Prior near ignorance for inferences in the k-parameter exponential family

A. Benavoli* and M. Zaffalon

IDSIA, Galleria 2, CH-6928 Manno (Lugano), Switzerland

This paper proposes a model of prior ignorance about a multivariate variable based on a set of distributions $\mathscr{M}$. In particular, we discuss four minimal properties that a model of prior ignorance should satisfy: invariance, near-ignorance, learning and convergence. Near-ignorance and invariance ensure that our prior model behaves as a vacuous model with respect to some statistical inferences (e.g., mean, credible intervals, etc.) and some transformation of the parameter space. Learning and convergence ensure that our prior model can learn from data and, in particular, that the influence of $\mathscr{M}$ on the posterior inferences vanishes with increasing numbers of observations. We show that these four properties can all be satisfied by a set of conjugate priors in the multivariate exponential families if the set $\mathscr{M}$ includes finitely additive probabilities obtained as limits of truncated exponential functions. The obtained set $\mathscr{M}$ is a model of prior ignorance with respect to the functions (queries) that are commonly used for statistical inferences and, because of conjugacy, it is tractable and easy to elicit. Applications of the model to some practical statistical problems show the effectiveness of the approach.

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

AMS Subject Classification: 62C10; 62F15.

1. Introduction

In performing a Bayesian analysis concerning the unknown value w of a variable $W \in \mathscr{W}$, one is required to express prior beliefs about w in the form of a prior distribution. However, we are often in a condition of prior ignorance about w and, thus, the problem is how to choose a prior distribution representing "ignorance". There are two main avenues to represent ignorance. The first assumes that ignorance can be modelled satisfactorily by a so-called noninformative prior, by which it is meant a prior that contains no information about w. The problem is how to define the meaning of "containing no prior information". It is common to call noninformative a prior that is invariant under certain re-parametrizations of the parameter space. For instance, when w is a location parameter, i.e., $\mathscr{W} = \mathbb{R}^k$, the prior should be invariant to translations. The justification is that, since the choice of an origin for a unit of measurement is often arbitrary, the noninformative prior should be independent of this choice. The uniform density on $\mathbb{R}^k$, i.e., $p(w) = 1$, is the improper prior that is invariant to translations. Other invariance properties that are often imposed to derive noninformative priors are: scale invariance when $\mathscr{W} = \mathbb{R}^+$, permutation invariance, symmetry, etc. A discussion about different ways to derive noninformative priors can be found in [1], [2, Ch. 6] and [3, Sec. 5.6.2]. Much criticism has been raised concerning the use of noninformative priors [2, Sec. 3.4.5]; for instance that they are often improper. A way to solve this problem is to use an approximation of these improper priors. For instance, in case of the uniform on $\mathbb{R}$ we can restrict the support of the prior to be a large closed interval, so that the resulting uniform truncated on this interval is proper. An issue of this approach is how to choose this interval; to do that we need some prior information that is not always available. Another possible way to justify the use of improper priors has been proposed in [4] and [5]. Noninformative priors can be regarded as limits of proper informative priors (a limit of proper truncated

priors when the truncation vanishes). The advantage of this interpretation is that it allows us to compute prior inferences (expectations) w.r.t. noninformative priors by computing this limit. This is not possible when we directly use improper priors, since the integrals diverge.

It can be shown that the probability measures that we obtain by this limit are finitely additive; this provides a justification of improper priors from a Bayesian perspective. The advantage of this interpretation of improper priors is that (i) we do not need to define a finite truncation; (ii) we can compute the prior expectation by calculating limits. A-posteriori, if the distribution that we obtain by combining the limit improper prior and the likelihood is proper, we can compute posterior expectations by moving the limit under the integral sign. In other words, we can compute first the limit which gives us the improper prior and then use this prior to compute the posterior. The presence of finitely additive measures is hidden in this case.

Other two criticisms to noninformative priors are that they are not unique, see for instance [2, Sec. 3.3.3], and that they do not really model lack of prior information, but only invariance [6]. To solve these issues, many authors [6–11] suggest that lack of prior information should be expressed in terms of a family $\mathscr{M}$ consisting of all prior distributions that are compatible with the available prior information. Inferences should then be carried out by considering the whole family $\mathscr{M}$. This approach is known as *Bayesian robustness* or *Bayesian sensitivity analysis*—examples are ε-contamination models [2, 12], restricted ε-contamination models [9], intervals of measures [10, 12], the density ratio class [6, 10], etc.

In case almost no prior information is available on W, $\mathscr{M}$ should be as large as possible in order to describe this state of prior ignorance. The natural candidate for $\mathscr{M}$ to represent complete ignorance is the set of all distributions. However, it turns out that the posterior inferences obtained from this set are *vacuous* [6, Sec. 7.3.7], i.e., the posterior set of distributions coincides with the prior set of distributions. This means that our prior beliefs do not change with experience, or, in other words, that there is no learning from data. Therefore, the vacuous prior model is not a statistically useful way to model our prior ignorance. There is then a compromise to be made. Walley suggests, as an alternative, the use of an almost vacuous model which he calls a"near-ignorance prior" [6, 11]. This is a model that behaves a priori as a vacuous model for some basic inferences (e.g., prior mean, prior credible regions) but it always provides non-vacuous posterior inferences. Near-ignorance models allow us to start a statistical analysis with very weak assumptions about the problem of interest and, thus, to implement a genuine objective-minded approach to inference. Near-ignorance models have been presented in [6, 11] to address inference problems in which: (i) the observations are discrete and the unknown W is a vector of variables with bounded components (the so-called *imprecise Dirichlet model*, or IDM); (ii) the observations are continuous and the unknown W is an unbounded but scalar variable (Gaussian or double exponential family models [11]). Near-ignorance often naturally leads to invariance. For instance, the IDM is known to satisfy permutation invariance and invariance to coarsening/refinement of the categories [6]. When this is not the case, invariance properties can be imposed to near-ignorance priors as suggested in [11].

In a recent paper [13] we have made a first proposal to generalize near-ignorance to the one-parameter exponential families of distributions. Considering the case that the likelihood model is a density belonging to the one-parameter exponential families, i.e., $w \in \mathscr{W} \subseteq \mathbb{R}$, and $\mathscr{M}$ includes the corresponding *proper conjugate exponential priors*; we have shown that there exists a choice of the parametrization of $\mathscr{M}$ which satisfies near-ignorance without leading to vacuous posterior inferences, 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 have proven that the set of priors $\mathscr{M}$ satisfying near-ignorance without leading to vacuous posterior inferences can be uniquely obtained by letting the parameters of the conjugate exponential prior vary jointly in suitable sets. This work allowed us to derive again the imprecise Beta model, which is the univariate version of the IDM, and the set of Gaussian priors discussed in [11] from a more general model of near-ignorance. Furthermore, we were able to derive new models of near-ignorance that were not available before such as, for instance, a model based on a set of Gamma priors.

In this paper, we extend the work in [13] in two main directions. First, we propose a way to reconcile the two approaches to design a prior model in case of lack of prior information, i.e., invariance and Walley's prior near-ignorance. We define four *minimal properties* that a prior ignorance model $\mathscr{M}$ should satisfy: prior-ignorance, invariance, learning from data and asymptotic convergence. To apply these properties to define a prior model, the analysts must choose: (i) w.r.t. which functions (e.g., mean, variance, credible intervals, etc.) she/he wants to be prior ignorant; (ii) w.r.t. which group transformations she/he wants to be invariant; (iii) w.r.t. which functions she/he wants to satisfy learning and convergence. These are general principles that allow the analyst to elicit a prior model in a condition of ignorance. Then we specialize these properties to the k-dimensional exponential families $\mathscr{W} = \mathbb{R}^k$, which is the main contribution of this work. For the multivariate exponential family in $\mathscr{W} = \mathbb{R}^k$, it is natural to impose: (i) prior-ignorance w.r.t. the function $b(w)$ (the mean of the observation variable); (ii) invariance to translations (w is a location parameter) and, depending on the exponential family we are considering, invariance to permutations, symmetry, invariance to coarsening/refinement, etc.; (iii) learning and convergence w.r.t. any (measurable) function.

Moreover, we show that translation invariance, near-ignorance and learning can all be satisfied by a set of conjugate priors $\mathscr{M}$ if it includes finitely additive probabilities obtained as limits of truncated increasing/decreasing exponential functions. As discussed previously, when the set of posteriors we obtain by combining the likelihood and the improper increasing/decreasing exponential functions includes only proper distributions, we can move the limit under the integral sign and compute inferences as we normally do with improper priors. When this is not the case, we must deal with this limit to obtain the lower and upper expectations. We will show that for the most important inferences in statistical analysis (mean, credible interval, etc.) computing this limit is easy. We discuss how the obtained near-ignorance models compare with other models that have been proposed for near-ignorance: the imprecise Dirichlet model [6], the bounded derivative model [14] and the nonparametric predictive inference model [15].

Finally, we show the application of our near-ignorance model to some practical statistical problems. In particular, to state the practical effectiveness of our prior ignorance model, we employ it to implement a prior near-ignorance version of the one-sample Bayesian t-test. Our new test has the peculiarity that it can suspend the judgement, which means that it stays indeterminate with respect to the choice of the null or the alternative hypothesis, in case it deems that there is not enough information in the data to take a reliable decision. By means of numerical simulations, we compare our test with the frequentist t-test, a Bayesian t-test based on an improper prior, a hierarchical Bayesian t-test based on an improper hyper-prior and show that our test is more robust. This means that if we take a random decision, with 50/50 chance, on the instances where our test is indeterminate (and take the recommended decision on the others), then we obtain a test that has the same performance (power) of the frequentist t-test and the Bayesian t-tests. This is an important result from a practical point of view, because it shows that using near-ignorance leads to isolating instances on which traditional tests have no clue. Finally, as another example, we apply our prior near-ignorance model to real data: the USA 2004 election poll.

Notation x, w, x_0, w_0, y, y_0 denote vectors in $\mathbb{R}^k$. Their components, for $i = 1, \ldots, k$, are denoted by $x_{\cdot i}, w_{\cdot i}$ for variables that do not have a subscript and by x_{0i}, w_{0i} for variables that have a subscript. g denotes a bounded real-valued function on a subset of $\mathbb{R}^k$. $E[\cdot]$ denotes the expectation operator, while $\underline{E}[\cdot]$ and $\overline{E}[\cdot]$ denote the lower and, respectively, upper expectation operators.

2. Prior Near-Ignorance and Invariance

The aim of this section is to define which minimal properties the set of priors $\mathscr{M}$ should satisfy in case there is (almost) no prior information about a random variable W taking values in $\mathscr{W} \subseteq \mathbb{R}^k$.

We start by defining prior ignorance.

Definition 2.1. A state of ignorance about a real-valued bounded function (RVBF) g of W can be modelled as follows: $\underline{E}[g] = \inf g$ and $\overline{E}[g] = \sup g$, where $\underline{E}, \overline{E}$ denote the lower and upper bounds of the expectation of g computed w.r.t. the set of priors $\mathscr{M}$.

It can be observed that the range of $E[g]$ under the set of priors $\mathscr{M}$ is the same as the original range of g. In other words, by specifying the set of priors $\mathscr{M}$, we are not giving any information on the value of the expectation of g. It corresponds to stating that we are in a state of ignorance.

In case of lack of prior information, another property that is desirable for $\mathscr{M}$ is that of satisfying some invariance principle. Consider for instance the case in which $\mathscr{W} = \mathbb{R}$. Then, in case of lack of prior information, it seems to be reasonable that, for all subsets of reals A and for all real numbers a, the prior probabilities of the set A and the shifted set $A + a$ should coincide. This property is called *translation invariance*. The principle of invariance to transformations can be formalized as follows [6, Sec. 3.5] (hereafter, we use $\mathscr{F}$ to denote a group of transformations of $\mathscr{W}$).[1]

Definition 2.2. A lower expectation $\underline{E}$ on $\mathscr{L}(\mathscr{W})$, where $\mathscr{L}(\mathscr{W})$ is the linear space of all RVBFs on $\mathscr{W}$, is called $\mathscr{F}$-invariant in $\mathscr{G} \subseteq \mathscr{L}(\mathscr{W})$ if $\underline{E}[g(f)] = \underline{E}[g]$ whenever $g \in \mathscr{G}$, $f \in \mathscr{F}$ and $g(f) \in \mathscr{G}$.

