for all $g \in \mathcal{G}_0$.

(A.3) $\mathcal{G}$-learning. For a chosen set of gambles $\mathcal{G} \supseteq \mathcal{G}_0$ and for each $g \in \mathcal{G}$ satisfying $\overline{E}[g] - \underline{E}[g] > 0$, there exists a finite $\delta > 0$ (possibly dependent on g) such that for each $n \geq \delta$ and sequence of observations $y^n = (y_1, \ldots, y_n)$, at least one of these two conditions is satisfied:

$$\underline{E}[g|y^n] \neq \underline{E}[g], \quad \overline{E}[g|y^n] \neq \overline{E}[g], \tag{1}$$

where $\underline{E}[\cdot|y^n]$ and $\overline{E}[\cdot|y^n]$ denote the posterior lower and upper expectations of g after having observed $y_1, \ldots, y_n$.

Property (A.2) states that $\mathcal{M}$ should be vacuous a priori w.r.t. some set of gambles $\mathcal{G}_0$, i.e., the lower and upper expectations of $g \in \mathcal{G}_0$ respectively coincide with the infimum and the supremum of g. In case $\mathcal{M}$ includes all possible distributions then (A.2) holds for any function g. Here, conversely, we require that (A.2) is satisfied for some subset of gambles $\mathcal{G}_0$. The subset of gambles $\mathcal{G}_0$ used in (A.2) should consist of the gambles g w.r.t. which we wish to state our condition of prior ignorance. Furthermore, in case of complete prior ignorance the set $\mathcal{G}_0$ should be as large as possible to guarantee that also $\mathcal{M}$ is as large as possible,[5] but no too large to be incompatible with the requirement (A.3) of learning. In fact, property (A.3) states that $\mathcal{M}$ should be non-vacuous a posteriori for any gamble $g \in \mathcal{G} \supseteq \mathcal{G}_0$, which is a condition for learning from the observations. The set of gambles $\mathcal{G}$ used in (A.3) should include the gambles g w.r.t. which we are interested in computing expectations (i.e., making inferences). The fact that $\mathcal{G}$ must include $\mathcal{G}_0$ is the only constraint on $\mathcal{G}$, meaning that (A.3) requires that $\mathcal{M}$ is not vacuous w.r.t. all these gambles for which the prior near-ignorance has been imposed.

Since $\mathcal{M}$ is a model of prior ignorance, it is also desirable that the influence of $\mathcal{M}$ on the posterior inferences vanishes with increasing number of observations n. This is captured by the following property.

(A.4) Convergence. For each gamble $g \in \mathcal{G}$ and non-empty sequence of observations y^n, the following conditions are satisfied for $n \to \infty$:

$$\underline{E}[g|y^n] \to \underline{E}^*[g|y^n], \qquad \overline{E}[g|y^n] \to \overline{E}^*[g|y^n], \tag{2}$$

where $\underline{E}^*[g|y^n], \overline{E}^*[g|y^n]$ are the posterior lower and upper expectations obtained as lower envelopes of standard expectations w.r.t. the posterior densities derived, via Bayes' rule, from the likelihood model and the improper uniform prior density on $\mathcal{W}$.

Property (A.4) states that, for $n \to \infty$, $\mathcal{M}$ should give the same lower and upper expectations of $g \in \mathcal{G}$ as those obtained from the improper uniform prior density on $\mathcal{W}$ (i.e., $p(w) = 1$). The fact that $\underline{E}^*[g|y^n] < \overline{E}^*[g|y^n]$ accounts for the general case in which the likelihood model is described by a set of likelihoods (for a single likelihood it would be $\underline{E}^*[g|y^n] = \overline{E}^*[g|y^n] = E^*[g|y^n]$). Although improper priors produce posteriors which are often incoherent with

[5] Note that if $\mathcal{G}_0$ included all gambles then $\mathcal{M}$ would be the set of all distributions.

the likelihood model, (A.4) does not conflict with the requirement of coherence in (A.1). In fact (A.4) is a limiting property that holds only for $n \to \infty$.[6] In order to better understand properties (A.1)–(A.4) we show their instantiation to the case of the set of exponential families we propose in Section 4. Before discussing these results, in the next section we introduce the exponential families of distributions and review their main properties [1, Ch. 5].

3. Exponential Families

Consider a sampling model where i.i.d. samples of a random variable Z are taken from a sample space $\mathcal{Z}$.

Definition 1. A probability density (or mass function) $p(z|x)$, parametrized by the continuous parameter x taking values in $\mathcal{X} \subseteq \mathbb{R}$, is said to belong to the one-parameter exponential family if it is of the form

$$p(z|x) = f(z)[g(x)]^{-1} \exp\left(c\phi(x)h(z)\right), \ z \in \mathcal{Z}, \tag{3}$$

where, given the real-valued functions f, h, ϕ and scalar c, it results that $g(x) = \int_{z\in\mathcal{Z}} f(z) \exp\left(c\phi(x)h(z)\right) dz < \infty$.

■

Sometimes it is more convenient to rewrite (3) in a different form.

Definition 2. The probability density (or mass function):

$$p(y|w) = a(y) \exp(yw - b(w)), \ y \in \mathcal{Y}_m, \tag{4}$$

derived from (3) via the transformations $y = h(z)$,[7] $\mathcal{Y}_m = h(\mathcal{Z})$, $w = c\phi(x)$, $b(w) = \ln(g(x))$ and $a(y) = f(z)$, is called the canonical form of representation of the exponential family; w is called the natural (or canonical) parameter and takes values in the parameter space

$$\mathcal{W} = \{w \in \mathcal{R} : \ b(w) < \infty\}. \tag{5}$$

[6] Furthermore, in Example 1, incoherence vanishes at the limit. This is also true for other members of the exponential families, since they become asymptotically Gaussian [1, Prop. 5.16] with a variance that goes to zero for $n \to \infty$.

[7] The Jacobian of the transformation should also be considered in case $p(z|x)$ is a density function.

■

It can be shown that $\mathcal{W}$ is an open convex set [14]. The canonical form has some useful properties. The mean and variance of Y are given by

$$E[Y|w] = \frac{db}{dw}(w),\ E[(Y - E[Y|w])^2|w] = \frac{d^2b}{dw^2}(w), \quad w \in \mathcal{W}, \tag{6}$$

where it has been assumed that $\frac{d^2b}{dw^2}(w) > 0$; from (6) it follows that $\frac{db}{dw}(w) \in Int(\mathcal{Y})$, i.e., the interior of $\mathcal{Y}$ [14], where $\mathcal{Y} \subseteq \mathbb{R}$ is the smallest closed or semi-closed convex set that includes the sample mean of an arbitrary sequence of observations (if this set exists, otherwise $\mathcal{Y} = \mathbb{R}$ and, thus, $Int(\mathcal{Y}) = \mathbb{R}$). Notice that the domain of the observations $\mathcal{Y}_m$ can be discrete or continuous (whether $p(y|w)$ is a mass function or a density function), while $\mathcal{Y}$ is always continuous. In the case of n i.i.d. observations $y_i = h(z_i)$, it follows that

$$p(y^n|w) = \prod_{i=1}^{n} p(y_i|w) = \exp(n(\hat{y}_n w - b(w))) \prod_{i=1}^{n} k(y_i), \quad w \in \mathcal{W}, \tag{7}$$

where $\hat{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean of the y_i which, together with n, is a sufficient statistic of y^n for inference about W under the i.i.d. assumption. Furthermore, by interpreting the density function in (7) as a likelihood function $L(W)$, with $y^n = (y_1, \ldots, y_n)$, we can define the corresponding conjugate prior.

Definition 3. A probability density $p(w|n_0, y_0)$, parametrized by $n_0 \in \mathbb{R}^+$ and $y_0 \in Int(\mathcal{Y})$, is said to be the canonical conjugate prior of (4) if

$$p(w|n_0, y_0) = k(n_0, y_0) \exp(n_0(y_0 w - b(w))), \tag{8}$$

where $w \in \mathcal{W}$, n_0 is the so-called number of pseudo-observations, y_0 is the so-called pseudo-observation and $k(n_0, y_0)$ is the normalization constant. ■

Proposition 1. *If $w \in \mathcal{W}$ the canonical conjugate prior is a well-defined probability density (it can be normalized) provided that $0 < n_0 < \infty$ and $y_0 \in Int(\mathcal{Y})$.* ■

For the proof of this proposition see [15, Sec. 4.18]. Some examples of one-parameter exponential (canonical) families and their conjugate densities and defined on $\mathcal{W} = \mathbb{R}$ follow.

Example 2. Gaussian with known variance. $z = y \in \mathcal{Y} = \mathbb{R}$, $x \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$,

$$p(y|x, \sigma^2) \propto \exp\left(-\tfrac{1}{2\sigma^2}(y-x)^2\right) \propto \exp\left(\tfrac{1}{\sigma^2}\left(yx - \tfrac{x^2}{2}\right)\right),$$

with $w = x/\sigma^2$ *and* $b(w) = x^2/2\sigma^2$. *The conjugate prior (8) is:*

$$p(x|n_0, y_0) \propto \exp\left(-\tfrac{n_0}{2\sigma^2}(x-y_0)^2\right),$$

which is a Gaussian with mean y_0 *and variance* σ^2/n_0. ■

Example 3. Binomial-Beta. $x \in \mathcal{X} = (0,1)$, $z = y \in \{0,1\}$,

$$\begin{aligned} p(y|x) &\propto x^y(1-x)^{(1-y)} \\ &= (1-x)\exp\left(y\ln\left(\frac{x}{1-x}\right)\right) \\ &= \exp\left(yw - b(w)\right), \end{aligned}$$

$w = \ln(x/(1-x))$, $b(w) = -\ln(1-x) = \ln(1+\exp(w))$. *Considering the change of variable* $dx = \exp(w)/(1+\exp(w))^2 dw$, *the conjugate prior (8) transformed back to the original domain* $\mathcal{X}$ *is:*

$$p(x|n_0, y_0) \propto x^{n_0 y_0 - 1}(1-x)^{n_0(1-y_0)-1},$$

which is a Beta density with $n_0 > 0$ *and* $y_0 \in (0,1)$. ■

The likelihood and conjugate prior pair in the canonical exponential family satisfies a set of interesting properties, most of them are particularly useful to represent the nature of the Bayesian learning process. A list of such properties is given in the following propositions, whose proofs are available in literature (see for instance [1, Ch. 5]).