For example, let $\mathscr{W} = \mathbb{R}$ and define the translations f_a by $f_a(w) = w + a$. The group of transformations $\mathscr{F} = \{f_a : a \in \mathbb{R}\}$ is called the translation group on $\mathbb{R}$. Since $g(f_a(w)) = g(w + a)$, $\mathscr{F}$-invariance is translation invariance in this case.[2] If $g = I_{\{A\}}$,[3] with $A \subseteq \mathscr{W}$ measurable, the above equality $\underline{E}[g(f)] = \underline{E}[g]$ means that the set A and the shifted set $A + a$ should have the same lower probability for any value of a.

The only set of priors which satisfies prior ignorance, i.e., $\underline{E}[g] = \inf g$ and $\overline{E}[g] = \sup g$, for all $g \in \mathscr{L}(\mathscr{W})$ and which is $\mathscr{F}$-invariant to all possible $\mathscr{F}$ is the set of all probabilities (vacuous prior model). However, it turns out that the posterior inferences obtained from this set are *vacuous* [6, Sec. 7.3.7], i.e., the posterior set of probabilities coincides with the prior set of probabilities. This means that our prior beliefs do not change with experience (i.e., there is no learning from data). Therefore, the vacuous prior model is not a useful way to model our prior ignorance in statistics. There is then a compromise to be made: imposing prior ignorance and invariance only on a subset of the possible RVBFs and group of transformations. The set of minimal properties that $\mathscr{M}$ or, equivalently, the lower and upper expectations it generates, should satisfy to be an invariant model of prior ignorance and produce consistent and meaningful posterior inferences are formalized hereafter.

(A1) $\{\mathscr{F}_i\}$**-prior invariance.** For a chosen set of RVBFs $\mathscr{G}_1$ and a class of groups of transformations $\{\mathscr{F}_i\}$, the prior upper and lower expectations are $\{\mathscr{F}_i\}$-invariant, i.e., $\underline{E}[g(f)] = \underline{E}[g]$[4] for any $\mathscr{F}_i \in \{\mathscr{F}_i\}$, $g \in \mathscr{G}_1$, $f \in \mathscr{F}_i$ and $g(f) \in \mathscr{G}_1$.

(A2) $\mathscr{G}_0$**-prior ignorance.** The prior upper and lower expectations of some suitable set of RVBFs $\mathscr{G}_0$ under $\mathscr{M}$ are vacuous, i.e., $\underline{E}[g] = \inf g$ and $\overline{E}[g] = \sup g$ for all $g \in \mathscr{G}_0$.

(A3) $\mathscr{G}$**-learning.** For a chosen set of RVBFs $\mathscr{G} \supseteq \mathscr{G}_0$ and for each $g \in \mathscr{G}$ 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

[1] More precisely, $\mathscr{F}$ denotes a semigroup of transformations of $\mathscr{W}$. That is, each $f \in \mathscr{F}$ maps $\mathscr{W}$ into itself, and the composition $f_1 f_2$ defined by $f_1(f_2(w))$ is in $\mathscr{W}$ whenever $f_1, f_2 \in \mathscr{F}$. The semigroup $\mathscr{F}$ is Abelian if $f_1 f_2 = f_2 f_1$ whenever $f_1, f_2 \in \mathscr{F}$.

[2] In this paper we mainly focus on translation invariance. However, for multivariate models, we will impose other invariance properties: invariance to permutations and invariance to representation.

[3] Note that $I_{\{A\}}$ is the indicator function of set A, i.e., $I_{\{A\}}(x) = 1$ if $x \in A$ and zero otherwise.

[4] Equivalently, if $-g$ and $-g(f)$ belong to $\mathscr{G}_1$ then $\underline{E}[-g(f)] = \underline{E}[-g]$, which implies that $\overline{E}[g(f)] = \overline{E}[g]$ being $\overline{E}[g] = -\underline{E}[-g]$ for any g.

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.

(A4) Convergence. For each RVBF $g \in \mathscr{G}$ and sequence of observations y^n, the following conditions are satisfied for $n \to \infty$:

$$|E^*[g|y^n] - \underline{E}[g|y^n]| \to 0, \quad |E^*[g|y^n] - \overline{E}[g|y^n]| \to 0, \tag{2}$$

where $E^*[g|y^n]$ is the posterior expectation w.r.t. the posterior density derived, via Bayes' rule, from the likelihood model and some prior on $\mathscr{W}$ that is dominated by the likelihood.[1]

Property (A1) states that $\mathscr{M}$ should be prior invariant w.r.t. some class of groups of transformations $\{\mathscr{F}_i\}$ and RVBFs $\mathscr{G}_1$. In general we expect $\mathscr{G}_1$ to be very large and coincide with $\mathscr{G}$ defined in (A3). Again in case $\mathscr{M}$ includes all possible distributions then (A1) holds for any $\mathscr{F}_i$ [6, Sec. 3.5.6]. Here, conversely, we require that (A1) is satisfied for some specific choice of $\{\mathscr{F}_i\}$. This choice depends on the aims and on the application of the set of priors $\mathscr{M}$.

Property (A2) states that $\mathscr{M}$ should be vacuous a priori w.r.t. some set of RVBFs $\mathscr{G}_0$, i.e., the lower and upper expectations of $g \in \mathscr{G}_0$ respectively coincide with the infimum and the supremum of g. In case $\mathscr{M}$ includes all possible distributions then (A1) holds for any function g. Here, conversely, we require that (A1) is satisfied for some subset of RVBFs $\mathscr{G}_0$. The subset of RVBFs $\mathscr{G}_0$ used in (A1) should include the RVBFs g w.r.t. which we state our condition of prior near-ignorance.

The sets $\mathscr{G}_0$ in (A2) and $\{\mathscr{F}_i\}$ in (A1) should be also as large as possible to guarantee that also $\mathscr{M}$ is as large as possible, but no too large to be incompatible with the requirement (A3) of learning. In fact, property (A3) states that $\mathscr{M}$ should be non-vacuous a-posteriori for any RVBF $g \in \mathscr{G} \supseteq \mathscr{G}_0$, which is a condition for learning from the observations. The set of RVBFs $\mathscr{G}$ used in (A3) should consist of the RVBFs g w.r.t. which we are interested in computing expectations (i.e., making inferences). The fact that $\mathscr{G}$ must include $\mathscr{G}_0$ is the only constraint on $\mathscr{G}$, meaning that (A3) requires that $\mathscr{M}$ is not vacuous w.r.t. all these RVBFs for which the prior near-ignorance has been imposed. Since $\mathscr{M}$ is a model of prior near-ignorance, it is also desirable that the influence of $\mathscr{M}$ on the posterior inferences vanishes with increasing number of observations n. This is a sort of ultimate agreement by the accumulation of evidence.

Property (A4) states that, for $n \to \infty$, $\mathscr{M}$ should give the same lower and upper expectations of $g \in \mathscr{G}$ as those obtained from some sufficiently regular prior on $\mathscr{W}$.

In order to better understand properties (A1)–(A4), in Section 4 we show their instantiation in the case of the exponential families. Before discussing these results, in the next section we introduce the exponential families of densities and review their main properties [3, 17–19].

3. Exponential Families

Consider a sampling model where i.i.d. samples of a k-dimensional random vector Y are taken from a sample space $\mathscr{Y}$. For $w \in \mathbb{R}^k$ define $b(w) = \ln \int_{\mathscr{Y}} u(y) \exp(y^T w) dy$, (a sum in case $\mathscr{Y}$ is discrete), where $u(\cdot) : \mathscr{Y} \to \mathbb{R}^+$ and T denotes transpose. Let $\mathscr{W}$ be $\{w : \ b(w) < \infty\}$ [20]; it can be shown that $\mathscr{W}$ is a convex set and it is called the *natural parameter space*. It is further

[1] We point the reader to [16, Ch. 20] for a general discussion about dominated priors. When the likelihood belongs to the exponential families (the focus of this paper), as dominated prior we may consider any proper conjugate prior, the improper uniform or other sufficiently regular priors. The posterior becomes asymptotically Normal in these cases.

assumed that $\mathscr{W}$ is a non-empty open set in $\mathbb{R}^k$, i.e., we restrict attention to regular exponential families [20]. The probability density

$$p(y|w) = u(y)\exp(y^T w - b(w)),\ y \in \mathscr{Y}, \tag{3}$$

is called the canonical form of representation of the exponential family; $y \in \mathscr{Y}$ is a so-called sufficient statistic; $w \in \mathscr{W}$ is a so-called natural parameter. The functions $u(\cdot) : \mathscr{Y} \to \mathbb{R}^+$ and $b(\cdot) : \mathscr{W} \to \mathbb{R}$ characterize each particular family of the exponential families of distributions. The canonical form has some useful properties. First, the mean and variance of Y are given by

$$E[Y|w] = \nabla b(w),\ \ Var[Y|w] = \nabla^2 b(w), \tag{4}$$

where $\nabla b = [\frac{\partial b}{\partial w_{\cdot 1}}, \frac{\partial b}{\partial w_{\cdot 2}}, \ldots, \frac{\partial b}{\partial w_{\cdot k}}]^T$ is the gradient (a column vector), $w_{\cdot i}$ denotes the i-th component of the vector and $\nabla^2 b(w) > 0$ for each $w \in \mathscr{W}$ is the Hessian matrix. Second, in the case of n i.i.d. observations y_i, it follows that

$$\prod_{i=1}^{n} p(y_i|w) = \exp\left(n\left(\hat{y}_n^T w - b(w)\right)\right)\prod_{i=1}^{n} u(y_i),\quad w \in \mathscr{W}, \tag{5}$$

where $\hat{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean of y_i which, together with n, is a sufficient statistic of $y_1, \ldots, y_n$ for inference about W under the i.i.d. assumption.

Furthermore, by interpreting the density function in (5) as a likelihood function, we can define the corresponding conjugate prior. A probability density $p(w|n_0, y_0)$, parametrized by $n_0 \in \mathbb{R}$ and $y_0 \in \mathbb{R}^k$, is said to be the canonical conjugate prior of (3) if

$$p(w|n_0, y_0) \propto \exp\left(n_0\left(y_0^T w - b(w)\right)\right),\ \ w \in \mathscr{W}, \tag{6}$$

where n_0 is the so-called number of pseudo-observations and y_0 is the so-called pseudo-observation. Consider the convex hull of the support set of the measure $u(dy)$, and then let $\mathscr{Y}_0$ be the interior of this convex set. From (4), it follows that $\mathscr{Y}_0$ is the set of possible values of ∇b [20].

Lemma 3.1. *Consider the prior $p(w|n_0, y_0) \propto \exp\left(n_0\left(y_0^T w - b(w)\right)\right)$. If $y_0 \in \mathscr{Y}_0$ and $0 < n_0 < \infty$, then the density is proper, i.e., there exists a normalization constant $k(n_0, y_0) > 0$ such that*

$$\int k(n_0, y_0)\exp\left(n_0\left(y_0^T w - b(w)\right)\right) dw = 1.$$

The lemma is proven in [20, Th. 1]. Lemma 3.1 states that if $n_0 \in \mathbb{R}^+$ and $y_0 \in \mathscr{Y}_0$ then the kernel $\exp\left(n_0\left(y_0^T w - b(w)\right)\right)$ is integrable and, thus, proper.

The likelihood and conjugate prior pair in the canonical exponential families satisfy 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 lemmas.

Lemma 3.2. *For a pair of likelihood and conjugate prior in the canonical exponential family, it holds that the posterior density for w is:*

$$p\left(w|n_p, y_p\right) = k(n_p, y_p)\exp\left(n_p\left({y_p}^T w - b(w)\right)\right),\ \ w \in \mathscr{W}, \tag{7}$$

where $n_p = n + n_0$ and $y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$, which is always proper if $n_0 > 0$ and $y_0 \in \mathscr{Y}_0$.

This follows by conjugacy, see [3, Ch. 5]. In (4), 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, it is interesting to compute the prior and posterior mean of ∇b. For sampling models such that $\nabla b(w) = x$, these expectations give also a closed formula for the prior and posterior expectations of x.

Lemma 3.3. *The prior mean of the function ∇b is $E[\nabla b|n_0, y_0] = y_0$ if $y_0 \in \mathscr{Y}_0$ and $n_0 > 0$ and the posterior mean: $E[\nabla b|n_p, y_p] = y_p = \frac{n_0 y_0 + n\hat{y}_n}{n+n_0}$.*

These results follow from the properties of exponential families [3, Ch. 5], [20]. Examples of conjugate exponential families can be found for instance in [3, Ch. 5].

4. Near-Ignorance For k-Parameter Exponential Families

Consider the problem of statistical inference about the real-valued parameter $w \in \mathscr{W} = \mathbb{R}^k$ 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 probability mass function (PMF) in the discrete case):

$$\prod_{i=1}^{n} p(y_i|w) = \exp\left(n\left(\hat{y}_n^T w - b(w)\right)\right) \prod_{i=1}^{n} u(y_i), \tag{8}$$

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 (the likelihood is precisely known, a single PDF/PMF specified by the parameters $\hat{y}_n$ and n). By conjugacy and following a Bayesian approach, as prior for w we may consider the proper PDF $p(w|n_0, y_0)$ defined in (6) for a given value of the parameter $y_0 \in \mathscr{Y}_0$ and $n_0 \in \mathbb{R}^+$. In the case there is not enough information about the value w of W to uniquely determine the values of the parameters y_0 and n_0, we can consider the family $\mathscr{M}$ of prior densities which represent prior-ignorance and satisfy some invariance principles, but without leading to vacuous posterior inferences. In other words, the question is how to construct a family $\mathscr{M}$ of prior densities which satisfies properties (A1)–(A4) discussed in Section 2. For tractability of inferences, we restrict this set of priors only to include densities such that the resulting set of posteriors is conjugate with the likelihood model. Before answering this question for the general multivariate case, $\mathscr{W} = \mathbb{R}^k$, we focus on the univariate case, $\mathscr{W} = \mathbb{R}$. We start proving the following lemma.

Lemma 4.1. *The lower expectation model*[1]

$$\underline{E}[g] = \liminf_{r\to\infty} \int g(w)p(w)dw, \tag{9}$$

with $p(w) = \frac{\ell}{\exp(\ell r)} \exp(\ell w) I_{\mathscr{W}_r}(w)$ for $\ell \neq 0$,

$$\mathscr{W}_r = \begin{cases} (-\infty, r] & \text{if } \ell > 0, \\ [-r, \infty) & \text{if } \ell < 0, \end{cases} \tag{10}$$

$r \in \mathbb{R}$ and $p(w) = \frac{1}{2r} I_{\mathscr{W}_r}(w)$ with $\mathscr{W} = [-r, r]$ for the case $\ell = 0$, satisfies $\underline{E}[g(f_a)] = \underline{E}[g]$ with $f_a(w) = w + a$ (translation invariance) for any g in the set of Borel-measurable RVBFs $\mathscr{K}$ and $a \in \mathbb{R}$.

[1] Let the least term ϕ of a sequence be a term which is smaller than all but a finite number of the terms which are equal to ϕ. Then ϕ is called the lower limit of the sequence.

Proof. For $\ell = 0$, $p(w) = \frac{1}{2r} I_{\mathscr{W}_r}(w)$ coincides with the uniform distribution on $\mathscr{W} = [-r, r]$, then

$$\begin{aligned}\underline{E}[g(f_a)] &= \liminf_{r\to\infty} \frac{1}{2r}\int_{-r}^{r} g(w+a)dw = \liminf_{r\to\infty} \frac{1}{2r}\int_{-r+a}^{r+a} g(w)dw \\ &= \liminf_{r\to\infty} \frac{1}{2r}\left(\int_{-r}^{r} g(w)dw + \int_{r}^{r+a} g(w)dw - \int_{-r}^{-r+a} g(w)dw\right) \\ &= \underline{E}[g] + \liminf_{r\to\infty} \frac{1}{2r}\left(\int_{r}^{r+a} g(w)dw - \int_{-r}^{-r+a} g(w)dw\right).\end{aligned}$$

Hence, being g bounded (it is a RVBF), it follows that both $\int_r^{r+a} g(w)dw$ and $\int_{-r}^{-r+a} g(w)dw$ are bounded for any r and, thus, the limit in the last equation above goes to zero for $r \to \infty$. Therefore, it results that $\underline{E}[g(f_a)] = \underline{E}[g]$. We have thus proven that the limit for $r \to \infty$ of the uniform distribution on $\mathscr{W} = [-r, r]$ satisfies translation invariance.

For $\ell \neq 0$ and $\mathscr{W}_r = (-\infty, r]$ (the case $\mathscr{W}_r = [-r, \infty)$ is similar),

$$\underline{E}[g(f_a)] = \inf_p \int g(w+a)p(w)dw = \inf_p \int g(w)p(w-a)dw,$$

where $\inf_p$ means $\liminf_{r\to\infty}$ with respect to $p(w) = \frac{\ell}{\exp(\ell r)} \exp(\ell w) I_{(-\infty,r]}(w)$ or, equivalently, to $p(w-a) = \frac{\ell}{\exp(\ell r)} \exp(\ell(w-a)) I_{(-\infty,r]}(w-a)$. Define $r_1 = r + a$, then $p(w-a) = \frac{\ell}{\exp(\ell r_1)} \exp(\ell w) I_{(-\infty,r_1]}(w)$ and, thus,

$$\underline{E}[g(f_a)] = \liminf_{r_1\to\infty} \frac{\ell}{\exp(\ell r_1)} \int g(w) \exp(\ell w) I_{(-\infty,r_1]}(w)dw,$$

which is equal to $\underline{E}[g]$ and, thus, satisfies translation invariance. □

Consider the kernel $\exp(\ell w)$; transformed back to the original parameter space, it becomes $\exp(\ell x)$ in the Normal case, $\theta^{\ell-1}(1-\theta)^{-\ell-1}$ in the Beta case and $\lambda^{\ell-1}$ in the Gamma case.[1] We can see these priors are generalizations of the translation invariant improper priors that are normally used in Bayesian analysis. These priors can be obtained by setting $\ell = 0$: they are respectively 1 (improper uniform), $\theta^{-1}(1-\theta)^{-1}$ (Haldane prior) and λ^{-1} (Jeffreys prior). All these priors are improper. Improper priors can be defined as a limit of truncations of proper priors as in (9). It can be shown that the probability measures that we obtain by this limit are finitely additive; this provides a justification of improper priors from a Bayesian perspective. We discuss this connection with more details in Appendix A. A-posteriori, if the distribution that we obtain by combining the limit improper prior and the likelihood is proper, we can compute posterior expectations by moving the limit under the integral sign. Hereafter, for notational convenience, we will often interchange the limiting procedure in $\underline{E}[g]$ with its limit kernel $\exp(\ell w)$, since they are in most of the cases equivalent in terms of posterior inferences. However, the reader should be aware that, for instance, in referring to the prior expected value of g w.r.t. $\exp(\ell w)$ we actually mean (9). Furthermore, again for notational convenience, we will denote the prior with $p(w) = \frac{\ell}{\exp(\ell r)} \exp(\ell w) I_{\mathscr{W}_r}(w)$ even in the case $\ell = 0$ where it should be $p(w) = \frac{1}{2r} I_{[-r,r]}(w)$ (this because the lower and upper expectations of RVBFs hereafter considered are obtained by priors of this form $p(w) = \frac{\ell}{\exp(\ell r)} \exp(\ell w) I_{\mathscr{W}_r}(w)$).

Theorem 4.2. *Assume that $\mathscr{W} = \mathbb{R}$ and consider properties (A1)–(A4), with $\mathscr{G} = \mathscr{G}_1$ including sufficiently smooth RVBFs,[2] $\mathscr{G}_0 = \{b'\}$, where b' is the first derivative of b, and $\mathscr{F}_i = \mathscr{F}$, where*

[1] The differences are due to the Jacobians of the transformations.

[2] With sufficiently smooth RVBFs, we mean integrable w.r.t. the kernel $\exp\left(n\left(\hat{y}_n^T w - b(w)\right)\right)\exp(\ell w)$ for any $\ell \in [-c, c]$, $n \in \mathbb{N}$ and $\hat{y}_n \in Cl(\mathscr{Y}_0)$, with support in $\mathscr{W}$ and continuous on a neighbourhood of the point where the posterior relative to the improper uniform prior concentrates for $n \to \infty$.

$\mathscr{F} = \{f_a : a \in \mathbb{R}\}$ and $f_a(w) = w + a$, i.e., a translation of the parameter space. Given the parameter $c \in \mathbb{R}^+$, the following set of priors:

$$\mathscr{M} \propto \{\exp(\ell w),\ \ell \in [-c,c]\}, \tag{11}$$

satisfies (A1)–(A4) and conjugacy between likelihood and posterior.

Proof. Consider first property (A1). Define $\underline{E}[g] = \inf\{\underline{E}_\ell[g] : \ell \in [-c,c]\}$, where $\underline{E}_\ell[g] = \liminf_{r\to\infty} \frac{\ell}{\exp(\ell r)} \int g(w)\exp(\ell w) I_{\mathscr{W}_r}(w) dw$. From Lemma 4.1, it follows that $\underline{E}_\ell[g(f_a)] = \underline{E}_\ell[g]$ for any ℓ and, thus, translation invariance holds for any kernel $\exp(\ell w)$. Thus, it holds also for the kernel ℓ which attains the infimum in $\underline{E}[g]$. Therefore, $\underline{E}[g(f_a)] = \underline{E}[g]$ for any $g \in \mathscr{G}_1$ and, thus, the set of priors $\mathscr{M}$ is invariant under translations. Furthermore, it can be seen that the elements of $\mathscr{M}$ satisfy conjugacy between posteriors and the likelihood model, in fact: $\exp(n(\hat{y}_n w - b(w)))\exp(\ell w) = \exp(n(\frac{n\hat{y}_n+\ell}{n} w - b(w)))$.

Consider now (A2). From Section 3 it follows that $\nabla b = b' \in \mathscr{Y}_0$ (we are in the scalar case, $\mathscr{W} = \mathbb{R}$) and the first derivative of b is always increasing in $\mathscr{W}$ since the second derivative b'' is positive. Therefore, since for $\ell > 0$, $\exp(\ell w)$ is always increasing in $\mathscr{W}$, it follows that:

$$\overline{E}\left[b'\right] = \sup_{\ell\in[-c,c]} \limsup_{r\to\infty} \frac{\ell}{\exp(\ell r)} \int b'(w)\ \exp(\ell w) I_{\mathscr{W}_r}(w) = \sup \mathscr{Y}_0,$$

where $\mathscr{W}_r = (-\infty, r]$. The last equality can be easily derived by the fact that $\exp(\ell w)$ has a maximum in $w = r$ and thus the mass of $\exp(\ell w)/\exp(\ell r)$ is concentrated on a small neighborhood of $w = r$. In fact, given r for any $r_1 < r$ it holds that

$$\begin{aligned}
&\frac{\ell}{\exp(\ell r)} \int_{-\infty}^{r} b'(w)\ \exp(\ell w) \\
&= \frac{\ell}{\exp(\ell(r-r_1))\exp(\ell r_1)} \int_{-\infty}^{r_1} b'(w)\ \exp(\ell w) + \frac{\ell}{\exp(\ell r)} \int_{r_1}^{r} b'(w)\ \exp(\ell w) \\
&> \frac{\ell}{\exp(\ell(r-r_1))\exp(\ell r_1)} \int_{-\infty}^{r_1} b'(w)\ \exp(\ell w) + \frac{\exp(\ell r) - \exp(\ell r_1)}{\exp(\ell r)} b'(r_1) \overset{r\to\infty}{=} b'(r_1).
\end{aligned}$$

Similarly, it can be shown that $\lim_{r\to\infty} \frac{\ell}{\exp(\ell r)} \int_{-\infty}^{r} b'(w) < \lim_{r\to\infty} b'(r)$. Hence, for any finite r_1,

$$b'(r_1) < \lim_{r\to\infty} \frac{\ell}{\exp(\ell r)} \int_{-\infty}^{r} b'(w)\ \exp(\ell w) < \lim_{r\to\infty} b'(r),$$

which by definition of supremum implies that $\overline{E}\left[b'\right] = \sup b'(r) = \sup \mathscr{Y}_0$.[1] For $r \to \infty$, the mass goes to infinity and, thus, concentrates on the value of w which gives the maximum of b'. In a similar way, since for $\ell < 0$, $\exp(\ell w)$ is always decreasing in $\mathscr{W}$, one has:

$$\underline{E}\left[b'\right] = \inf_{\ell\in[-c,c]} \liminf_{r\to\infty} \frac{\ell}{\exp(\ell r)} \int b'(w)\ \exp(\ell w) I_{[-r,\infty)}(w) = \inf \mathscr{Y}_0.$$

[1] This holds for any $\ell \in (0,c]$. All $\ell \in (0,c]$ are equivalent w.r.t. this property, since all $\exp(\ell w)$ are increasing in $\mathscr{W}$ for $\ell > 0$.