Proposition 2. *For a pair of likelihood and conjugate prior in the canonical exponential family, it holds that:*

(i) *the posterior density for w is:*

$$p\left(w|n_p, y_p\right) = k(n_p, y_p)\exp(n_p(y_p w - b(w))), \tag{9}$$

where $n_p = n + n_0$ and $y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$;

(ii) *the predictive density for future observations $(y_{n+1}, \ldots, y_{n+m})$ is*

$$p(y_{n+1}, \ldots, y_{n+m}|y_1, \ldots, y_n) = \prod_{j=1}^{m} k(y_{n+j}) \frac{k\left(n_0 + n, \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}\right)}{k\left(n_0 + n + m, \frac{n_0 y_0 + (n+m)\hat{y}_{n+m}}{n+m+n_0}\right)}. \tag{10}$$

■

Proposition 3. *The prior mean of the function $b'(w) = \frac{d}{dw}b(w)$ is $E[b'|n_0, y_0] = y_0$ and the posterior mean is:*

$$E\left[b'\middle|n_p, y_p\right] = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0}. \tag{11}$$

■

In (6), it has been shown that b' is the mean of Y conditional on w. Hence, b' is the quantity about which we will have prior beliefs before seeing the data y and posterior beliefs after observing the data. Hence, the results in Proposition 3 are particularly important, because they provide us with a closed formula for the prior and posterior expectation of b'. For sampling models such that $\frac{d}{dw}b(w) = x$ (e.g., Gaussian, Beta and Gamma density), Proposition 3 gives thus a closed formula for the prior and posterior expectation of X.

Proposition 4. *Suppose that the canonical conjugate prior family is such that $\frac{d}{dw}b(w) = x$, then y_0 and (11) are the prior and, respectively, posterior mean of X.* ■

Hereafter, we give the instantiation of the above propositions in the case of the conjugate Gaussian and Binomial model.

Example 4. Gaussian with known variance. In Example 3, it has already been shown that

$$\mathcal{N}(x; y_0, \sigma^2/n_0) = p(w|n_0, y_0) = k(n_0, y_0) \exp(n_0(y_0 w - b(w))),$$

with $w = x/\sigma^2$ *and* $b(w) = x^2/2\sigma^2$*. By considering the likelihood,*

$$L(y^n|w) = \prod_{i=1}^{n} \mathcal{N}(y_i; x, \sigma^2) = \exp(n(\hat{y}_n w - b(w))) \prod_{i=1}^{n} k(y_i)$$

and assuming that the variance σ^2 *is known, then* $\mathcal{N}(x; y_0, \sigma^2/n_0)$ *and* $L(y^n|w)$ *are conjugate and, thus, the posterior density is*

$$\mathcal{N}(x; y_p, \sigma_p^2) = k(n_p, y_p) \exp(n_p(y_p w - b(w)))$$

where

$$y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0}, \quad \sigma_p^2 = \frac{\sigma^2}{n_p} = \frac{\sigma^2}{n + n_0}.$$

Since $\frac{d}{dw} b(w) = w = x$ *then from Proposition 4 it follows that* y_p *is also the posterior mean of* x*. It is straightforward to verify that* σ_p^2 *is also the posterior variance of* x*.* ■

Example 5. Binomial-Beta. For the Binomial-Beta conjugate model, in Example 3 it has been shown that

$$Beta(x|s, t) = k(n_0, y_0) \exp(n_0(y_0 w - b(w))),$$

with $w = \ln(x/(1-x))$ *and* $b(w) = -\ln(1-x) = \ln(1 + \exp(w))$*. By considering a binomial density as the likelihood,*

$$Bi(y^n|x) = \exp(n(\hat{y}_n w - b(w))) \prod_{i=1}^{n} k(y_i)$$

then $Beta(x|s, t)$ *and* $Bi(y^n; x)$ *are conjugate and, thus, the posterior density is*

$$Beta\left(x \middle| n_0 + n, \frac{n_0 y_0 + \sum_{i=1}^{n} y_i}{n_0 + n}\right) = k(n_p, y_p) \exp(n_p(y_p w - b(w))),$$

where $n_p = n + s$ *and* $y_p = \frac{n_0 y_0 + \sum_{i=1}^n y_i}{s+n}$. *In this case, the posterior mean and variance are*

$$y_p = \frac{n_0 y_0 + \sum_{i=1}^n y_i}{n_0 + n}, \qquad \sigma^2 = \frac{y_p(1-y_p)}{n_0 + n + 1}.$$

Since $\frac{d}{dw} b(w) = x$ *then from Proposition 4 it follows again that* $\hat{\mu}$ *is also the posterior mean of* x. *It is straightforward to verify that* $\hat{\sigma}^2$ *is also the variance of* x. ■

4. Sets of Conjugate Priors for Exponential Families

Consider the problem of statistical inference about the real-valued parameter w from noisy measurements $\{y_1, \ldots, y_n\}$ and assume that the likelihood is completely described by an exponential family probability density function (PDF) (or mass function (PMF) in the discrete observations case):

$$\prod_{i=1}^n p(y_i|w) = \exp(n(\hat{y}_n w - b(w))) \prod_{i=1}^n k(y_i), \tag{12}$$

where the parameters of the likelihood, i.e., sample mean $\hat{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ and $n \in \mathbb{N}$, are known.[8] By conjugacy and following a Bayesian approach, as prior for w we may consider the PDF $p(w|n_0, y_0)$ defined in (8) for a given value of the parameter y_0 and n_0.

In the case there is not enough information about w to uniquely determine the values of the parameters y_0 and n_0, we can consider the family of priors $p(w|n_0, y_0)$ obtained by letting y_0 vary in $\mathcal{Y}' \subseteq Int(\mathcal{Y})$ and n_0 in some set $\mathcal{A}_{y_0} \subseteq \mathbb{R}^+$, which may depend on y_0. The question to be addressed is when such a family of priors satisfies the properties (A.1)–(A.4) discussed in Section 2. The answer to this question is given in the next theorem.

Theorem 1. *Consider as set of priors* $\mathcal{M}$ *the family of proper conjugate priors* $p(w|n_0, y_0)$ *with:*

- y_0 *spanning the set* $\mathcal{Y}' \subseteq Int(\mathcal{Y}) \subseteq \mathbb{R}$,

[8] Also the likelihood may be modelled by a set of PDFs (or PMFs) (for instance in case of interval data). However, in this paper we assume that the likelihood is a single PDF (or PMF), i.e., the sufficient statistics $\hat{y}_n$ and n are known.

- n_0 *spanning the set* $\mathcal{A}_{y_0} \subset \mathbb{R}^+$*, with* $\mathcal{A}_{y_0}$ *possibly dependent on* y_0.[9]

If and only if the following conditions hold:

(a) $\mathcal{Y}' = Int(\mathcal{Y})$*,*

(b) $\mathcal{A}_{y_0}$ *satisfies the following constraints:* $\sup \mathcal{A}_{y_0} \leq \min(\overline{n}_0, \frac{c}{|y_0|})$ *for each* $y_0 \in Int(\mathcal{Y})$ *and given the parameters* $\overline{n}_0, c > 0$*,*

then, given the design parameters $\overline{n}_0$ *and* c*,* $\mathcal{M}$ *is the largest set which satisfies properties (A.1)–(A.4), with* $\mathcal{G}_0 = \{b'\}$ *and* $\mathcal{G}$ *including sufficiently smooth gambles.*[10] ■

Before giving the proof of the above theorem, we discuss its meaning. Theorem 1 states that properties (A.1)–(A.4) can be satisfied by the set of conjugate priors $\mathcal{M}$ if and only if y_0 is free to vary in the convex set $Int(\mathcal{Y})$ and n_0 is bounded above by the function $\min(\overline{n}_0, \frac{c}{|y_0|})$, which depends on y_0 and the design parameters $\overline{n}_0$ and c. Since $y_0 \in Int(\mathcal{Y})$ and $0 < n_0 < \infty$, from Proposition 1, it follows that all priors in $\mathcal{M}$ are proper and well-defined PDFs. Observe that the result of Theorem 1 has been derived by imposing (A.2) only on the gamble b', i.e., $\mathcal{G}_0 = \{b'\}$. In Section 4.1, it will be shown that the obtained set of priors satisfies (A.2) also for other functions of interest in statistical analysis (i.e., indicators over intervals that are used to compute one- and two-sided hypothesis testing and credible intervals). Therefore, the choice of imposing prior ignorance only on b' is not limiting for exponential families. This choice has also been motivated by the meaning of b' for exponential families. Remember in fact from Section 3 that b' is the mean of Y and, thus, is the quantity about which we will have prior beliefs before seeing the data and posterior beliefs after observing the data.

Proof: *The proof is organized as follows. First we prove the necessity of the conditions (a)–(c) for (A2)–(A4). Second we prove their sufficiency. Then we show that* $\mathcal{M}$ *is the largest set which satisfies these properties. Finally,*

[9] The set $\mathcal{A}_{y_0} \times \mathcal{Y}' \subseteq \mathbb{R}^+ \times \mathbb{R}$ can have arbitrary form.

[10] With sufficiently smooth gambles, we mean integrable w.r.t. the exponential family density functions with support in $\mathcal{W}$ and continuous on a neighborhood of the point where the posterior relative to the improper uniform prior on $\mathcal{W}$ ($p(w) = 1$) concentrates for $n \to \infty$.

we prove (A.1).
Property (A.2): prior ignorance, in case $g = b'$. Consider the prior $p(w|n_0, y_0)$ and the function $\frac{db}{dw}(w)$. From Proposition 3 it follows that $E[b'|n_0, y_0] = y_0$. Since the codomain of b' is $Int(\mathcal{Y})$, because of (6), a necessary condition for (A.2) to be satisfied is $\mathcal{Y}' = Int(\mathcal{Y})$.[11] *In fact, in this case, since $\underline{E}[b'|n_0, y_0] = y_0$, it follows that, for $g = b'$, $\underline{E}[g] = \inf \mathcal{Y}' = \inf g$ and $\overline{E}[g] = \sup \mathcal{Y}' = \sup g$. This proves that $\mathcal{Y}' = Int(\mathcal{Y})$ is a necessary condition for (A.2) to hold in the case $g = b'$.*
Property (A.3): learning. To prove this property, we exploit the fact that the posterior density in (9) belongs to the exponential family and, thus, is fully described by the parameters $n_p = n + n_0$ and $y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$. For the proof, we distinguish three cases $\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = [a, \infty)$ (or $\mathcal{Y} = (-\infty, a]$) with $a \in \mathbb{R}$, and $\mathcal{Y} \subset \mathbb{R}$ bounded. In the last two cases w.l.o.g. it can be assumed that $\mathcal{Y} = [0, \infty)$ (or $\mathcal{Y} = (-\infty, 0]$) and, respectively, $\mathcal{Y} = [0, 1]$ (by shifting and scaling $\mathcal{Y}$); since $\mathcal{Y}$ has been assumed to be convex, these three cases account for all the possibilities.