Thus, (A2) is also satisfied. Consider property (A3). The set of posteriors resulting from (11) is:

$$\mathscr{M}_p \propto \left\{ \exp\left(n\left(\frac{\ell + n\hat{y}_n}{n} w - b(w) \right) \right),\ \ell \in [-c,c] \right\}. \tag{12}$$

Since $\mathscr{Y}_0$ is an open convex subset of $\mathbb{R}$ [20] (and thus dense in $\mathbb{R}$) and since $\hat{y}_n \in Cl(\mathscr{Y}_0)$, there exists a $\delta > 0$ such that for $n > \delta$ either $\frac{\ell+n\hat{y}_n}{n} \in \mathscr{Y}_0$ or $\frac{-\ell+n\hat{y}_n}{n} \in \mathscr{Y}_0$ for $\ell \in (0,c]$. Thus, for any $\hat{y}_n$ and for a sufficiently large n, there are always probabilities in $\mathscr{M}_p$ that are countably additive (they have proper PDFs) for either $\ell \in (0,c]$ or $\ell \in [-c,0)$. By applying Proposition 3.3 with $n_0 y_0 = \ell$ and $n_0 \to 0$ to the proper PDFs in $\mathscr{M}_p$, this implies that $\underline{E}[b'|n,\hat{y}_n] = \frac{-c+n\hat{y}_n}{n}$ and/or $\overline{E}[b'|n,\hat{y}_n] = \frac{c+n\hat{y}_n}{n}$ for $n \geq \delta$ (in case $\hat{y}_n \in \mathscr{Y}_0$ but $\hat{y}_n \notin Cl(\mathscr{Y}_0)\backslash\mathscr{Y}_0$ all probabilities in $\mathscr{M}_p$ are countably additive for all $\ell \in [-c,c]$ for a suitably large n). This means that (A3) holds w.r.t. the lower and/or the upper for the RVBF $g = b'$ for which we have stated our condition of prior ignorance (A2). Furthermore, since for $n \to \infty$, $\frac{\ell+n\hat{y}_n}{n} \to \hat{y}_n$ and $\mathscr{M}_p$ reduces to the single posterior obtained from the improper prior $p(w) = 1$, it follows that the lower and upper posterior expectations of any RVBF $g \in \mathscr{G}$ converge to the posterior expectation obtained from the improper uniform prior $p(w) = 1$. This proves (A4) but also (A3); in fact since, for $n \to \infty$, $\underline{E}[g|n,\hat{y}_n] = \overline{E}[g|n,\hat{y}_n]$, this implies that either $\underline{E}[g|n,\hat{y}_n] \neq \underline{E}[g]$ or $\overline{E}[g|n,\hat{y}_n] \neq \overline{E}[g]$ for all g such that $\overline{E}[g] - \underline{E}[g] > 0$. □

Summarizing, Theorem 4.2 states that:

- Translation invariance and prior ignorance are satisfied by the set of finitely additive probabilities obtained from the kernels in (11): that is, as lower limit for $r \to \infty$ of the truncated densities $\exp(\ell w) I_{\mathscr{W}_r}(w)$, see Appendix A.
- A-posteriori, for any $n > 0$, there are always probabilities in $\mathscr{M}_p$ that are countably additive and, at the increase of n these probabilities give either the lower or upper expectation of g or both (since either $\frac{c+n\hat{y}_n}{n} \in \mathscr{Y}_0$ or $\frac{c+n\hat{y}_n}{n} \in \mathscr{Y}_0$, while if they are both in $\mathscr{Y}_0$ then all the probabilities in $\mathscr{M}_p$ are countably additive). This explains why also (A3)–(A4) are satisfied.
- The result in Theorem 4.2 is different from the one we have obtained in [13], because of the additional requisite of translation invariance that is not satisfied by the set of conjugate priors in [13]. The advantage of imposing translation invariance is twofold. First, it implies that the lower and upper posterior inferences are invariant to translations of the sample mean—this aspect will be clarified in Section 5. Second, it allows us to reduce to one (the constant c) the number of parameters that specify the set of priors, while in [13] there are two parameters. We compare the two models with more details in Section 7.1.

A consequence of Theorem 4.2 is given in the following corollary.

Corollary 4.3. *The lower and upper posterior expectation of b' obtained from the set of posteriors (12) is:*

$$\underline{E}[b'|n,\hat{y}_n] = \max\left(\inf\mathscr{Y}_0, \tfrac{n\hat{y}_n - c}{n}\right),\ \ \overline{E}[b'|n,\hat{y}_n] = \min\left(\sup\mathscr{Y}_0, \tfrac{n\hat{y}_n + c}{n}\right), \tag{13}$$

and there exists $\delta > 0$ such that, for any $n > \delta$ either the lower or the upper is different from $\inf\mathscr{Y}_0$ *or, respectively,* $\sup\mathscr{Y}_0$.[1]

Proof. In case $\frac{n\hat{y}_n+\ell}{n} \in \mathscr{Y}_0$ for any $\ell \in [-c,c]$ and any n, any posterior in (12) is proper and, thus, the result follows by Proposition 3.3 with $n_0 y_0 = \ell$ and $n_0 \to 0$. Conversely, when this is not the

[1] Notice that this behavior in general is not monotone and depends on how $\hat{y}_n$ converges with the number of observations.

case, observe that:

$$\frac{d\exp\left(n\left(y_p w - b(w)\right)\right)}{dw} = n\left(y_p - b'(w)\right)\exp\left(n\left(y_p w - b(w)\right)\right),$$

where $y_p = \frac{n\hat{y}_n+\ell}{n}$. Since $b''(w) > 0$ for each $w \in \mathscr{W}$, the function b' is always increasing in $\mathscr{W}$ and obtains its infimum $\inf \mathscr{Y}_0$ and supremum $\sup \mathscr{Y}_0$ for $w \to \inf \mathscr{W}$ and, respectively, $w \to \sup \mathscr{W}$. Since $y_p \in Cl(\mathscr{Y}_0)$ and $b' \in \mathscr{Y}_0$, then if $y_p \leq \inf \mathscr{Y}_0$, it follows that $y_p - b'(w) \leq 0$ for any $w \in \mathscr{W}$ and, thus, the kernel $\exp\left(n\left(y_p w - b(w)\right)\right)$ is always decreasing in w. Similarly it holds that if $y_p \geq \sup \mathscr{Y}_0$ then $\exp\left(n\left(y_p w - b(w)\right)\right)$ is always increasing in w. Therefore, at the limit $r \to \infty$, the posterior mean computed w.r.t. the prior $\exp(\ell w) I_{(-\infty,r]}(w)$ if $\ell > 0$ or, $\exp(\ell w) I_{[-r,\infty)}(w)$ if $\ell < 0$, will be equal to $\sup \mathscr{Y}_0$ in the case $y_p > \sup \mathscr{Y}_0$ or, respectively, $\inf \mathscr{Y}_0$ in the case $y_p < \inf \mathscr{Y}_0$. For the second part of the corollary, as in the proof of Theorem 4.2, observe that, in case $\hat{y}_n \in Cl(\mathscr{Y}_0)\backslash\mathscr{Y}_0$, for sufficiently large n either $\frac{\ell+n\hat{y}_n}{n} \in \mathscr{Y}_0$ or $\frac{-\ell+n\hat{y}_n}{n} \in \mathscr{Y}_0$ for $\ell \in (0,c]$. This means that, for a sufficiently large n, if $\underline{E}[b'|n,\hat{y}_n] = \inf \mathscr{Y}_0$ then $\overline{E}[b'|n,\hat{y}_n] = \frac{n\hat{y}_n+c}{n}$ while if $\overline{E}[b'|n,\hat{y}_n] = \sup \mathscr{Y}_0$ then $\underline{E}[b'|n,\hat{y}_n] = \frac{n\hat{y}_n-c}{n}$. □

We illustrate the results in Theorem 4.2 and Corollary 4.3 with the following example.

Example 4.4. In the Poisson case, the likelihood is $\prod_{i=1}^n p(y_i|\theta) = \prod_{i=1}^n P(y_i;\lambda) = \lambda^{\sum_{i=1}^n y_i}\exp(-n\lambda)$. Assume we do not have prior information about the unknown parameter λ. To model this condition of prior ignorance, we can employ the set of priors (11), that transformed to the original parameter space reduces to $\mathscr{M} \propto \{\lambda^{\ell-1},\ \ell \in [-c,c]\}$, where λ^{-1} is the determinant of the Jacobian. If $c > 1$, it includes Jeffreys' prior $\lambda^{-\frac{1}{2}}$ and the improper prior λ^{-1} and it satisfies the prior-ignorance property $\underline{E}[\lambda] = 0$ and $\overline{E}[\lambda] = \infty$. The resulting set of posteriors obtained by combining priors and likelihood is:

$$\mathscr{M}_p \propto \left\{\lambda^{\sum_{i=1}^n y_i+\ell-1}\exp(-n\lambda),\ \ell \in [-c,c]\right\}. \tag{14}$$

It can be noticed that if $\sum_{i=1}^n y_i + \ell > 0$ for any ℓ, $\mathscr{M}_p$ includes proper Gamma densities (it can thus be normalized), otherwise $\mathscr{M}_p$ includes densities that again concentrate on $\lambda = 0$ for $\ell < 0$. Thus, from Corollary 4.3 it follows that a-posteriori:

$$\underline{E}[\lambda|n,\hat{y}_n] = \max\left(0, \frac{-c+\sum_{i=1}^n y_i}{n}\right),\ \ \overline{E}[\lambda|n,\hat{y}_n] = \frac{c+\sum_{i=1}^n y_i}{n}.$$

□

We are now ready to extend the results of Theorem 4.2 to the multivariate case. Before doing that, we state the following lemma that will be exploited in the theorem.

Lemma 4.5. *Let $\mathbb{L}$ be a bounded closed convex subset of $\mathbb{R}^k$ strictly including the origin (i.e., the origin is in the interior of $\mathbb{L}$), and let $\ell = [\ell_{\cdot 1},\ldots,\ell_{\cdot k}]^T$ belong to $\mathbb{L}$; then it holds that*

$$\min_{\ell\in\mathbb{L}} \ell_{\cdot i} < 0, \qquad \max_{\ell\in\mathbb{L}} \ell_{\cdot i} > 0. \tag{15}$$

Proof. This results follows straightforwardly by considering any ball $\mathscr{B}(\mathbb{L}) \subset \mathbb{L}$ centred at the origin and observing that (15) holds for $\ell \in \mathscr{B}(\mathbb{L})$. □

Theorem 4.6. *Assume that $\mathscr{W} = \mathbb{R}^k$ and consider properties (A1)–(A4), with $\mathscr{G}_0 = \{\frac{\partial b}{\partial w_{\cdot 1}},\ldots,\frac{\partial b}{\partial w_{\cdot k}}\}$, $\mathscr{G} = \mathscr{G}_1$ including sufficiently smooth RVBFs and $\mathscr{F}_i = \mathscr{F}$, where $\mathscr{F} = \{f_a : a \in$*

$\mathbb{R}^k\}$ *and* $f_a(w) = w + a$. *The following set of priors:*

$$\mathscr{M} \propto \left\{\exp(\ell^T w),\ \ell = [\ell_{\cdot 1}, \ldots, \ell_{\cdot k}]^T \in \mathbb{L}\right\}, \tag{16}$$

where $\mathbb{L}$ *is a bounded closed convex subset of* $\mathbb{R}^k$ *strictly including the origin, satisfies (A1)–(A4) and conjugacy.*

Proof. Consider first property (A1). Define $\underline{E}[g] = \inf\{\underline{E}_\ell[g] : \ell \in \mathbb{L}\}$, where

$$\underline{E}_\ell[g] = \liminf_{r_{\cdot 1}\to\infty} \ldots \liminf_{r_{\cdot k}\to\infty} \int \cdots \int g(w) \exp(\ell^T w) \prod_{i=1}^{k} I_{\mathscr{W}_{r_{\cdot i}}}(w_{\cdot i}) \frac{\ell_{\cdot i}}{\exp(\ell_{\cdot i} r_{\cdot i})} dw_{\cdot i},$$

and $\mathscr{W}_{r_{\cdot i}}$ defined in (10). Observe that Proposition 4.1 also holds for $\exp(\ell^T w)$ in the multivariate case (it is enough to proceed component-wise). In fact, since $g(f_a) = g(w + a)$ then after the change of variables $w' = w - a$,

$$\underline{E}_\ell[g(f_a)] = \liminf_{r_{\cdot 1}\to\infty} \ldots \liminf_{r_{\cdot k}\to\infty} \int \cdots \int g(w') \exp(\ell^T w') \prod_{i=1}^{k} I_{\{\mathscr{W}_{r_{\cdot i}} + a_{\cdot i}\}}(w'_{\cdot i}) \frac{\ell_{\cdot i}}{\exp(\ell_{\cdot i}(r_{\cdot i} + a_{\cdot i}))} dw'_{\cdot i}.$$

Thus, by redefining $r_{\cdot i} = r_{\cdot i} + a_{\cdot i}$ and $\mathscr{W}_{r_{\cdot i}} + a_{\cdot i} = \mathscr{W}_{r_{\cdot i}}$, it follows that $\underline{E}_\ell[g(f_a)] = \underline{E}_\ell[g]$ for any ℓ and, thus, translation invariance holds for any kernel $\exp(\ell^T w)$. Then it holds also for the kernel ℓ which attains the infimum in $\underline{E}[g]$. Therefore, $\underline{E}[g(f_a)] = \underline{E}[g]$ for any $g \in \mathscr{G}_1$ and, thus, the set of priors $\mathscr{M}$ is invariant under translations.

Property (A2) follows from Theorem 4.2 by proceeding component-by-component for each $\frac{\partial b}{\partial w_{\cdot i}}$ for $i = 1, \ldots, k$. In fact,

$$\begin{aligned}
\underline{E}_\ell\left[\frac{\partial b}{\partial w_{\cdot i}}\right] &= \liminf_{r_{\cdot 1}\to\infty} \ldots \liminf_{r_{\cdot k}\to\infty} \int \cdots \int \frac{\partial b}{\partial w_{\cdot i}} \exp(\ell^T w) \prod_{i=1}^{k} I_{\mathscr{W}_{r_{\cdot i}}}(w_{\cdot i}) \frac{\ell_{\cdot i}}{\exp(\ell_{\cdot i} r_{\cdot i})} dw_{\cdot i} \\
&= \liminf_{r_{\cdot i}\to\infty} \int \frac{\partial b}{\partial w_{\cdot i}} \exp(\ell_{\cdot i} w_{\cdot i}) I_{\mathscr{W}_{r_{\cdot i}}}(w_{\cdot i}) \frac{\ell_{\cdot i}}{\exp(\ell_{\cdot i} r_{\cdot i})} dw_{\cdot i}.
\end{aligned}$$

Since $\mathbb{L}$ includes the origin, then from Proposition 4.5 it follows that $\ell_{\cdot i}$ can assume both positive and negative values. Thus, we are back to Theorem 4.2 (univariate case) in which $\ell_{\cdot i}$ is free to vary in a closed interval including the origin. Thus, the infimum and supremum of $\frac{\partial b}{\partial w_{\cdot i}}$ are obtained for $w_{\cdot i} \to -\infty$ and, respectively, $w_{\cdot i} \to +\infty$. By combining likelihood and the priors in (16), one obtains the posterior kernels:

$$\mathscr{M}_p = \left\{ \exp\left(n \left(\left(\frac{\ell + n\hat{y}_n}{n} \right)^T w - b(w) \right) \right),\ \ell \in \mathbb{L} \right\}. \tag{17}$$

Observe that $\frac{\ell + n\hat{y}_n}{n} = \hat{y}_n + \frac{\ell}{n}$. Since $\mathbb{L}$ is bounded, the effect of $\frac{\ell}{n}$ vanishes for $n \to \infty$, which proves (A4) and thus (A3). □

Theorem 4.6 requires $\mathbb{L}$ to be bounded and to include the origin in order to satisfy (A1)–(A4). For symmetry reasons, we restrict $\mathbb{L}$ to be symmetric w.r.t. the origin so to reduce to the set $[-c, c]$ in the $k = 1$ case. It is clear that while in the one-dimensional case, convexity, boundness and symmetry are satisfied only by the interval $[-c, c]$ in the $(k > 1)$-dimensional case there are many subsets of $\mathbb{R}^k$ that satisfy these three characteristics. In other words, Theorem 4.6 does not specify uniquely $\mathbb{L}$ and, thus, the modeller can add further characteristics to $\mathbb{L}$ in order to satisfy additional properties, for instance to determine the shape of the set of posterior means.

Corollary 4.7. *The expectation of* ∇b *calculated from the set of posteriors (17) satisfies:*

$$E\left[\nabla b|n,\hat{y}_n\right] \in \left\{\tfrac{\ell+n\hat{y}_n}{n},\ \ \ell \in \mathbb{L}\right\}, \tag{18}$$

provided that all the posteriors in (17) are proper.

Proof. The result follows directly from Lemma 3.3 assuming that the set $\mathscr{M}_p$ contains all proper distributions. This always holds for a large enough n because (17) satisfies (A4). □

Corollary 4.7 is weaker than Corollary 4.3, because it provides the shape of the posterior set of means only in the case all the posteriors in (17) are proper. In the multivariate case it is difficult to characterize in general the shape of the set of posterior means when (17) can include some improper distribution. This difficulty is due to the interaction among the components of w, which is determinate by the function $b(w)$, and changes depending on the exponential family we are considering. Although we cannot provide a general characterization, we can extend Corollary 4.7 case-by-case, as it is shown in the next examples about the multivariate Normal and categorical distributions.

4.1 *Multivariate Normal with Known Variance*

In the multivariate Normal case with known variance, $\prod_{i=1}^{n} p(y_i|x) = \prod_{i=1}^{n} N(y_i;x,V)$ with $x, y_i \in \mathbb{R}^d$, $0 < V \in \mathbb{R}^{d\times d}$ and $k = d$. Let us assume that $\mathbb{L} = \{\ell \in \mathbb{R}^d : \ell_{\cdot i} \in [-c_{\cdot i}, c_{\cdot i}],\ c_{\cdot i} > 0,\ i = 1,\ldots,k\}$, this models a sort of "uncorrelated" prior uncertainty (i.e., $\ell_{\cdot i}$ is free to vary in $[-c_{\cdot i}, c_{\cdot i}]$ independently of the values that $\ell_{\cdot j}$ with $j \neq i$ assumes). Notice in fact that $\mathbb{L}$ corresponds to a box in $\mathbb{R}^d$. In this case, the set of posteriors

$$\mathscr{M}_p = \left\{N\left(x; \tfrac{V\ell+n\hat{y}_n}{n}, \frac{1}{n}V\right),\ \ \ell \in \mathbb{L}\right\}$$

includes Normal PDFs with variance V/n and means free to vary in a hyper-rectangle whose vertices are the extremes of $\frac{\ell+n\hat{y}_n}{n}$ for $\ell_{\cdot i} \in [-c_{\cdot i}, c_{\cdot i}]$, which implies:

$$E[X_{\cdot i}|n,\hat{y}_n] \in \left[\frac{-c_{\cdot i}+n\hat{y}_{ni}}{n}, \frac{c_{\cdot i}+n\hat{y}_{ni}}{n}\right]. \tag{19}$$

Observe that, provided that all the components of x are observed, $\mathscr{M}_p$ includes only proper priors. Thus (19) follows by Corollary 4.7. Conversely, if some component of x was not observed, because of marginalization invariance of the Normal distribution, (19) would hold for all observed components, while for the unobserved ones $E[X_{\cdot i}|n,\hat{y}_n] \in (-\infty,\infty)$. We are prior ignorant until we observe x_i.

Notice that the set of constraints $\ell_{\cdot i} \in [-c_{\cdot i}, c_{\cdot i}]$ for $i = 1,\ldots,k$ are equivalent to $\frac{1}{c_{\cdot i}}|\ell_{\cdot i}| \leq 1$ for $i = 1,\ldots,k$ and can be replaced by the unique constraint $\max_i \frac{1}{c_{\cdot i}}|\ell_{\cdot i}| \leq 1$ or, equivalently, $||C^{-1}\ell||_\infty \leq 1$, where $\ell = [\ell_{\cdot 1},\ldots,\ell_{\cdot k}]^T$ and $C = diag(c_{\cdot 1},\ldots,c_{\cdot k})$. From Theorem 4.6, it follows that prior ignorance and translation invariance are also satisfied in the case $||C^{-1}\ell||_\infty \leq 1$ is replaced by $||C^{-1}\ell||_1 \leq 1$ or $||C^{-1}\ell||_2 \leq 1$. It can also be noticed that, in the case the components of x have the same "physical" meaning, there is no reason to use different values of the parameters $c_{\cdot i}$, which can be thus set all equal, $c_{\cdot i} = c$. Furthermore, when $c_{\cdot i} = c$ and $||C^{-1}\ell||_p \leq 1$ for some p-norm, the resulting set of priors is also invariant to permutations of the components of x. This means that all the components have the same posterior imprecision, i.e., $\overline{E}[x_{\cdot i}|n,\hat{y}_n] - \underline{E}[x_{\cdot i}|n,\hat{y}_n] = 2c/n$.

4.2 *Categorical Distribution*

In the multinomial case, $\prod_{i=1}^{n} p(y_i|x) = \prod_{i=1}^{n} Ca(y_i;x)$ with $x \in (0,1)^d$, $\sum_i x_{\cdot i} = 1$ and $y_{ij} \in \{0,1\}$, the set of posteriors obtained based on the parametrization [17, Ch.1] assuming $\mathbb{L} = \{\ell \in \mathbb{R}^d : \ell_{\cdot i} \in [-c_{\cdot i}, c_{\cdot i}],\ c_{\cdot i} > 0,\ i = 1,\dots,k\}$ is:

$$\mathscr{M}_p = \left\{ \theta_{\cdot 1}^{n\hat{y}_{n1}+\ell_{\cdot 1}-1} \theta_{\cdot 2}^{n\hat{y}_{n2}+\ell_{\cdot 2}-1} \cdots \theta_{\cdot d}^{n(1-\sum_{i=1}^{d-1}\hat{y}_{ni})-\sum_{i=1}^{d-1}\ell_{\cdot i}-1},\ \ell_{\cdot i} \in [-c_{\cdot i}, c_{\cdot i}] \right\}. \tag{20}$$

Observe that in this case Theorem 4.6 can be applied since $\mathscr{W} = \mathbb{R}^{d-1}$ and thus $\mathscr{W} = \mathbb{R}^k$ with $k = d-1$. It can be noticed that in (20) the exponents of the components $\theta_{\cdot i}$ with $i \neq d$ can vary between $n\hat{y}_{ni} - c_{\cdot i} - 1$ and $n\hat{y}_{ni} + c_{\cdot i} - 1$, while the exponent of the component $\theta_{\cdot d}$ can vary between $n(1-\sum_{i=1}^{d-1}\hat{y}_{ni}) - \sum_{i=1}^{d-1} c_{\cdot i} - 1$ and $n(1-\sum_{i=1}^{d-1}\hat{y}_{ni}) + \sum_{i=1}^{d-1} c_{\cdot i} - 1$. This means that the imprecision on $\theta_{\cdot d}$ is larger than the imprecision on the other components even if $c_{\cdot i} = c$ for $i = 1,\dots,d-1$. This asymmetry is due to the choice of the parameter transformation [17, Ch. 1], a different parameter transformation would result in asymmetry for a different component of θ. This is not desirable for a model of prior ignorance in which, because of lack of prior information, we expect the ignorance to be *symmetric*.