Consider the posterior density in (9), i.e,

$$p\left(w|n_p, y_p\right) = k(n_p, y_p) \exp(n_p(y_p w - b(w))), \quad w \in \mathcal{W}, \tag{13}$$

where $n_p = n + n_0$ and $y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$. Then a necessary condition for (A.3) to hold is that n_0 is bounded from above (by some $\overline{n}_0 < \infty$). In fact, assume that this does not hold and consider the gamble $g = b' \in \mathcal{G}_0$. Since $E[b'|n_p, y_p] = y_p$, it results that for $n_0 \to \infty$, then $y_p \to y_0$, which means no learning for any value of y_0 and $\hat{y}_n$. In the case $n_0 \leq \overline{n}_0 < \infty$ holds, condition (A.3) can still be violated if $\mathcal{Y} = [0, +\infty)$ (or $\mathcal{Y} = (-\infty, 0]$) or $\mathcal{Y} = (-\infty, +\infty)$. For $\mathcal{Y} = [0, +\infty)$, assume that $\hat{y}_n = 0$, then

$$0 \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0} < \infty, \tag{14}$$

where the lower bound has been obtained for $y_0 = 0$ and the upper bound for $n_0 y_0 \to \infty$.[12] *In the case $\mathcal{Y} = (-\infty, +\infty)$, then, for any $\hat{y}_n$,*

$$-\infty \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0} < \infty, \tag{15}$$

[11] Actually it is enough that $\mathcal{Y}'$ include neighbourhoods of the extremes of the set $\mathcal{Y}$.
[12] The limit $n_0 y_0 \to \infty$ means that the increase of y_0 is faster than the decrease of n_0.

where the right and left bounds have been obtained for $n_0 y_0 \to \pm\infty$. Therefore, $n_0|y_0| \leq c$ for some $0 < c < \infty$ is also a necessary condition for learning. Since there cannot be convergence without learning, the conditions $n_0 \leq \overline{n}_0 < \infty$ and $n_0|y_0| \leq c$ are also necessary for (A.4).

Consider now sufficiency. For (A.2), the condition $\mathcal{Y}' = Int(\mathcal{Y})$ is clearly also sufficient. Consider (A.3). Since for $n_0 \leq \overline{n}_0 < \infty$, $n_0|y_0| \leq c$ and for $n \to \infty$ (i.e., the number of observations goes to infinity), it results that $n_p \approx n$ and $y_p \approx \hat{y}_n$. Then, since it has been assumed a priori that $\overline{E}[g] - \underline{E}[g] > 0$, both conditions in (1) cannot be false at the same time. Assume, by contradiction, that there exists a gamble $g \in \mathcal{G}$ w.r.t. which the lower and upper prior and posterior expectations are, respectively, always equal for each n. Since for $n \to \infty$, it results that $y_p \to \hat{y}_n$ and $n_p \to n$, then the lower and upper posterior expectations converge, i.e., $\underline{E}[g|y_1, \ldots, y_n] \to \overline{E}[g|y_1, \ldots, y_n]$. This implies that $\underline{E}[g] = \overline{E}[g]$, which is a contradiction because by hypothesis we have assumed that $\overline{E}[g] - \underline{E}[g] > 0$. A consequence is that there exists a $\delta > 0$ such that for all $n > \delta$ either $\underline{E}[g|y_1, \ldots, y_n] \neq \underline{E}[g]$ or $\overline{E}[g|y_1, \ldots, y_n] \neq \overline{E}[g]$. To prove the second part of (A.3), consider the gamble b'. In this case the lower and upper bound of $E[b'|y_1, \ldots, y_n]$ for the case $\mathcal{Y} = [0, 1]$ are:

$$\frac{n\hat{y}_n}{n + \overline{n}_0} \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0} \leq \frac{\overline{n}_0 + n\hat{y}_n}{n + \overline{n}_0}, \tag{16}$$

which are obtained for $y_0 = 0$, $n_0 = \overline{n}_0$ (lower bound) and $y_0 = 1$, $n_0 = \overline{n}_0$ (upper bound). Observe that $\hat{y}_n \leq 1$ because $\hat{y}_n \in \mathcal{Y}$. In case $\mathcal{Y} = [0, +\infty)$ (or $\mathcal{Y} = (-\infty, 0]$)[13] one has:

$$\frac{n\hat{y}_n}{n + \overline{n}_0} \leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0} \leq \frac{c + n\hat{y}_n}{n}, \tag{17}$$

which are obtained for $y_0 = 0$, $n_0 = \overline{n}_0$ (lower bound) and $n_0 = c/y_0$, $y_0 \to \infty$ (upper bound). For the case $\mathcal{Y} = (-\infty, +\infty)$, one has

$$\begin{aligned} \min\left(\frac{-c + n\hat{y}_n}{n + \overline{n}_0}, \frac{-c + n\hat{y}_n}{n}\right) &\leq y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0} \\ &\leq \max\left(\frac{c + n\hat{y}_n}{n + \overline{n}_0}, \frac{c + n\hat{y}_n}{n}\right), \end{aligned} \tag{18}$$

[13] Since $\hat{y}_n \geq 0$, it follows that $\frac{c+n\hat{y}_n}{n} \geq \frac{c+n\hat{y}_n}{n+\overline{n}_0}$ and, thus, $\frac{c+n\hat{y}_n}{n}$ is a upper bound for y_p. In the case, $\mathcal{Y} = (-\infty, 0]$, since $\hat{y}_n \leq 0$, it follows that $\frac{c+n\hat{y}_n}{n} \leq \frac{c+n\hat{y}_n}{n+\overline{n}_0}$ and, thus, $\frac{c+n\hat{y}_n}{n}$ is a lower bound for y_p.

which are obtained for $n_0 = c/|y_0|$, $y_0 \to -\infty$ *or* $n_0 = \overline{n}_0$, $y_0 = -c/\overline{n}_0$ *(left bound) and* $n_0 = c/|y_0|$, $y_0 \to +\infty$ *or* $n_0 = \overline{n}_0$, $y_0 = c/\overline{n}_0$ *(right bound).*

In all three cases, for $n > 0$ *and independently of the value of* $\hat{y}_n$, *at least one between the lower and upper bound differs from its a priori value* $\underline{E}[b']$ *or, respectively,* $\overline{E}[b']$. *This is obvious for the upper bound of (17) and the lower and upper bounds of (18), since a priori* $\underline{E}[b'] = -\infty$ *and* $\overline{E}[b'] = \infty$. *For the lower and upper bounds of (16), it follows that if* $\hat{y}_n = 0$ *then* $\underline{E}[b'|y^n] = \underline{E}[b'] = 0$ *but* $\overline{E}[b'|y^n] \neq \overline{E}[b'] = 1$ *and, vice versa, if* $\hat{y}_n = 1$ *then* $\overline{E}[b'|y^n] = \overline{E}[b'] = 1$ *but* $\underline{E}[b'|y^n] \neq \underline{E}[b'] = 0$. *Thus, the second part of (1) holds for any* $\delta > 0$. *This proves that* $n_0 \leq \overline{n}_0 < \infty$ *and* $n_0|y_0| \leq c$ *are necessary and sufficient for (A.3) to be satisfied.*

Since for $n_0 \leq \overline{n}_0 < \infty$, $n_0|y_0| \leq c$ *and* $n \to \infty$ *it follows that* $y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$ *and* $n_p = n + n_0$ *do not depend on* n_0 *and* y_0 *but only on* n *and* $\hat{y}_n$, *the sufficiency for (A.4) follows also straightforwardly. In fact it can be verified that, for* $n \to \infty$, *it follows that* $\underline{E}[g|y^n], \overline{E}[g|y^n] \to E^*[g|y^n]$, *where* $E^*[g|y^n]$ *is the posterior expectation obtained from the improper uniform prior density on* $\mathcal{W}$ *(*$p(w) = 1$*).*

To sum up, necessary and sufficient conditions for (A2)–(A4) are $\mathcal{Y}' = Int(\mathcal{Y})$ *and* $0 < n_0 < \min(\overline{n}_0, c/|y_0|)$ *and the corresponding set of priors* $\mathcal{M}$ *is also the largest set which satisfies (A2)–(A4). This proves the theorem for (A2)–(A4). Consider now property (A.1), coherence. Notice that, for each fixed value of the parameters* $0 < n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$ *and* $y_0 \in Int(\mathcal{Y})$, *the set of priors* $\mathcal{M}$ *includes only proper densities [14]. Thus, since the set of posteriors is obtained by applying Bayes' rule to each pair likelihood-prior in* $\mathcal{M}$, *the strong coherence of priors, likelihood and posteriors follows by the application of the lower envelope theorem [2, Theorem 7.8.1].*[14]

■

Some remarks on Theorem 1.

1. In order to better understand the intuition behind the theorem, consider the case in which the observations belong to $\mathbb{R}$ and the likelihood is a Gaussian density with known variance, so that $\mathcal{Y} = (-\infty, +\infty)$. The conjugate model under considerations is thus a Gaussian-Gaussian model (see Examples 2 and 4). In this case, the set of priors $\mathcal{M}$ is equal

[14] Walley' s theory is defined only for bounded gambles. An open question is whether the strong coherence of the model in Theorem 1 extends to the unbounded case.

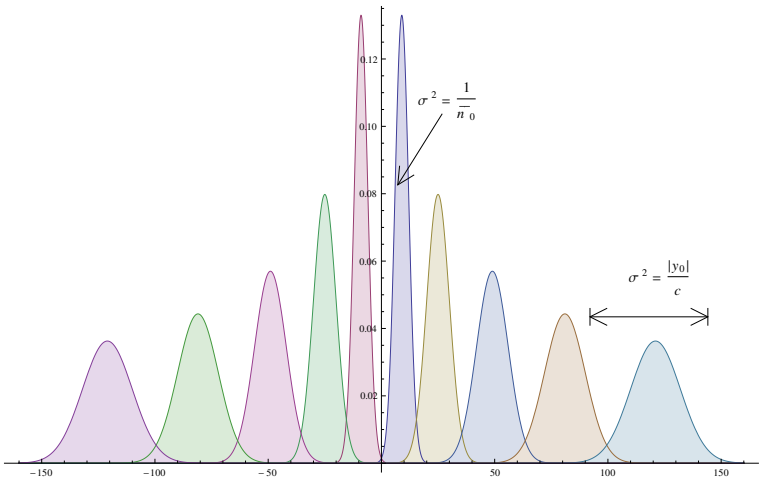

Figure 1: The set of priors $\mathcal{M}$ for the Gaussian-Gaussian conjugate model with $c = 1$ and $\overline{n}_0 = 1/9$.

to:

$$\Big\{\mathcal{N}\left(w; y_0, \sigma_0^2\right) : \quad y_0 \in (-\infty, +\infty), \\ \max(1/\overline{n}_0, |y_0|/c) < \sigma_0^2 < \infty\Big\}, \tag{19}$$

where y_0 is the prior mean and $\sigma_0^2 = 1/n_0$ the prior variance. Hence, $\mathcal{M}$ includes all the Gaussian densities with mean free to vary in $\mathbb{R}$ and variance lower bounded by $1/\overline{n}_0$ but linearly increasing with $|y_0|$. Figure 1 shows some densities belonging to $\mathcal{M}$. Notice, in fact, that if $|y_0| > c/\overline{n}_0$, then $\sigma_0^2 \geq |y_0|/c$. Hence, considering the likelihood $\mathcal{N}(y_i; w, \sigma^2)$ for $i = 1, \ldots, n$ and the derivations in Example 4, the corresponding set of posteriors is equal to:

$$\Big\{\mathcal{N}\left(w; y_p, \sigma_p^2\right) : \quad y_p = \sigma_p^2 \left(\frac{y_0}{\sigma_0^2} + \frac{n\hat{y}_n}{\sigma^2}\right), \\ \sigma_p^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}, \quad y_0 \in (-\infty, +\infty), \\ \max(1/\overline{n}_0, |y_0|/c) < \sigma_0^2 < \infty\Big\}, \tag{20}$$

where y_p is the posterior mean. Since $y_p = (n_0 y_0 + n\hat{y}_n)/(n + n_0)$ then, fixing $n_0 = 1/\sigma_0^2$, for $|y_0| \to \infty$ it follows that $|y_p| = |n_0 y_0 + n\hat{y}_n|/(n +$