A way to overcome this problem is that of imposing to the set of priors not only to satisfy translation invariance but also permutation invariance: that is, that the set of priors $\mathscr{M}$ is invariant to permutations of the components of θ.

Corollary 4.8. *Consider Theorem 4.6 in the multinomial case, the set of posteriors satisfy (A1)–(A4) and permutation invariance of the components of θ provided that $\mathbb{L}$ is chosen as follows:*

$$\mathbb{L} = \left\{ \ell \in \mathbb{R}^d : \ell_{\cdot i} \in [-c, c],\ \sum_{i=1}^{d-1} \ell_{\cdot i} \in [-c, c],\ i = 1,\dots,k \right\}.$$

This follows by Theorem 4.6 imposing the additional constraint of permutation invariance. The above choice of $\mathbb{L}$ implies that the set of priors is:

$$\mathscr{M}_{SYM} = \left\{ \theta_{\cdot 1}^{\ell_{\cdot 1}-1} \theta_{\cdot 2}^{\ell_{\cdot 2}-1} \cdots \theta_{\cdot d}^{-\sum_{i=1}^{d-1}\ell_{\cdot i}-1},\ \ell_{\cdot i} \in [-c, c],\ \sum_{i=1}^{d-1} \ell_{\cdot i} \in [-c, c] \right\}, \tag{21}$$

which satisfies permutation invariance of the components of θ. For instance in the case $d = 3$, the PDFs with extreme exponents that belong to (21) are of the following type: $\theta_{\cdot 1}^{c-1}\theta_{\cdot 2}^{-c-1}\theta_{\cdot 3}^{-1}$, $\theta_{\cdot 1}^{c-1}\theta_{\cdot 2}^{-1}\theta_{\cdot 3}^{c-1}$, $\theta_{\cdot 1}^{-c-1}\theta_{\cdot 2}^{c-1}\theta_{\cdot 3}^{-1}$, $\theta_{\cdot 1}^{-c-1}\theta_{\cdot 2}^{-1}\theta_{\cdot 3}^{-c-1}$, $\theta_{\cdot 1}^{-1}\theta_{\cdot 2}^{-c-1}\theta_{\cdot 3}^{c-1}$ and $\theta_{\cdot 1}^{-1}\theta_{\cdot 2}^{c-1}\theta_{\cdot 3}^{-c-1}$. For the case $d = 4$, one has $\theta_{\cdot 1}^{c-1}\theta_{\cdot 2}^{-c-1}\theta_{\cdot 3}^{c-1}\theta_{\cdot 4}^{-c-1}$, $\theta_{\cdot 1}^{c-1}\theta_{\cdot 2}^{c-1}\theta_{\cdot 3}^{-c-1}\theta_{\cdot 4}^{-c-1}$, etc. This means that $\sum_{i=1}^{d} s_{\cdot i} c = 0$, where $s_{\cdot i}$ is the sign of c as it appears in the exponent of $\theta_{\cdot i}$.

Corollary 4.9. *The lower and upper posterior means resulting from (21) are:*

$$\underline{E}[\theta_{\cdot i}|n,\hat{y}_n] = \max\left(0, \frac{-c + n\hat{y}_{ni}}{n}\right),\ \ \overline{E}[\theta_{\cdot i}|n,\hat{y}_n] = \min\left(1, \frac{c + n\hat{y}_{ni}}{n}\right)\ \forall\, i.$$

In general, the lower and upper mean of $\sum_{i\in J}\theta_{\cdot i}$, where J is a subset of $\{1,\dots,d\}$, is equal to:

$$\begin{aligned} \underline{E}[\textstyle\sum_{i\in J}\theta_{\cdot i}|n,\hat{y}_n] &= \max\left(0, \tfrac{1}{n}\left[\textstyle\sum_{i\in J} n\hat{y}_{ni} + \max\left(-c|J|, -c(d-|J|)\right)\right]\right), \\ \overline{E}[\textstyle\sum_{i\in J}\theta_{\cdot i}|n,\hat{y}_n] &= \min\left(1, \tfrac{1}{n}\left[\textstyle\sum_{i\in J} n\hat{y}_{ni} + \min\left(c|J|, c(d-|J|)\right)\right]\right), \end{aligned} \tag{22}$$

with $|J|$ denoting the cardinality of J.

Proof. Consider the general form (22), the posterior kernel is:

$$p(\theta|n,\hat{y}_n) \propto \theta_{\cdot 1}^{n\hat{y}_{n1}+\ell_{\cdot 1}-1}\theta_{\cdot 2}^{n\hat{y}_{n2}+\ell_{\cdot 2}-1}\cdots\theta_{\cdot d}^{n(1-\sum\hat{y}_{ni})-\sum_{i=1}^{d-1}\ell_{\cdot i}-1}.$$

We prove it for $J=\{1,2\}$ and $d=4$ but the steps can be repeated in general. We need to compute

$$\begin{aligned}&\tfrac{1}{K}\int_0^1\int_0^{1-\theta_{\cdot 1}}(\theta_{\cdot 1}+\theta_{\cdot 2})\theta_{\cdot 1}^{n\hat{y}_{n1}+\ell_{\cdot 1}-1}\theta_{\cdot 2}^{n\hat{y}_{n2}+\ell_{\cdot 2}-1}\\&\int_0^{1-\theta_{\cdot 1}-\theta_{\cdot 2}}\theta_{\cdot 3}^{n\hat{y}_{n3}+\ell_{\cdot 3}-1}(1-\theta_{\cdot 1}-\theta_{\cdot 2}-\theta_{\cdot 3})^{n(1-\sum_{i=1}^3\hat{y}_{ni})-\sum_{i=1}^3\ell_{\cdot i}-1}d\theta_{\cdot 1}d\theta_{\cdot 2}\theta_{\cdot 3},\end{aligned}$$

where K is the normalization constant. Now consider the transformation of coordinate $\theta'_{\cdot 3}=\theta_{\cdot 3}/(1-\theta_{\cdot 1}-\theta_{\cdot 2})$ (in general $\theta'_{\cdot k}=\theta_{\cdot k}/(1-\sum_{i\in J}\theta_{\cdot i})$ for $k\notin J$):

$$\begin{aligned}&\tfrac{1}{K}\int_0^1\int_0^{1-\theta_{\cdot 1}}(\theta_{\cdot 1}+\theta_{\cdot 2})\theta_{\cdot 1}^{n\hat{y}_{n1}+\ell_{\cdot 1}-1}\theta_{\cdot 2}^{n\hat{y}_{n2}+\ell_{\cdot 2}-1}\\&(1-\theta_{\cdot 1}-\theta_{\cdot 2})^{n(1-\hat{y}_{n1}-\hat{y}_{n2})-\sum_{i=1}^2\ell_{\cdot i}-1}\int_0^1\theta'^{\,n\hat{y}_{n3}+\ell_{\cdot 3}-1}_{\cdot 3}(1-\theta_{\cdot 3})^{n(1-\sum\hat{y}_{ni})-\sum_{i=1}^3\ell_{\cdot i}-1}d\theta_{\cdot 1}d\theta_{\cdot 2}\theta_{\cdot 3}\\&=\tfrac{1}{K}\int_0^1\int_0^{1-\theta_{\cdot 1}}(\theta_{\cdot 1}+\theta_{\cdot 2})\theta_{\cdot 1}^{n\hat{y}_{n1}+\ell_{\cdot 1}-1}\theta_{\cdot 2}^{n\hat{y}_{n2}+\ell_{\cdot 2}-1}(1-\theta_{\cdot 1}-\theta_{\cdot 2})^{n(1-\hat{y}_{n1}-\hat{y}_{n2})-\sum_{i=1}^2\ell_{\cdot i}-1}d\theta_{\cdot 1}d\theta_{\cdot 2}.\end{aligned}$$

Now consider the transformation of coordinates $\theta_{\cdot 1}=uv$, $\theta_{\cdot 2}=u(1-v)$ with $0<u,v<1$ (in general $\theta_{\cdot i}=uv_i\ \forall i\in J$ and $1-\sum_{i\in J}\theta_{\cdot i}=u(1-\sum_{i\in J}v_i)$):

$$\begin{aligned}&\tfrac{1}{K}\int_0^1\int_0^{1-\theta_{\cdot 1}}(\theta_{\cdot 1}+\theta_{\cdot 2})\theta_{\cdot 1}^{n\hat{y}_{n1}+\ell_{\cdot 1}-1}\theta_{\cdot 2}^{n\hat{y}_{n2}+\ell_{\cdot 2}-1}(1-\theta_{\cdot 1}-\theta_{\cdot 2})^{n(1-\hat{y}_{n1}-\hat{y}_{n2})-\sum_{i=1}^2\ell_{\cdot i}-1}d\theta_{\cdot 1}d\theta_{\cdot 2}\\&=\tfrac{1}{K}\int_0^1\int_0^1 u^{n\hat{y}_{n1}+\ell_{\cdot 1}+n\hat{y}_{n2}+\ell_{\cdot 2}-1}(1-u)^{n(1-\hat{y}_{n1}-\hat{y}_{n2})-\sum_{i=1}^2\ell_{\cdot i}-1}v^{n\hat{y}_{n1}+\ell_{\cdot 1}-1}(1-v)^{n\hat{y}_{n2}+\ell_{\cdot 2}-1}dudv\\&=\tfrac{1}{K}\int_0^1 u^{n\hat{y}_{n1}+\ell_{\cdot 1}+n\hat{y}_{n2}+\ell_{\cdot 2}-1}(1-u)^{n(1-\hat{y}_{n1}-\hat{y}_{n2})-\sum_{i=1}^2\ell_{\cdot i}-1}du.\end{aligned}$$

Note that we are reduced to a univariate expectation w.r.t. a Beta distribution. We can therefore apply Corollary 4.3, with the constraints $\ell_{\cdot i}\in[-c,c]$, $\sum_{i=1}^{d-1}\ell_{\cdot i}\in[-c,c]$, to obtain (22). □

It can be noticed that, fixed J, the posterior imprecision of $\sum_{i\in J}\theta_{\cdot i}$ increases with d, i.e., the components of θ. For some problem this can be undesirable, especially because the choice of the number of components of θ is often arbitrary. We would like to make the posterior imprecision invariant to the number of components of θ.

Corollary 4.10. *Consider Theorem 4.6. The set of posteriors satisfy (A1)–(A4), permutation invariance of the components of θ and invariance of the posterior imprecision w.r.t. the number of components of θ provided that $\mathbb{L}$ is chosen as follows:*

$$\mathbb{L}=\left\{\ell\in\mathbb{R}^d:\ ||\ell||_1\leq 2c,\ \sum_{i=1}^{d-1}\ell_{\cdot i}\in[-c,c]\right\},$$

where $\ell=[\ell_{\cdot 1},\ldots,\ell_{d-1}]^T$.

This follows by Theorem 4.6 imposing the additional constraint of permutation invariance and invariance w.r.t. the number of components of θ. This implies that the set of priors in this case

is:

$$\mathscr{M}_{RIP} = \left\{ \theta_{\cdot 1}^{\ell_{\cdot 1}-1} \theta_{\cdot 2}^{\ell_{\cdot 2}-1} \cdots \theta_{\cdot d}^{-\sum_{i=1}^{d-1} \ell_{\cdot i}-1},\ ||\ell||_1 \leq 2c,\ \sum_{i=1}^{d-1} \ell_{\cdot i} \in [-c,c] \right\}. \tag{23}$$

For the case $d = 2$ and $d = 3$, (23) coincides with (21), but in the case $d = 4$, the PDFs with extreme exponents are $\theta_{\cdot 1}^{c-1}\theta_{\cdot 2}^{-c-1}\theta_{\cdot 3}^{-1}\theta_{\cdot 4}^{-1}$, $\theta_{\cdot 1}^{c-1}\theta_{\cdot 2}^{-1}\theta_{\cdot 3}^{-c-1}\theta_{\cdot 4}^{-1}$, $\theta_{\cdot 1}^{c-1}\theta_{\cdot 2}^{-1}\theta_{\cdot 3}^{-1}\theta_{\cdot 4}^{-c-1}$, $\theta_{\cdot 1}^{-1}\theta_{\cdot 2}^{c-1}\theta_{\cdot 3}^{-c-1}\theta_{\cdot 4}^{-1}$ etc. In this case, there are at most two components whose exponents assume value $\pm c-1$. With this constraint, the set of posteriors satisfies the one-step ahead representation invariance principle (RIP) [6], i.e., the posterior upper and lower probabilities assigned to the event "the next observation belongs to the categories in J" should not depend on the sample space in which the event and the previous observations are represented. In fact, one has that:

$$\underline{E}[\textstyle\sum_{i\in J} \theta_{\cdot i}|n,\hat{y}_n] = \max\left(0, \frac{1}{n}\left[\sum_{i\in J} n\hat{y}_{ni} - c\right]\right),\ \ \overline{E}[\textstyle\sum_{i\in J} \theta_{\cdot i}|n,\hat{y}_n] = \min\left(1, \frac{1}{n}\left[\sum_{i\in J} n\hat{y}_{ni} + c\right]\right), \tag{24}$$

and, thus, it can be noticed that the difference between the upper and lower expectation does not depend on the number of categories in J.