$n_0) = |y_0| \to \infty$. Similarly, fixing y_0, for $n_0 \to \infty$ it follows that $|y_p| = |y_0|$. In other words, $n_0|y_0| = \infty$ implies a vacuous posterior mean and, thus, no learning and no convergence. Theorem 1 states that a necessary and sufficient condition to guarantee near-ignorance without preventing learning and convergence to take place is by imposing the constraint:

$$|n_0 y_0| < c < \infty,$$

which means that n_0 must in general depend on y_0.[15] In this case in fact for $|y_0| \to \infty$, it follows that $|y_p| = |n_0 y_0 + n\hat{y}_n|/(n + n_0) < \infty$. That is, the contribution of y_0 to y_p must decrease as $|y_0| \to \infty$, otherwise the observations do not contribute to y_p (learning cannot take place). This is essentially the meaning of the constraint $|y_0|/c < \sigma_0^2$ in (20), i.e., the variance of the Gaussians in $\mathcal{M}$ must be greater than $|y_0|/c$. Furthermore, $n_0 < \infty$ or, equivalently, the variance must also be greater than zero otherwise the Gaussian density would coincide with a Dirac delta; this is the reason for the constraint $\sigma_0^2 > 1/\overline{n}_0 > 0$. Under these constraints, it can be verified that y_p satisfies:

$$\begin{aligned} \min\left(\frac{-c + n\hat{y}_n}{n + \overline{n}_0}, \frac{-c + n\hat{y}_n}{n}\right) \leq \\ y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0} \leq \max\left(\frac{c + n\hat{y}_n}{n + \overline{n}_0}, \frac{c + n\hat{y}_n}{n}\right), \end{aligned} \tag{21}$$

and converges to $\hat{y}_n$ (maximum likelihood estimate) for $n \to \infty$ (convergence property (A.4)).

2. The family of priors $\mathcal{M}$ defined in Theorem 1 is completely determined by the two parameters $c > 0$ and $\overline{n}_0 > 0$ (actually just by $\overline{n}_0 > 0$ in the case $\mathcal{Y} = [0, 1]$). The larger these parameters are the larger the family of priors $\mathcal{M}$ is and, thus, the more conservative the posterior inferences are. The choice of these parameters will be discussed in Section 5.

3. In case the observations are binary, i.e., $\mathcal{Y} = [0, 1]$, the set of priors $\mathcal{M}$ transformed back to the original parameter space $\mathcal{X}$ reduces to

[15] Walter and Augustin [16] propose a functional relationship between n_0 and y_0 in the exponential families with a different aim w.r.t. that of the present paper; that is highlighting prior-data conflict in the case of inference drawn from a set of informative priors, i.e., near-ignorance is not satisfied.

the Imprecise Beta Model discussed by Walley [2, Section 5.3.1] and Bernard [17]:

$$\mathcal{M} = \Big\{ Beta(x; st, s(1-t)) : \quad t \in [0,1] \Big\}, \tag{22}$$

where $x \in (0,1)$, $n_0 = s = \overline{n}_0$ is a positive fixed value and $Beta(x; \alpha, \beta)$ is the Beta density with parameters α and β. The Imprecise Beta Model (22) and its multidimensional extension [18] have been applied effectively in classification [19] and system reliability problems [20].

4. In the case the observations belong to the real line then, by taking the limit for $\overline{n}_0 \to 0$ of the bounds in (21), one obtains the same posterior inferences as the ones derived from the family of Gaussian priors with infinite variance discussed by Pericchi and Walley [12, Section 3.3]. In fact, the bounds in (18) become equal to:

$$\frac{-c + n\hat{y}_n}{n} \leq \frac{n_0 y_0 + n\hat{y}_n}{n + n_0} \leq \frac{c + n\hat{y}_n}{n}. \tag{23}$$

The case $\overline{n}_0 \to 0$ is excluded by our model in Theorem 1, since we admit only proper priors in $\mathcal{M}$. Observe in fact that the set of Gaussian priors with infinite variance [12, Section 3.3] is a set of improper priors. The advantage of excluding improper priors is that the set $\mathcal{M}$ defined in Theorem 1 can be proved to be strongly coherent, while no proof of coherence is given for the model in [12, Section 3.3]; the coherence of this model is still an open problem. Another advantage is that prior expectations can be defined directly and no limit procedures are necessary. However, in some case, it can be convenient to consider small values of $\overline{n}_0$, i.e., $\overline{n}_0 \approx 0$, as it will be explained in the next Sections.

4.1. Additional properties for prior near ignorance

It is worth comparing properties (A.1)–(A.4) with the properties for prior near-ignorance discussed in [12] for the case $\mathcal{W} = \mathbb{R}$ and $\frac{db}{dw}(w) = w$. Hereafter, for the convenience of the reader, we summarize these properties.

(B) Prior Near Ignorance The following conditions on the upper and lower expectations generated by $\mathcal{M}$ are necessary for $\mathcal{M}$ to be sufficiently large.

(B1) If $A = [a - \zeta, a + \zeta] \subseteq \mathbb{R}$ is a finite interval for some $a \in \mathbb{R}$ and $\zeta > 0$, then $\underline{E}(I_{\{A\}}) = 0$ and $\overline{E}(I_{\{A\}}) \to 1$ as $2\zeta \to \infty$.

(B2) If $A = [a, +\infty)$ or $A = (-\infty, a]$ is a semi-infinite interval, then $\underline{E}(I_{\{A\}}) = 0$ and $\overline{E}(I_{\{A\}}) = 1$.

(B3) The upper and lower mean under $\mathcal{M}$ is $\underline{E}(W) = -\infty$ and $\overline{E}(W) = +\infty$.

(C) Translation invariance When there is no prior information on the value of W, the set $\mathcal{M}$ should be invariant under translations of the scale on which measurements of W are made. Equivalently, upper and lower expectations generated by $\mathcal{M}$ should be translation-invariant.

(D) Dependence on sample size Upper and lower probabilities of the standard γ-intervals should converge to the nominal probability γ as $n \to \infty$.[16]

Other properties for $\mathcal{M}$ defined in [12] are: easy elicitation, tractability, variety of shapes and weak coherence of posteriors and likelihood.

Concerning the coherence property (A.1) defined in Section 4, this is more restrictive than the property of weak coherence defined in [12]: (A.1) implies weak coherence, while the converse is not true. We point the reader to [2, Ch. 7] for more details. Concerning properties (B1)–(B3), they can be obtained from (A.2) for different choices of g (the meaning of the limit in (B1) will be discussed later). Hence, (A.2) includes (B1)–(B3) as special cases. However, notice that the functions g that we use to model our state of prior near-ignorance about W are usually those listed in (B1)–(B3), i.e., W and $I_{\{A\}}$, where A can be a finite interval or a semi-infinite interval (for $A = (-\infty, a]$, $E[I_{\{A\}}]$ coincides the cumulative distribution of w). Thus, (B2) and (B3) are exactly the property (A.2) defined in Section 2 in case $\mathcal{G}_0 = \{W, I_{\{A\}}\}$. Concerning property (B1) if, for a moment, we forgot the limit condition $2\zeta \to \infty$ in the upper and we assumed $\overline{E}(I_{\{A\}}) = 1$ to hold for any ζ, (B1) would again coincide with (A.2) by setting $g = I_{\{A\}}$, i.e., $\underline{E}(I_{\{A\}}) = \inf I_{\{A\}} = 0$ and $\overline{E}(I_{\{A\}}) = \sup I_{\{A\}} = 1$ for any ζ. However, if

[16] For Gaussian densities, the standard γ-intervals are $[\mu - z_\gamma \sigma^2, \mu + z_\gamma \sigma^2]$, where μ, σ^2 are the posterior mean and variance and $z_\gamma \in \mathbb{R}^+$ depends on γ. For sets of densities, we can define the γ-interval as the smallest interval whose probability of enclosing w is at least γ.

(A.2) holds for $g = I_{\{A\}}$ with A arbitrary, then (A.3) cannot be satisfied. In fact, the requirement $\overline{E}(I_{\{A\}}) = 1$ implies that $\mathcal{M}$ must include a proper density with support in A. However, since A can be arbitrarily small, this density can be arbitrarily close to a Dirac delta centered in A. Thus, $\mathcal{M}$ must include densities approaching all possible Dirac's delta in the real line. Intuitively this implies that, no matter the observations, the posterior upper and lower expectations of $g = I_{\{A\}}$ would coincide with the prior upper and lower expectations and, thus, (A.3) cannot be satisfied. (B1) is thus a way to relax the requirement of prior ignorance w.r.t. $g = I_{\{A\}}$ with $A = [a-\zeta, a+\zeta]$, which is also compatible with (A.3).

Corollary 1. *Under the hypotheses of Theorem 1 and with $\mathcal{W} = \mathcal{R}$, the family of priors $\mathcal{M}$ satisfies (B1) and (B2).* ■

Proof: *Concerning (B1), consider the case in which $A = [w_0 - \zeta, w_0 + \zeta]$, for some $w_0 \in \mathcal{W}$ and $\zeta < \infty$. The expected value w.r.t. w of the indicator $I_{[w_0-\zeta,w_0+\zeta]}$ is given by*

$$P(A) = \int_{\mathcal{W}} I_{[w_0-\zeta,w_0+\zeta]}(w)p(w|n_0,y_0)dw = \int_{w_0-\zeta}^{w_0+\zeta} k(n_0,y_0)\exp(n_0(y_0w-b(w)))dw.$$

It is clear that for $\zeta \to \infty$, $\overline{E}\left[I_{[w_0-\zeta,w_0+\zeta]}\right] \to 1$, as $p(w|n_0,y_0)$ is a proper density for some $0 < n_0 \leq \overline{n}_0$ and $y_0 \in Int(\mathcal{Y})$. In order to prove that $\underline{E}\left[I_{[w_0-\zeta,w_0+\zeta]}\right] = 0$ for any $\zeta < \infty$, consider the first and second derivatives of $p(w|n_0,y_0)$, i.e., for any $w \in \mathcal{W}$,

$$\begin{aligned}
\frac{dp(w|n_0,y_0)}{dw} &= n_0\left(y_0 - \frac{db}{dw}(w)\right)p(w|n_0,y_0), \\
\frac{d^2p(w|n_0,y_0)}{dw^2} &= -\frac{d^2b}{dw^2}(w)n_0p(w|n_0,y_0) + n_0^2\left(y_0 - \frac{db}{dw}(w)\right)^2 p(w|n_0,y_0).
\end{aligned} \tag{24}$$

When $\frac{db}{dw}(w) = y_0$ the first derivative is zero and the second derivative is negative (as $p(w|n_0,y_0) > 0$ and $\frac{d^2b}{dw^2}(w) > 0$; the positiveness of $\frac{d^2b}{dw^2}$ follows from the fact that it is equal to the variance of Y see (6)), the value w^* such

that $\frac{db}{dw}(w) = y_0$ *is a maximum of* $p(w|n_0, y_0)$*. Since* $\frac{db}{dw}(w), y_0 \in Int(\mathcal{Y})$ *and since* $\frac{d^2b}{dw^2}(w) > 0$*, the function* $\frac{db}{dw}(w)$ *is always increasing in* $\mathbb{R}$ *and obtains its infimum* $\inf \mathcal{Y}$ *and supremum* $\sup \mathcal{Y}$ *for* $w \to \pm\infty$*. Thus, since* $\frac{db}{dw}$ *and* $p(w|n_0, y_0)$ *are continuous in* w *it follows that by moving* y_0 *between* $\inf \mathcal{Y}$ *and* $\sup \mathcal{Y}$*,* w^* *can be shifted arbitrarily in* $\mathbb{R}$*. Then assume, by absurd, that*

$$\int_{w_0-\zeta}^{w_0+\zeta} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw \geq \delta \tag{25}$$

for all y_0*, where* $\delta > 0$ *is some (arbitrarily small) constant. Then, in the case* $w^* = w_0 + 3\zeta$*, it follows that*

$$\begin{aligned}
&\int_{-\infty}^{\infty} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw \\
&> \int_{w_0-\zeta}^{w_0+\zeta} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw + \int_{w_0+\zeta}^{w_0+3\zeta} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw \\
&> 2 \int_{w_0-\zeta}^{w_0+\zeta} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw \geq 2\delta,
\end{aligned}$$

in fact, since $w^* = w_0 + 3\zeta$ *is a maximum of* $p(w|n_0, y_0)$*,* $p(w|n_0, y_0)$ *is increasing in* $(-\infty, w^*)$ *and, thus,*

$$\int_{w_0-\zeta}^{w_0+\zeta} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw \leq \int_{w_0+\zeta}^{w_0+3\zeta} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw.$$

Hence, in general if $w^* = w_0 + (2r+1)\zeta$ *with* $r = 1, 2, 3, \ldots$*, it follows that*

$$\int_{-\infty}^{\infty} k(n_0, y_0) \exp(n_0(y_0 w - b(w)))dw \geq (r+1)\delta.$$

Thus, being $\delta > 0$*, if (25) holds and since* w^* *is free to vary in* $\mathbb{R}$*, we can always find a value of* w^* *such that* $\int_{-\infty}^{\infty} p(w|n_0, y_0) > 1$*, which contradicts* $p(w|n_0, y_0)$ *being a PDF. Therefore, for* $y_0 \to \sup \mathcal{Y}$*, it must hold that* $\delta \to 0$ *and thus* $p(w|n_0, y_0) \to 0$ *in* $[w_0 - \zeta, w_0 + \zeta]$*. A similar conclusion can be derived by moving* y_0 *towards* $\inf \mathcal{Y}$*. This implies that* $P(A) = 0$ *for*

$y_0 \to \sup \mathcal{Y}$ (or $y_0 \to \inf \mathcal{Y}$), which proves (B1) w.r.t. w. For the lower, the same result can be proven for any semi-infinite interval such as $A = [w_0, +\infty)$ or $A = (-\infty, w_0]$. Furthermore, since $\underline{E}[I_{\{A\}}] = 0$ implies $\overline{E}[I_{\{\mathbb{R}\setminus A\}}] = 1$ [2], (B2) follows straightforwardly.

■

To be invariant to re-parametrizations [1, Sec. 5] of the parameter space is a desirable property for a model of prior ignorance. Since, in this paper, we consider the one-parameter exponential family, we have formulated the prior near-ignorance property w.r.t. the natural parameter w. This means that our prior near-ignorance property holds only for a specific parametrization of the parameter space, i.e., the one which transforms the original parameter $x \in \mathcal{X} \subseteq \mathbb{R}$ into the natural parameter $w \in \mathcal{W} \subseteq \mathbb{R}$. The advantage of this approach is that likelihood and prior, under this parametrization, belong to the natural exponential family and this considerably simplifies the computations of the posterior inferences (e.g., the expectations of some functions of interest can be computed in closed form).

Observe that in case $\mathcal{W} = \mathbb{R}$, we are considering parametrizations that transform x in a location parameter (in the Gamma density case, x is a scale parameter ($x > 0$) while in the Beta density case x is an unknown chance ($x \in [0, 1]$), but for both cases $\mathcal{W} = \mathbb{R}$). Hence, since w is defined in $\mathbb{R}$, the invariance property that we would like to respect on $\mathcal{W}$ is translation invariance. This means that for all $g : \mathcal{W} \to \mathbb{R}$, $w_0 \in \mathcal{W}$ and $g_t(w) = g(w - w_0)$, the family $\mathcal{M}$ is translation invariant if $\underline{E}[g] = \underline{E}[g_t]$. Since this property must hold for each function g, it is satisfied if for each $w_0 \in \mathcal{W}$, $n_0 \in \mathcal{A}_{y_0}$ and $y_0 \in \mathcal{Y}_0$ there exist $\tilde{n}_0 \in \mathcal{A}_{y_0}$ and $\tilde{y}_0 \in \mathcal{Y}_0$ such that:

$$\exp\left(n_0\left(y_0(w - w_0) - b\left(w - w_0\right)\right)\right) \propto \exp\left(\tilde{n}_0\left(\tilde{y}_0 w - b(w)\right)\right). \tag{26}$$

The validity of the above equality depends on the functional form of b and, thus, on the particular member of the exponential family under consideration. In particular, (26) can be satisfied if the member of the exponential family has a location parameter [21]. From the results by Ferguson [21] for the one-parameter exponential family, it follows that the densities which have a location parameter are either Gamma or Gaussian densities. For this family, translation invariance can be satisfied only if the parameters of the densities y_0 and n_0 are free to vary independently. This is in fact the only way to satisfy (26) for each pair n_0 and y_0. In the case considered in the present

paper, since y_0 and n_0 are not free to vary independently (because of the constraint $n_0 \leq \min(\overline{n}_0, \frac{c}{|y_0|})$) translation invariance cannot be guaranteed in general.

However, it can be noticed that, at the decreasing of $\overline{n}_0$, the lower and right side member in (26) become closer and closer and, thus, $\mathcal{M}$ gets closer and closer to a translation invariant model. In fact, as observed in Section 4, in the case $\mathcal{Y} = (-\infty, +\infty)$ and for $\overline{n}_0 \to 0$, the set of priors $\mathcal{M}$ reduces to the family of Gaussian priors with infinite variance discussed in [12, Section 3.3], which is indeed translation invariant. Also in the case $\mathcal{Y} = [0, +\infty)$, the set of priors $\mathcal{M}$ reduces to a translation invariant model for $\overline{n}_0 \to 0$.[17] The same happens for the case $\mathcal{Y} = [0, 1]$ as even then, for $\overline{n}_0 \to 0$, the set of priors $\mathcal{M}$ collapses to a single improper prior (since $\overline{n}_0 \to 0$ is not contrasted by $|y_0| \to \infty$).

As discussed by Pericchi and Walley [12, Section 3.3], an important consequence of translation invariance is that the imprecision (that is, the size) of the posterior set of densities does not depend on the sample mean $\hat{y}_n$; the only effect of shifting $\hat{y}_n$ is to shift the posterior set by the same amount. Translation invariance is a sufficient condition for the independence to the sample mean $\hat{y}_n$ of the imprecision of the posterior set of densities, but it is not necessary. In fact, there are families $\mathcal{M}$ that are not translation invariant but for which this independence holds such as, for instance, the family $\mathcal{M}$ in Theorem 1 in the case $\mathcal{Y} = [0, 1]$ and $\overline{n}_0 > 0$, see the bounds in (16). This independence is used in Section 5 to choose the value of the parameter c of the set of priors $\mathcal{M}$ based on the desirable imprecision of the posterior set of densities.

Concerning property (D), this is a condition for learning, convergence of the lower and upper probability to the nominal probability γ. It states that, with increasing n, the effects of the lack of prior information should become less and less important (learning) and that lower and upper probabilities should converge to the nominal probability γ (convergence). Property (D) is thus similar to the properties (A.3)–(A.4) defined in Section 2. Notice that

[17] The kernels $\exp(n_0(y_0 w - b(w))$ of the priors in $\mathcal{M}$ tend to the set of improper kernels $\{\exp(\ell w) : \ \ell \in [0, c]\}$, which implies that a-priori $\overline{E}[g] = 0$ for all gambles g that are absolutely integrable w.r.t. $\exp(\ell w)$. $\overline{E}[g] = 0$ follows from the fact that the normalization constant goes to zero for $\overline{n}_0 \to 0$ and $y_0 \to 0$ or $y_0 \to \infty$, and $\int g(w) \exp(\ell w) dw < \infty$ since g is absolutely integrable w.r.t. $\exp(\ell w)$.

there cannot be convergence without learning. Thus, (D) requires (A.3). The convergence defined in (A.4) is more general than that in (D). In [12], it is in fact assumed that there is a single likelihood distribution which describes the observation model and, thus, with increasing number of observations we expect that lower and upper posterior expectations get closer and closer. The convergence to the nominal probability γ is then a consequence of the central limit theorem for distributions. Conversely, property (A.4) only states that the influence of the prior set of distributions should vanish with increasing number of observations. However, in the case of the exponential family discussed in Section 4, it can easily be proved that not only (A.4) but also (D) holds.

Corollary 2. *Under the hypotheses of Theorem 1, the family of priors $\mathcal{M}$ satisfies (D).* ■

Proof: *Consider the quantity*

$$y_p = \frac{n_0 y_0 + n\hat{y}_n}{n + n_0}.$$

Assuming the constraint $n_0 < \min(\overline{n}_0, \frac{c}{|y_0|})$, then $y_p \to \hat{y}_n$ and $n_p = n + n_0 \to n$ as $n \to \infty$. Thus, the family of posterior densities converges to

$$p(w|n, \hat{y}_n) \approx k(n, \hat{y}_n) \exp(n(\hat{y}_n w - b(w))), \ w \in \mathcal{W}. \tag{27}$$

Hence, for any function g, the convergence of the lower towards the upper expectation of g follows straighforward. The convergence of the upper and lower probabilities of the standard γ-intervals to the nominal probability γ is then a consequence of the central limit theorem. ■

The family of priors $\mathcal{M}$ resulting from Theorem 1 is also easy to elicit and tractable. About elicitation only two parameters $(\overline{n}_0, c)$ must be specified by the modeller. Guidelines for the choice of these parameters are given in Section 5. Tractability is instead guaranteed by conjugacy as it will be shown in Section 5.