4.3 *How to Choose Parameters* $c_{\cdot i}$

The aim of this section is to give guidelines for the choice of the parameters $c_{\cdot i}$. The parameters $c_{\cdot i}$ determine the precision of posterior inferences and, thus, their robustness. To understand the meaning of $c_{\cdot i}$, we can compare the expression of the posterior expectation of $\partial b/\partial w_{\cdot i}$ obtained by considering a standard Bayesian analysis using the conjugate prior $\exp(n_0(y_0^T w - b(w)))$ and the lower and upper posterior expectations obtained using our model, i.e.,

$$\frac{n\hat{y}_{ni} + n_0 y_{0i}}{n + n_0} \quad \longleftrightarrow \quad \frac{n\hat{y}_{ni} \pm c_{\cdot i}}{n}. \tag{25}$$

Looking at the expression on the left, it is evident that the prior parameter n_0 attached to the prior mean y_{0i} plays the same role to the sample size n attached to the data mean $\hat{y}_{ni}$. Since $n\hat{y}_{ni} = \sum_j y_{ji}$, we can also interpret $n_0 y_0$ as the sum of n_0 pseudo-observations that are added to $\sum_j y_{ji}$. Matching the two expectations in (25), it implies that $|n_0 y_{0i}| = c_{\cdot i}$ and $n_0 = 0$. In the case on the right we are not adding additional (pseudo) observations, but we are considering a scenario in which an amount $c_{\cdot i}$ is summed/subtracted to the sum $\sum_j y_{ji}$ but without changing the sample size. For instance, when y_{ij} are frequency data from an election poll (this example will be discussed in Section 8.2), we can interpret the lower and upper expectations as the result of a swing scenario in which we test what happens if $c_{\cdot i}$ votes are subtracted (lower) to a candidate and assigned to another one (upper). The interpretation of $c_{\cdot i}$ as a pseudo-observation is the main avenue to be followed for the choice of a value for $c_{\cdot i}$.

The choice of $c_{\cdot i}$ can also be based on measures of posterior robustness such as [6]: (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. We will give an example of this approach in Section 8.1. Another possible requirement for the choice of $c_{\cdot i}$ is that the family of priors $\mathscr{M}$ should be large enough to encompass frequentist or objective Bayesian inferences, but not too large to avoid obtaining too weak inferences. These are the approaches that we have followed in [13] for the univariate one-parameter exponential family. In the multivariate case the only difference is that the value of k parameters have to be fixed. However, as discussed previously, for symmetry reasons in many cases it can be assumed $c_{\cdot i} = c$ for (all) a subset of parameters $i = 1,\ldots,k$. Finally, we

can choose c to have some desirable frequentist properties (e.g., robust credible intervals to be calibrated frequentist intervals, hypothesis tests to be calibrated for the Type I error, etc.). We may also place an upper bound on the value of c based on considerations regarding the power of the hypothesis test.

5. Prior Near-Ignorance: Some Further Useful Properties

Notice that in all the above examples the set of priors (16) guarantees prior ignorance w.r.t. ∇b. This choice has also been motivated by the meaning of ∇b for exponential families. Remember in fact from Section 3 that ∇b (equal to b' in the univariate case) 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. Furthermore, to state prior ignorance, invariance, learning and converge in the natural space $\mathscr{W}$ allows us to treat jointly all the members of the exponential families and, thus, to derive general properties. The natural parametrization is in fact a preferred parametrization for exponential families. In this section, it will be shown that the set of priors (16) satisfies (A2) also for other functions of interest in statistical analysis (i.e., one- and two-sided hypotheses testing and credible intervals). Therefore, the choice of imposing prior ignorance only on $\mathscr{G}_0 = \{\nabla b\}$, is not very restrictive for exponential families.

Corollary 5.1. *The family of priors $\mathscr{M}$ in Theorem 4.6 satisfies the following properties:*

(B1) *For each ball $B_\zeta(\mathscr{W}) = \{w : ||w - w_0|| \leq \zeta\} \subset \mathscr{W}$, $w_0 \in \mathscr{W}$ and $\zeta > 0$, it holds that $\underline{E}[I_{B_\zeta(\mathscr{W})}] = 0$ and $\overline{E}[I_{B_\zeta(\mathscr{W})}] = 0$ but $\underline{E}[\mathbb{R}^d] = 1$.*

(B2) *For each subset $\mathscr{W}' \subset \mathscr{W}$ such that $\mathscr{W}' = \times_i \Pi_i(\mathscr{W}')$, where $\Pi_i(\mathscr{W}') = [a_i, \sup \Pi_i(\mathscr{W}))$ or $\Pi_i(\mathscr{W}') = (\inf \Pi_i(\mathscr{W}), a_i]$ and $a_i \in \mathbb{R}$, with $\Pi_i(\mathscr{W}')$ denoting the orthogonal projection of $\mathscr{W}'$ w.r.t. the i-th component, then $\underline{E}[I_{\{\mathscr{W}'\}}] = 0$ and $\overline{E}[I_{\{\mathscr{W}'\}}] = 1$.*

Proof. The first part of (B1) follows straightforwardly from the fact the set of priors in Theorem 4.6 concentrate their mass on the border of $\mathscr{W}$. In fact, the set of priors is composed by the following truncated densities

$$\exp(\ell^T w) \prod_{i=1}^{k} I_{\mathscr{W}_{r_{\cdot i}}}(w_{\cdot i}) \frac{\ell_{\cdot i}}{\exp(\ell_{\cdot i} r_{\cdot i})}.$$

Then, for any ball in $\mathscr{W}$, it follows that

$$\overline{E}[I_{B_\zeta(\mathscr{W})}] = \limsup_{r_{\cdot 1} \to \infty} \ldots \limsup_{r_{\cdot k} \to \infty} \int \cdots \int I_{B_\zeta(\mathscr{W})} \exp(\ell^T w) \prod_{i=1}^{k} I_{\mathscr{W}_{r_{\cdot i}}}(w_{\cdot i}) \frac{\ell_{\cdot i}}{\exp(\ell_{\cdot i} r_{\cdot i})} dw_{\cdot i} = 0,$$

since the integral of $I_{B_\zeta(\mathscr{W})}$ is bounded and $1/\exp(\ell_{\cdot i} r_{\cdot i})$ goes to zero for $|r_{\cdot i}| \to \infty$. Concerning (B2), $\mathscr{W}'$ is the Cartesian product of $\Pi_i(\mathscr{W}')$. Thus, if we take $\ell_{\cdot i} > 0$ for $\Pi_i(\mathscr{W}') = [a_i, \sup \Pi_i(\mathscr{W}))$ or $\ell_{\cdot i} < 0$ for $\Pi_i(\mathscr{W}') = (\inf \Pi_i(\mathscr{W}), a_i]$ for all $i = 1, \ldots, k$, then $\overline{E}[I_{\{\mathscr{W}'\}}] = 1$. While if we take $\ell_{\cdot i} > 0$ for $\Pi_i(\mathscr{W}') = (\inf \Pi_i(\mathscr{W}), a_i]$ or and $\ell_{\cdot i} < 0$ for $\Pi_i(\mathscr{W}') = [a_i, \sup \Pi_i(\mathscr{W}))$ for some i, then $\underline{E}[I_{\{\mathscr{W}'\}}] = 0$. The second part of (B1) follows from the fact that any truncated prior integrates on its support to 1 and so it is the limit when its support tends to $\mathscr{W}$. □

Corollary 5.1 has several important implications in statistical inference. Since it is common to represent credible regions with balls $B_\zeta(\mathscr{W}) \subset \mathscr{W}$, from (B1) it follows that a priori for any $\zeta < \infty$, $\underline{E}[I_{B_\zeta(\mathscr{W})}] = 0$. That is, the lower probability of the true w_0 to be in $B_\zeta(\mathscr{W})$ is zero, and is only greater than zero for $\zeta \to \infty$. The only convex set that has prior lower probability greater than zero is $\mathscr{W}$, which means that a priori we only known that $w_0 \in \mathscr{W}$. Concerning (B2), it

implies that the lower and upper distribution of w are vacuous and also that the marginal lower and upper distribution of any component $w_{\cdot i}$ are vacuous.

6. The Case $\mathscr{W} \neq \mathbb{R}^k$

In Theorems 4.2–4.6 we have defined a model of prior ignorance which satisfies (A1)–(A4) in case $\mathscr{W} = \mathbb{R}^k$. For instance in case $k = 1$, we have seen that translation invariance and prior ignorance can be satisfied because the set of priors $\mathscr{M}$ includes limits of truncated increasing/decreasing exponentials which, at the limit, accumulate to $\pm\infty$. However, for many exponential families of interest it holds that $\mathscr{W} \neq \mathbb{R}^k$, e.g., the centred-normal, multivariate-normal with unknown variance and the exponential distribution, etc. Since $\mathscr{W}$ is semi-bounded in these three cases, translation invariance is not well defined. For example, in case $\mathscr{W} = \mathbb{R}^-$ then $w + a \notin \mathscr{W}$ for $a > 0$ and $|w| < a$. Also prior ignorance does not hold in these cases for the set of priors (16). In fact, for $\mathscr{W} = \mathbb{R}^-$ the prior $\exp(\ell w)$ accumulates to minus infinity for $\ell < 0$, but it is integrable for $\ell > 0$, since $\int_{-\infty}^{0} \exp(\ell w) = 1/\ell$ for $\ell > 0$. Thus, in these cases, properties (A1)–(A2) cannot be satisfied by the set of priors (16). Observe in fact that for the centred normal $b' = \lambda^{-1}$ and $\int_{-\infty}^{0} \lambda^{-1} \exp(-\ell\lambda) d\lambda = \infty$ for both $\ell > 0$ and $\ell < 0$. Thus, prior ignorance is not satisfied in this case, being $\inf b' = 0$ and $\sup b' = \infty$. Translation invariance is also not satisfied for $\ell > 0$, since $0 \neq \int_A \exp(-\ell\lambda) d\lambda \neq \int_{A+a} \exp(-\ell\lambda) d\lambda \neq 0$ for any bounded set $A \subset \mathbb{R}^+$ and $a > 0$. How can we impose prior near-ignorance in case $\mathscr{W} \neq \mathbb{R}^k$?

In Proposition 3.2, it is stated that the prior kernel $\exp(n_0 y_0 w - n_0 b(w))$ is integrable if $y_0 \in \mathscr{Y}_0$ and $n_0 > 0$. Diaconis and Ylvisaker in [20] give more conditions on the integrability of $\exp(n_0 y_0 w - n_0 b(w))$. In particular, they prove the following result: (i) if $n_0 > 0$ and $y_0 \in \mathscr{Y}_0$ then $\exp(n_0 y_0 w - n_0 b(w))$ is integrable; (ii) if $\exp(n_0 y_0 w - n_0 b(w))$ is integrable and $\mathscr{W} = \mathbb{R}^d$ then $n_0 > 0$ and (iii) if $\exp(n_0 y_0 w - n_0 b(w))$ is integrable and $n_0 > 0$, then $y_0 \in \mathscr{Y}_0$.