4.2. Comparison with other models for ignorance

It is also interesting to compare the set of priors $\mathcal{M}$ in Theorem 1 with another model for ignorance, the Bounded Derivative Model (BDM) [22]. In

the BDM, $\mathcal{M}_{BDM}$ includes all continuous proper probability density functions for which the derivative of the log-density is bounded by a positive constant, i.e., $\left|\frac{d}{dw}\log p(w)\right| \leq c$ for all $w \in \mathcal{W}$. It can be verified that BDM satisfies all the properties (A1)–(A4), with $\mathcal{G}_0$ and $\mathcal{G}$ defined as in Theorem 1. BDM is a non-parametric model and, in this sense, is more general than the model resulting from Theorem 1 that is restricted to the one-parameter exponential family only. A drawback of this generality is that inferences with BDM can in general be difficult to compute [22, Sec. 6], while this is often not the case for the model resulting from Theorem 1 because of conjugacy.

Conversely, a model for statistical inferences for exponential family likelihoods is presented by Quaeghebeur and De Cooman [23, Ch.4], [24]. The main difference with respect to the present work is that the model in [24] is not a model of prior ignorance, as pointed out by the authors, i.e., the set $\mathcal{Y}'$ in Theorem 1 is chosen in [24] to reflect the prior information on y_0 and, thus, the posterior inferences depend on this information. Since no constraint between n_0 and y_0 is assumed, the model in [24] can also violate (A.3)–(A.4) in the case $\mathcal{Y}' = Int(\mathcal{Y})$, and hence it can produce vacuous inferences. The case $\mathcal{Y}' = Int(\mathcal{Y})$ is thus excluded by the authors, see [23, pag. 175 (ii),(iii)]. Quaeghebeur [23, pag. 176 (vi)] motivates the choice of keeping n_0 fixed as follows: (i) n_0 represents an hypothetical number of observations determining the learning speed and (ii) it is mathematical convenient because it makes the computations simpler. In this paper, we have advocated that, in case of lack of prior information, one should let y_0 to vary in $\mathcal{Y}' = Int(\mathcal{Y})$, which allows to satisfy the prior ignorance property (A.2). Hence, in order to also satisfy learning (A.3) and convergence (A.4), n_0 must depend on $|y_0|$. In our case the learning speed is determined by the constants c and $\overline{n}_0$ About tractability the fact that n_0 depends on $|y_0|$ does not increase significatively the computational cost w.r.t. the model discussed by Quaeghebeur and De Cooman.

Boratynska [25] defines a set of priors by specifying bounds for the product $n_0 y_0$. However, since n_0 is kept constant, this is equivalent to define bounds for y_0, as in the work of Quaeghebeur and De Cooman. Thus also this set of priors is not a model of prior ignorance. However it can be used to draw robust inferences in case the prior information only allows to specify bounds for the value of y_0. Based on the bounds for y_0, Boratynska in fact derives robust estimates (stable and Γ-minimax) of w.

5. How to choose $\overline{n}_0$ and c

From Theorem 1, it follows that the family of priors $\mathcal{M}$ is completely determined by the two parameters $\overline{n}_0$ and c. The aim of this section is to give guidelines for the choice of these parameters. First of all, notice that larger values of $\overline{n}_0$ and c imply larger sets of priors $\mathcal{M}$ and, thus, more robustness to the lack of prior information in the posterior inferences. Thus, measures of robustness can be used to select $\overline{n}_0$ and c such as proposed by Walley [2]: (a) the convergence rate of the lower and/or upper expectations to suitable limits; (b) the convergence rate of the posterior imprecision, i.e., the difference between upper and lower expectations. Here the expectations are computed w.r.t. some function of interest g and the convergence is defined w.r.t. the number of samples n.

Another possible requirement for the choice of $\overline{n}_0$ and c is that the family of priors $\mathcal{M}$ should be large enough to encompass frequentist or objective Bayesian inferences, but not too large to avoid obtaining too weak inferences.

As in the previous sections we distinguish three cases $\mathcal{Y} = [0, 1]$: $\mathcal{Y} = (-\infty, \infty)$ and $\mathcal{Y} = [0, \infty)$ (or $\mathcal{Y} = (-\infty, 0]$). Consider the case $\mathcal{Y} = [0, 1]$. Since $\mathcal{Y}$ is finite, it follows that $c/|y_0|$ is bounded below by c. Then, by selecting $c > \overline{n}_0$, the set of priors $\mathcal{M}$ depends only on $\overline{n}_0$ (remember in fact that in this case $0 < n_0 \leq \overline{n}_0 < \infty$ is a necessary and sufficient condition for learning and convergence). Thus, only the parameter $\overline{n}_0$ must be specified. As already discussed in Section 4, in the case $\mathcal{Y} = [0, 1]$ the set of priors $\mathcal{M}$ coincides with the Imprecise Beta Model. The choice of the parameter $\overline{n}_0 = s$ for this model is widely discussed by Bernard [17] and Walley [18]. Several arguments support the choice of $\overline{n}_0 = 2$. In fact, in this case, the Imprecise Beta Model encompasses the three Beta densities that are commonly used by Bayesians to model prior ignorance about an unknown chance: the uniform prior ($n_0 = 2$ and $y_0 = 1/2$), Jeffreys' prior ($n_0 = 1$ and $y_0 = 1/2$) and Haldane's prior ($n_0 = 0$). Furthermore, for $\overline{n}_0 = 2$, it is also guaranteed that the 95% one- and two-sided credible intervals for x are also (at least) 95% confidence intervals in the frequentist sense [18].

We can also use the posterior imprecision, i.e., $\overline{E}[b'] - \underline{E}[b']$, which is equal to $\overline{n}_0/(n + \overline{n}_0)$, to select $\overline{n}_0$. Thus, by selecting for instance $\overline{n}_0 = 1$, we can impose that the posterior imprecision reduces to $1/2$ its initial value after $n = 1$ observations (for $\overline{n}_0 = 2$ the imprecision reduces to $2/3$ its initial value after $n = 1$ observations, etc.). Smaller values of $\overline{n}_0$ produces faster convergence and stronger conclusions, whereas larger values produces more

cautious inferences.

In case $\mathcal{Y} = (-\infty, \infty)$ it follows from Equation (18) that for b' the posterior imprecision is equal to:

$$\begin{cases} \dfrac{2c}{n} & \text{if } |n\hat{y}_n| < c, \\ \dfrac{2c}{n} + \dfrac{\overline{n}_0}{n+\overline{n}_0}\left(\hat{y}_n - \dfrac{c}{n}\right) & \text{if } n\hat{y}_n - c > 0, \\ \dfrac{2c}{n} + \dfrac{\overline{n}_0}{n+\overline{n}_0}\left(-\hat{y}_n - \dfrac{c}{n}\right) & \text{if } n\hat{y}_n + c < 0. \end{cases} \tag{28}$$

Thus, apart from the first case, the imprecision does in general depend on the observations through $\hat{y}_n$. A way to overcome this problem is by selecting a small value for $\overline{n}_0$. In fact, when $\overline{n}_0$ approaches zero, the part of the imprecision that depends on $\hat{y}_n$ vanishes and, thus, the imprecision reduces to $2c/n$. Hence, as in the case $\mathcal{Y} = [0, 1]$, we can select a-priori the width of the posterior imprecision by fixing a suitable value of c. Notice that in the case $\mathcal{Y} = (-\infty, \infty)$, both the frequentist and noninformative Bayesian inferences are encompassed by $\mathcal{M}$ for each $c > 0$. In fact, the inferences obtained by Jeffreys' (improper uniform) prior are included in the set of posteriors inferences obtained by $\mathcal{M}$ for each $c > 0$.

In the case $\mathcal{Y} = [0, \infty)$, from (17) it follows that

$$\underline{E}[b'] = \frac{n\hat{y}_n}{n+\overline{n}_0}, \quad \overline{E}[b'] = \frac{c + n\hat{y}_n}{n}. \tag{29}$$

Thus, the posterior imprecision is equal to

$$\frac{c}{n} + \frac{\overline{n}_0}{n+\overline{n}_0}\hat{y}_n. \tag{30}$$

In this case, the posterior imprecision depends on $\hat{y}_n$ apart from the case $\overline{n}_0 \approx 0$ where it goes to c/n. However, for $\overline{n}_0 \approx 0$, the lower expectation in (29) reduces to $\hat{y}_n$ and, thus, the lower bound is always equal to $\hat{y}_n$. Thus, in the case $\overline{n}_0 \approx 0$, the posterior imprecision is strongly asymmetric w.r.t. $\hat{y}_n$. However, it is interesting to observe that, for $\overline{n}_0 \approx 0$ and $c = 1$, the set of posteriors obtained by $\mathcal{M}$ includes the two posteriors densities that are obtained by the two priors which are commonly used by Bayesians to model prior ignorance about a Poisson parameter: the positive uniform density ($n_0 = 0$ and $n_0 y_0 = 1$) and Jeffreys' prior ($n_0 = 0$ and $n_0 y_0 = 1/2$).

To sum up, in the case $\mathcal{Y} = (-\infty, \infty)$, for inferences in problems where there is almost no prior information about the parameters, we can consider the set of posteriors that one obtains for $\overline{n}_0 > 0$ but suitably small.[18] In fact, in this case, the set of priors $\mathcal{M}$ is close to a translation invariant model (as discussed in Section 4.1), encompasses frequentist and objective Bayesian inferences and produces self-consistent inferences (i.e., it satisfies (A.1)). The same holds in the case $\mathcal{Y} = [0, \infty)$ if c is selected to encompass the two noninformative priors which are commonly used by Bayesians to model prior ignorance, i.e., $c > 1$. However, in this case, also a non-negligible value of $\overline{n}_0$ seems to be convenient. In fact, a value of $\overline{n}_0$ that is not too small avoids that the lower bound of (29) reduces to $\hat{y}_n$, guaranteeing thus more robust inferences w.r.t. the maximum likelihood estimator $\hat{y}_{ML} = \hat{y}_n$.

The price to be paid for this increase in robustness is the need of specifying two (in some case just one) parameters. This is the main difference w.r.t. frequentist and Bayesian inferences based on improper priors that require no parameter elicitation (frequentist inferences or objective Bayesian inferences employing the improper uniform prior $p(w) = 1$ in the Gaussian case) or the elicitation of just one parameter (objective Bayesian employing the Jeffreys' prior in the Poisson case or the uniform or Jeffreys' prior in the Bernoulli case). However, observe that to model prior ignorance we need the same number of parameters necessary to specify an informative prior with the advantage that our inferences are robust in case of lack of prior information.

6. Application: inferences for a Poisson model

The Poisson distribution belongs to the one-parameter exponential family and it is used to count the number of occurrences of rare events which are occurring randomly through time (or space) at a constant rate. Suppose we have a random sample of n independent observations from a Poisson

[18] With suitably small, we mean close to the precision of the machine where the computations are performed. Another possibility could be to consider the posterior inferences obtained for $\overline{n}_0 \to 0$. However, for $\overline{n}_0 \to 0$, the priors in $\mathcal{M}$ becomes improper and, thus, we cannot use the lower envelope theorem to prove (A.1) as shown in the proof of Theorem 1. This means that inferences obtained for $\overline{n}_0 \to 0$ may be incoherent. A question to be addressed in future is whether the posterior model obtained for $\overline{n}_0 \to 0$ and likelihood model are coherent.

distribution with parameter x. The Poisson density is:

$$p(y^n|x) = \prod_i \frac{x^{y_i}\exp(-x)}{y_i!} \propto x^{n\hat{y}_n}\exp(-nx),$$

where y_i is the number of occurrences of an event in the i-th observation and $n\hat{y}_n = \sum_i y_i$. It is well known [1, Sec. 5.2] that the conjugate prior of a Poisson density is the Gamma density $p(x|\alpha,\beta) \propto x^{\alpha-1}\exp(-\beta x)$ with $\alpha,\beta \in \mathbb{R}^+$, which belongs to the exponential family. Hence, likelihood and prior can be rewritten in the canonical form

$$p(y^n|w) \propto \exp\left(n\left(\hat{y}_n w - b(w)\right)\right), \qquad p(w|n_0,y_0) \propto \exp\left(n_0\left(y_0 w - b(w)\right)\right), \tag{31}$$

where $n_0 = \beta$, $y_0 = \alpha/\beta$, $w = \ln(x)$, $w \in \mathcal{W} = \mathbb{R}$ and $b(w) = \exp(w) = x$.

In the case no prior information is available about w, the following two densities are used by Bayesians to model prior ignorance about the Poisson parameter: the positive uniform density ($\beta = n_0 = 0$ and $\alpha = n_0 y_0 = 1$) and Jeffreys' prior ($\beta = n_0 = 0$ and $\alpha = n_0 y_0 = 1/2$). Although both densities are improper, the posteriors obtained from these two priors are proper.

In the following we compare the frequentist and Bayesian inferences (the latter obtained using the above noninformative priors) with the inferences produced by the model in Section 4, which considers the following set of priors:

$$\mathcal{M} = \Big\{p(w|n_0,y_0):\ y_0 \in (0,+\infty),\ 0 < n_0 \le \min(\overline{n}_0, c/|y_0|)\Big\}. \tag{32}$$

As demonstrated in Section 4 and 4.1, the set $\mathcal{M}$ includes only proper priors and is really a model of prior ignorance w.r.t. the functions that are commonly used for statistical inferences (mean and credible intervals). The set of posteriors resulting from (32) is:

$$\mathcal{M}_p = \Big\{p(w|n_p,y_p) : y_p = (n_0 y_0 + n\hat{y}_n)/(n+n_0),\ n_p = n+n_0,\\ y_0 \in (0,+\infty),\ 0 < n_0 \le \min(\overline{n}_0, c/|y_0|)\Big\}. \tag{33}$$

Figures 2–3 show the set of priors and posteriors for the particular case $c = 2$, $\overline{n}_0 = 1$, $n = 5$ and $\hat{y}_n = 5$. The sets of priors and posteriors in (31)–(32) can be used to derive prior and posterior inferences as described below.

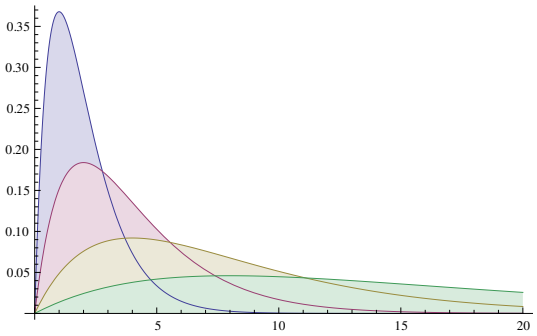

Figure 2: The set of priors $\mathcal{M}$ for $c = 2, \overline{n}_0 = 1$ and $y_0 = 2, 4, 8, 16$ (from left to right).

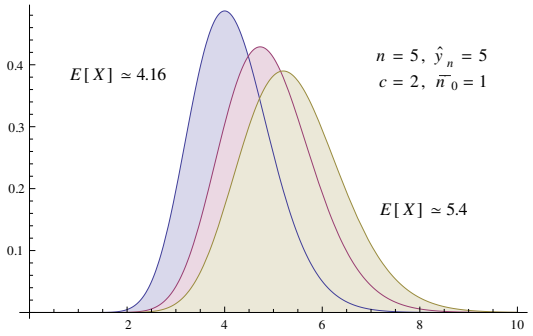

Figure 3: The set of posteriors for $c = 2$, $\overline{n}_0 = 1$, $n = 5$, $\hat{y}_n = 5$ and $y_0 = 0, 4, \infty$ (from left to right).

Mean: Prior lower and upper expectations of X are zero and, respectively, ∞. This follows from Theorem 1, since $E[b'] = y_0$. Posterior expectations of X are included in the interval $[n\hat{y}_n/(n+\overline{n}_0), (n\hat{y}_n + c)/n]$, see Figure 3 for the case $c = 2$, $\overline{n}_0 = 1$, $n = 5$, $\hat{y}_n = 5$. The lower expectation is less than the maximum likelihood estimator of X for any $\overline{n}_0 > 0$, while the upper, for $c > 1$, is greater than the Bayesian posterior means obtained with the positive uniform and Jeffreys' improper priors.

Variance: Prior lower and upper variance are defined as follows [2, Appendix G]:

$$\underline{V} = \inf_{\substack{y_0 \in (0,+\infty), \\ 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|)}} \frac{y_0}{n_0} = 0, \tag{34}$$

$$\overline{V} = \sup_{\substack{y_0 \in (0,+\infty), \\ 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|)}} \frac{y_0}{n_0} = \infty, \tag{35}$$

being $V = y_0/n_0$ the variance of a Gamma density. Notice that (34)–(35) is again a state of complete ignorance.[19] Posterior lower and upper variances are:

$$\underline{V} = \inf_{\substack{y_0 \in (0,+\infty) \\ 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|)}} \frac{y_p}{n_p} = \frac{n\hat{y}_n}{(n+\overline{n}_0)^2}, \tag{36}$$

$$\overline{V} = \sup_{\substack{y_0 \in (0,+\infty) \\ 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|)}} \frac{y_p}{n_p} = \frac{n\hat{y}_n + c}{n^2}, \tag{37}$$

which are, for $c > 1$, bounds for the Bayesian variances with noninformative priors.

One-sided hypothesis testing: Prior lower and upper probabilities of the hypotheses

- H_0: the Poisson population parameter $X \leq x_0$,
- H_1: the Poisson population parameter $X > x_0$,

[19] The fact that $\mathcal{M}$ is a model of prior ignorance for the variance is true in the specific case of Gamma-like densities, but it is not true in general, e.g., for Gaussian densities.

are $\underline{P}(H_i) = 0, \quad \overline{P}(H_i) = 1, \quad i = 0, 1$. This follows from the property (B2) discussed in Section 4.1 and reflects our state of prior ignorance about the probability of H_i. Lower and upper posterior probabilities of H_0 are:

$$\begin{aligned}
\underline{P}(H_0|y^n) &= \min_{\alpha\in[0,\overline{n}_0]} \int_{-\infty}^{\ln(x_0)} k\left(n+\alpha, \frac{c+n\hat{y}_n}{n+\alpha}\right) \exp\left((n+\alpha)\left(\frac{c+n\hat{y}_n}{n+\alpha}w - b(w)\right)\right) dw \\
\overline{P}(H_0|y^n) &= \int_{-\infty}^{\ln(x_0)} k(n''', y''') \exp(n'''(y'''w - b(w)))dw,
\end{aligned}$$

where $y''' = \frac{n\hat{y}_n}{n+\overline{n}_0}$, $n''' = n+\overline{n}_0$ (which is obtained for $y_0 = 0$). Figure 4 shows the value of $\underline{P}(H_0|y^n)$ as a function of n in the case $\hat{y}_n = 0.75$, $\overline{n}_0 = 10^{-6}$, $x_0 = 1$ and for three different values of c, i.e., $\{0.5, 1, 2\}$. The lower probabilities $\underline{P}(H_0|y^n)$ for $c = 0.5$ and $c = 1$ coincide with the Bayesian probabilities $P(H_0|y^n)$ obtained with Jeffreys and, respectively, uniform priors. It can be noticed that, based on Bayesian inferences, the hypothesis H_1 can be rejected after just 2 observations, being $P(H_0|y^n) > 0.5$ for $n \geq 2$ (for the uniform prior, while for Jeffreys' prior this already happens at $n = 1$). Conversely, by using the method proposed in this paper, the hypothesis H_1 can be rejected only after $n \geq 7$ observations in the case $c = 2$, i.e., when $\underline{P}(H_0|y^n) > 0.5$. Figure 5 shows the value of $\underline{P}(H_0|y^n)$ as a function of y^n in the case $n = 1$, $\overline{n}_0 = 10^{-6}$, $x_0 = 1$ and for the previous three values of c. It can be noticed again that, based on Bayesian inferences, the hypothesis H_0 is rejected after $n = 1$ observation if $y > 1$ (for Jeffreys' prior, while for the uniform prior this already happens at $y = 0.3$). Conversely, by using the method proposed in this paper, H_0 is rejected for $y > 1.5$, i.e., when $\overline{P}(H_0|y^n) < 0.5$.

Credible interval: A $100(1-\gamma)\%$ credible interval is the smallest interval $\mathcal{X}$ that has at least probability $(1-\gamma)$ of including the true x, i.e., $\underline{E}[I_{\{\mathcal{X}\}}] > (1-\gamma)$. For the model $\mathcal{M}$, it can be verified that $\mathcal{X} = (0, \infty)$ for any $\gamma > 0$, which is again a state of complete ignorance. This follows from the same arguments used to prove property (B1) discussed in Section 4.1. Credible intervals are also used for two-sided hypothesis testing:

- H_0: the Poisson population parameter $X = x_0$;
- H_1: the Poisson population parameter $X \neq x_0$.

The hypothesis H_0 is rejected with probability $(1-\gamma)$ if x_0 is not included in the $100(1-\gamma)\%$ credible interval. Since $\mathcal{X} = (0, \infty)$, no hypothesis $x = x_0$ can be rejected a-priori for any $\gamma > 0$. The posterior $100(1-\gamma)\%$ credible interval for x is

$$\begin{aligned}\mathcal{X} = & \bigcup \Big\{ [w_1, w_2] : \int_{w_1}^{w_2} p(w|n_p, y_p)\ dw = 1-\gamma, \\ & y_p = (n_0 y_0 + n\hat{y}_n)/(n+n_0),\ n_p = n + n_0, \\ & y_0 \in (0, +\infty),\ 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|) \Big\}, \end{aligned} \tag{38}$$

where, for each n_p, y_p, only the intervals with minimum size are considered in (38). In the case $\overline{n}_0 \approx 0$ and suitably large n, the left extreme of $\mathcal{X}$ is approximatively equal to the value w_1 obtained for $y_p = \hat{y}_n$, $n_p = n$ (i.e., $y_0 \to 0$) and the upper extreme to the value w_2 obtained for $y_p = \hat{y}_n + c/n$, $n_p = n$ (i.e., $y_0 \to \infty$).[20] Again for $c > 1$, $\mathcal{X}$ includes the $100(1-\gamma)\%$ Bayesian credible interval computed with the noninformative priors. For two-sided hypothesis testing, H_0 can be rejected with posterior probability $1-\gamma$ if $x_0 \notin \mathcal{X}$. Figure 6 compares the width of 95% credible interval $\mathcal{X}$, for different values of n, $\gamma = 0.05$, $\hat{y}_n = 1$, $\overline{n}_0 = 10^{-6}$, and $c \in \{0.5, 1, 2\}$, with the width of the 95% Bayesian credible interval obtained with the improper uniform prior. From the figure it can be noticed that the credible interval $\mathcal{X}$ includes the Bayesian one for any $c > 1$ and, that the difference in width between the Bayesian and the proposed credible interval is particularly large for small n.

Summing up the previous results:

- $\mathcal{M}$ is really a model of prior ignorance w.r.t. the functions (queries) that are commonly used for statistical inferences, since the lower and upper expectations of such functions are vacuous a-priori;
- because of conjugacy, the computation of the posterior set of densities

[20] For $\overline{n}_0 \approx 0$ and suitably large n the set of gamma posteriors is close to a set of Gaussian posteriors with same variance and mean between $\hat{y}_n$ and $\hat{y}_n + c/n$. Then, the left extreme of the credible interval is determined by the Gamma density corresponding to the left extreme mean and the right extreme of the credible interval by the Gamma density with the right extreme mean.

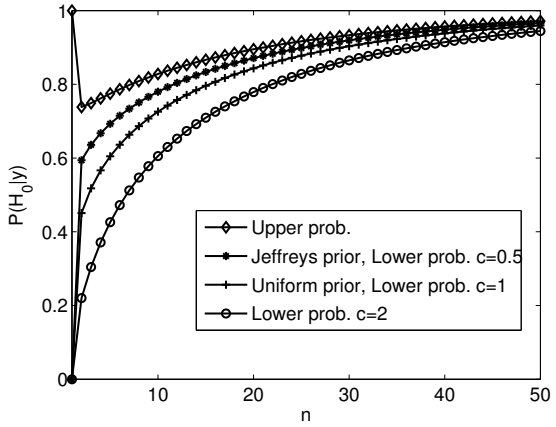

Figure 4: Lower and upper probability of H_0 as a function of n for different values of c.

and, thus, of the posterior inferences is as straightforward as for the prior ones (as in the Bayesian case);

- the inferences drawn with the model $\mathcal{M}$ encompass, for any $\overline{n}_0 > 0$ and for $c > 1$, the frequentist inferences and objective Bayesian inferences with improper priors;

- inferences drawn with the model $\mathcal{M}$ for $c > 1$ are more robust than noninformative priors, when only few observations are available.

Thus, $\mathcal{M}$ can be regarded as a model of prior ignorance that can be used as an alternative to the noninformative priors.

7. Conclusions

In this paper, we have proposed a model of prior ignorance about a scalar variable based on a set of distributions $\mathcal{M}$. In particular, we have defined some *minimal properties* that a set $\mathcal{M}$ of distributions should satisfy to be a model a prior ignorance without producing vacuous inferences. When the likelihood model belongs to the one-parameter exponential family of distributions, we have shown that, by letting the parameters of the conjugate exponential prior vary in suitable sets, it is possible to define a set of conjugate priors $\mathcal{M}$ that is the largest one which is equivalent to imposing the above properties. The set $\mathcal{M}$ satisfies the following properties:

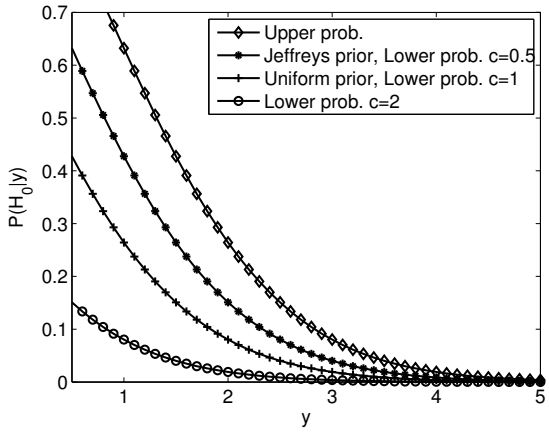

Figure 5: Lower and upper probability of H_0 as a function of y for different values of c.

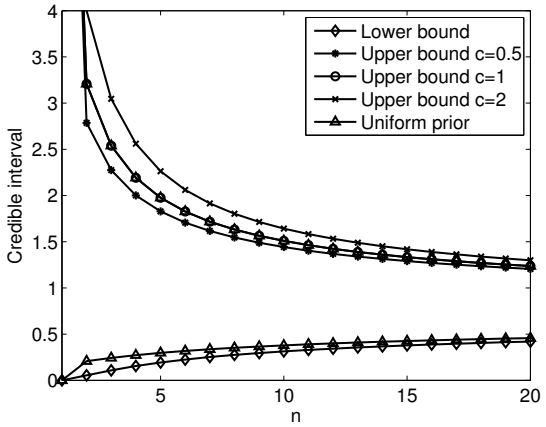

Figure 6: Credible intervals as a function of n for different values of c. The upper bound obtained from the uniform prior and the upper bound obtained from the set of priors with $c = 1$ coincide.

- it is by definition a model of prior ignorance w.r.t. the functions (queries) that are commonly used for statistical inferences (mean, one and two sided hypothesis testing, credible intervals, cumulative distributions);
- it is easy to elicit since only two (in some cases just one) parameters must be specified (as in Bayesian inferences from informative priors) and, because of conjugacy, tractable;
- it produces self-consistent (or *coherent*) probabilistic models (it only includes proper priors);
- it encompasses frequentist and objective Bayesian inferences with improper priors and, thus, it is more robust.

Future work will address the following issues: (1) application of the model to real data; (2) extension from the one-parameter exponential families to k-parameter exponential families, i.e., the multivariate case. Extending near-ignorance to the multivariate case is fundamental because the most important applications of statistical inference refer to the multivariate case such as classification, linear regression, time series models, survival analysis etc. The main problem of this extension is about how to model the prior dependence between the variables. We believe that a straightforward way to generalize the univariate model described in this paper to the multivariate case is by assuming independence. The multivariate model simply becomes a collection of independent univariate models, one for each variable in the vector. We plan to address this issue modelling the relationships between the variables that allow near-ignorance using the ℓ_∞ norm. Consider for instance the Gaussian model and assume that the mean y_0 is a vector. If the constraint $|y_0| \leq c\sigma^2$ is replaced by the constraint $||y_0||_\infty \leq c\sigma^2$, then the relative set of Gaussians $\mathcal{N}(x; y_0, \sigma^2 I)$ satisfies near-ignorance and guarantees learning and convergence. We believe that this can be proven from the results derived in this paper. This can be a starting point model for the generalization of the univariate model to the multivariate case and fast way to derive near-ignorance models useful for practical applications, such as linear regression.

Linear regression is an important area of statistical data analysis, where near-ignorance can play an important role. Consider for instance a normal linear model. In this case, the probabilistic relationship between observations and variables of interests are expressed via a multivariate normal density. The Bayesian approach to normal linear regression assumes that the prior

information on the variables of interest is expressed with a Gaussian-inverse Gamma distribution (or Gaussian-inverse Wishart distribution) which belongs to k-parameters exponential families. In the case of lack of prior information, prior ignorance is commonly modelled by selecting the parameters of the prior to make it noninformative. As pointed out in this paper, it is no possible to express prior ignorance with noninformative priors. In our opinion, a better approach is that of using a set of priors satisfying the property of near-ignorance. Notice that the Gaussian-inverse Gamma distribution belongs to k-parameters exponential families. Therefore, we can employ the multivariate models of near-ignorance for k-parameters exponential families to develop new robust regressors based on near-ignorance priors. For instance, the above discussed model based on the ℓ_∞ norm can be directly used in linear regression to model near-ignorance on the the vector of variables X in the case the variance of the observations is known. Conversely, if also the variance was unknown, we could use a Gaussian-Gamma prior model, where the Gaussian model is essentially the one derived based on the ℓ_∞ norm and for the Gamma model we could use the near-ignorance prior discussed in this paper for the one-parameter Gamma family.

Acknowledgements

This work has been partially supported by the Swiss NSF grants n. 200020-137680/1.

References

[1] J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. John Wiley & Sons, 1994.

[2] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[3] J.O. Berger, E. Moreno, L.R. Pericchi, M.J. Bayarri, Bernardo, et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

[4] L. Wasserman. Invariance properties of density ratio priors. *The Annals of Statistics*, 20(4):2177–2182, 1992.

[5] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, New York, 1985.

[6] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[7] S. Sivaganesan and J.O. Berger. Ranges of posterior measures for priors with unimodal contaminations. *The Annals of Statistics*, pages 868–889, 1989.

[8] L. DeRoberts and J.A. Hartigan. Bayesian inference using intervals of measures. *The Annals of Statistics*, 9(2):235–244, 1981.

[9] J.M. Keynes. *A treatise on probability*. Macmillan & Co., Ltd., 1921.

[10] C.A.B. Smith. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1):1–37, 1961.

[11] P.M. Williams. Coherence, strict coherence and zero probabilities. In *Fifth Int. Congress of Logic, Methodology and Philos. Sci*, pages 29–30, 1975.

[12] L.R. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, pages 1–23, 1991.

[13] B. de Finetti. *Theory of Probability*. John Wiley & Sons, Chichester, 1974–1975. Two volumes.

[14] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

[15] L.D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. 1986.

[16] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009.

[17] J.M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *Int. Journal of Approximate Reasoning*, pages 123–150, 2005.

[18] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):3–57, 1996.

[19] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.

[20] A. Wilson, A. Huzurbazar, and K. Sentz. The Imprecise Dirichlet Model for Multilevel System Reliability. *Journal of Statistical Theory and Practice*, 3(1):211–223, 2009.

[21] T.S. Ferguson. Location and scale parameters in exponential families of distributions. *The Annals of Mathematical Statistics*, 33(3):986–1001, 1962.

[22] P. Walley. A bounded derivative model for prior ignorance about a real-valued parameter. *Scandinavian Journal of Statistics*, 24(4):463–483, 1997.

[23] E. Quaeghebeur. Learning from samples using coherent lower previsions. PhD thesis, Ghent University, 2009.

[24] E. Quaeghebeur and G. De Cooman. Imprecise probability models for inference in exponential families. In *Proc. of ISIPTA'05*, pages 287–296, 2005.

[25] Agata Boratyńska. Stability of Bayesian inference in exponential families. *Statistics & Probability Letters*, 36:173–178, 1997.