Thus, even in the case $y_0 \in \mathscr{Y}_0$, $\exp(n_0 y_0 w - n_0 b(w))$ may not be integrable if $n_0 < 0$ (for $\mathscr{W} = \mathbb{R}^d$ the kernel is certainly not integrable if $n_0 < 0$, see condition (ii) of Diaconis and Ylvisaker, while for $\mathscr{W} \neq \mathbb{R}^d$ there is not a general result, i.e., the integrability of the kernel in case $n_0 < 0$ depends on the value of n_0 and on the particular exponential family). This means that for some value of $n_0 < 0$, it can happen that $\int_{\mathscr{W}} \exp(n_0 y_0 w - n_0 b(w)) dw = \infty$. Therefore, in this case, if we can find a sequence of subsets $\mathscr{W}_1, \mathscr{W}_2, \ldots, \mathscr{W}_r$ of $\mathscr{W}$ converging to $\mathscr{W}$ from below and such that the truncated prior satisfies

$$\int_{\mathscr{W}} I_{\mathscr{W}_i} \exp(n_0 y_0 w - n_0 b(w)) dw < \infty,$$

for any $\mathscr{W}_i$, then we can define a sequence of countably additive truncated densities on $\mathscr{W}_r$ whose lower limit for $r \to \infty$ defines a lower expectation model as in (9). Since $\overline{E}[A] = 0$ for any bounded set A (as it is shown in Appendix A), the set of priors $\mathscr{M}$ associated to this lower expectation model includes only finitely additive probabilities.

Therefore, for $\mathscr{W} \neq \mathbb{R}^d$ we may be able to satisfy (A1)–(A4) with a family of finitely additive probabilities obtained as limit of truncated conjugate exponential priors $\exp(n_0 y_0 w - n_0 b(w))$ obtained by choosing $\ell = n_0 y_0 \in \mathbb{L}$ and $n_0 < 0$. In the following sections, we show that this is possible for the centred-normal and multivariate-normal with unknown variance.[1] Our conjecture is that this result may be extended to other distributions in the exponential families by following the same approach. However, we cannot develop a general theory because the non-integrability of $\exp(n_0 y_0 w - n_0 b(w))$ for $\mathscr{W} \neq \mathbb{R}^d$ and $n_0 < 0$ depends on the value of n_0 and on the particular exponential family. Thus, a general treatment does not seem to be possible.

[1] This also holds for the exponential distribution for $n_0 < -1$.

6.1 *Centred-Normal Distribution*

Assume that the likelihood is a centred-normal distribution and consider the following set of priors

$$\mathscr{M} \propto \left\{ \lambda^{-\frac{\nu}{2}} \exp(-\ell\tfrac{\lambda}{2}),\ \ell \in [-c,c] \right\}, \tag{26}$$

defined by the parameters $c > 0$, $\nu \in \mathbb{R}^+$ and $n_0 = -\nu < 0$. It can be verified that for any value of ℓ and $\nu \geq 2$ all the kernels in (26) are not integrable. The kernels are integrable in $\mathscr{W}_r = [1/r, r]$ which, for $r \to \infty$, defines an infinite sequence of subsets of $\mathbb{R}^+$ converging from below to $\mathbb{R}^+$. By truncating these kernels in $[1/r, r]$ and taking the limit of the expectations for $r \to \infty$, one can verify that for $\nu > 2$ and $\ell > 0$

$$\overline{E}[\lambda^{-1}] = \limsup_{r\to\infty} \frac{\int_{1/r}^{r} \lambda^{-\frac{\nu+2}{2}} \exp(-\ell\frac{\lambda}{2}) d\lambda}{\int_{1/r}^{r} \lambda^{-\frac{\nu}{2}} \exp(-\ell\frac{\lambda}{2}) d\lambda} = \infty,$$

while for $\ell < 0$

$$\underline{E}[\lambda^{-1}] = \liminf_{r\to\infty} \frac{\int_{1/r}^{r} \lambda^{-\frac{\nu+2}{2}} \exp(-\ell\frac{\lambda}{2}) d\lambda}{\int_{1/r}^{r} \lambda^{-\frac{\nu}{2}} \exp(-\ell\frac{\lambda}{2}) d\lambda} = 0.$$

Notice that we are interested in the expectation of λ^{-1} because $1/\lambda = \sigma^2$. For any bounded subset A of $\mathbb{R}^+$, it holds that $\overline{E}[I_{\{A+a\}}] = \overline{E}[I_{\{A\}}] = 0$ for any $a > 0$ (which is a weaker version of translation invariance). Finally, a-posteriori by combining the centred-normal likelihood with the priors in (26), one obtains

$$\mathscr{M} \propto \left\{ \lambda^{\frac{n-\nu+2}{2}-1} \exp\left(-n\frac{\ell + n\hat{y}_n}{n}\frac{\lambda}{2} \right),\ \ell \in [-c,c] \right\}. \tag{27}$$

It can be noticed that for $n > \nu$ and $\frac{\ell+n\hat{y}_n}{n} > 0$ for all $\ell \in [-c,c]$, all the kernels in (27) are integrable and the expected value of $1/\lambda = \sigma^2$ is equal to $(\ell + n\hat{y}_n)/(n-\nu)$ in this case. Thus, it follows that for $n > \nu$:

$$\underline{E}[\lambda^{-1}|n,\hat{y}_n] = \max\left(0, \tfrac{-c+n\hat{y}_n}{n-\nu}\right),\ \ \overline{E}[\lambda^{-1}|n,\hat{y}_n] = \tfrac{c+n\hat{y}_n}{n-\nu}, \tag{28}$$

while if $n < \nu$, the posterior expectation of λ^{-1} is vacuous. Thus, there exists a $\delta > 0$ such that for any $n > \delta$ it holds that $n > \nu$, which implies that properties (A3)–(A4) hold for inferences derived from (26).

6.2 *Normal with Unknown Mean and Variance*

In the case the variance is unknown, $\prod_{i=1}^n p(y_i|x) = \prod_{i=1}^n N(y_i; x, \sigma^2)$, with $\sigma^2 > 0$ and $x, y_i \in \mathbb{R}$ (the following results can be easily generalized to the case $x, y_i \in \mathbb{R}^d$). The likelihood model can

be expressed in the canonical form by setting $w = [\lambda x, -\frac{1}{2}\lambda]^T$,

$$b(w) = \frac{\lambda x^2 - \ln(\lambda)}{2}, \;\; \nabla b(w) = [x, x^2 + \lambda^{-1}]^T,$$

where $\lambda = 1/\sigma^2$ and $\mathscr{W} = [\mathbb{R}, \mathbb{R}^-]^T$. Unfortunately in this case $\mathscr{W} \neq \mathbb{R}^2$ and, thus, Theorem 4.6 cannot be applied directly. However, this issue can be solved by instead considering the kernel

$$\mathscr{M} = \left\{ \exp(x\lambda\ell_{\cdot 1})\lambda^{\frac{1}{2}} \exp(-\tfrac{1}{2}\lambda\ell_{\cdot 2})\lambda^{-\frac{\nu}{2}}, \; \ell_{\cdot i} \in [-c_{\cdot i}, c_{\cdot i}] \right\},$$

with $\nu \geq 2$. Notice in fact that $\exp(-\frac{1}{2}\lambda\ell_{\cdot 2})\lambda^{-\frac{\nu}{2}}$ is the same model discussed in (26). The likelihood can be rewritten as:

$$p(data|x,\lambda) \propto \exp(n\hat{y}_{n1}\lambda x - n\tfrac{1}{2}\lambda x^2) \exp(-n\hat{y}_{n2}\tfrac{1}{2}\lambda)\lambda^{\frac{n}{2}},$$

where $n\hat{y}_{n1} = \sum_{i=1}^n y_i$ and $n\hat{y}_{n2} = \sum_{i=1}^n y_i^2$. By multiplication of likelihood and prior, one gets:

$$p(data|x,\lambda)p(x,\lambda) \propto \lambda^{\frac{1}{2}} \exp((n\hat{y}_{n1} + \ell_{\cdot 1})\lambda x - n\tfrac{1}{2}\lambda x^2) \exp(-(n\hat{y}_{n2} + \ell_{\cdot 2})\tfrac{1}{2}\lambda)\lambda^{\frac{n-\nu+2}{2}-1},$$

which gives the posterior

$$N\left(x; \hat{y}_{p1}, (n\lambda)^{-1}\right) G(\lambda; \tfrac{n-\nu+2}{2}, \tfrac{n}{2}(\hat{y}_{p2} - \hat{y}_{p1}^2)), \tag{29}$$

where $\hat{y}_{p1} = \frac{n\hat{y}_{n1} + \ell_{\cdot 1}}{n}$ and $\hat{y}_{p2} = \frac{n\hat{y}_{n2} + \ell_{\cdot 2}}{n}$. Observe that (29) makes sense only if $\hat{y}_{p2} - \hat{y}_{p1}^2 > 0$ and $n - \nu + 2 > 0$, i.e., the posterior is a proper Normal-Gamma density. For $n - \nu + 2 > 0$, by marginalizing out λ in (29), one can compute the posterior distribution of X

$$p(x|data) \propto \left[1 + \frac{(x - \hat{y}_{p1})^2}{\hat{y}_{p2} - \hat{y}_{p1}^2} \right]^{-(n-\nu+3)/2}, \tag{30}$$

which is a Student t distribution. Thus, for $n - \nu + 2 > 0$, it follows that

$$\underline{E}[X|n,\hat{y}_n] = \frac{n\hat{y}_{n1} - c_{\cdot 1}}{n}, \;\; \overline{E}[X|n,\hat{y}_n] = \frac{n\hat{y}_{n1} + c_{\cdot 1}}{n},$$

$$\begin{aligned}
\underline{E}[\lambda^{-1}|n,\hat{y}_n] &= \max\left(0, \tfrac{n}{n-\nu}\left[\left(\tfrac{\sum_{i=1}^n y_i^2 - c_{\cdot 2}}{n}\right) - \max_{|c|<c_{\cdot 1}}\left(\tfrac{\sum_{i=1}^n y_i + c}{n}\right)^2\right]\right), \\
\overline{E}[\lambda^{-1}|n,\hat{y}_n] &= \tfrac{n}{n-\nu}\left(\tfrac{\sum_{i=1}^n y_i^2 + c_{\cdot 2}}{n} - \min_{|c|<c_{\cdot 1}}\left(\tfrac{\sum_{i=1}^n y_i + c}{n}\right)^2\right).
\end{aligned}$$

Observe that for $\ell_{\cdot 1} = \ell_{\cdot 2} = 0$ and $\nu = 3$, one obtains $E[x|n,\hat{y}_n] = \hat{y}_{n1}$ and $E[\lambda^{-1}|n,\hat{y}_n] = \frac{1}{n-3}\sum_{i=1}^n (y_i - \hat{y}_{n1})^2$ which are the estimates of X and σ^2 one would obtain from the improper prior $p(x,\sigma^2) = (\sigma^2)^{-1}$, while for $\nu = 2$ one obtains the estimates from the improper prior $p(x,\sigma^2) = (\sigma)^{-1}$, which gives the same inferences as the frequentist ones. The advantage of this model over the ones based on improper priors is that it satisfies prior near-ignorance.

7. Comparison with Other Prior Near-Ignorance Models

The aim of this section is to compare the model in Theorems 4.2–4.6 with other near-ignorance models available in the literature.

7.1 *Near-Ignorance with a Set of Proper Priors in the One-Parameter Exponential Families*

In [13], we have defined a model of prior ignorance for one-parameter exponential families based on a set of proper conjugate priors:

$$\mathscr{M} = \left\{ k(n_0, y_0) \exp\left(n_0 \left(y_0 w - b(w)\right)\right) : \; y_0 \in \mathscr{Y}_0, \; 0 < n_0 \leq \min(\overline{n}_0, c/|y_0|) \right\}. \tag{31}$$

We have proven that this set satisfies prior ignorance (A2) w.r.t. $g = b'$, learning (A3) and convergence (A4).[1] More precisely, we have proven that prior ignorance and learning/convergence can be both satisfied provided that y_0 is free to vary in $\mathscr{Y}_0$ and $|n_0 y_0| \leq c$, i.e., n_0 must depend on y_0. It can be observed that the set of priors in Theorem 4.2 can be obtained from (31) by relaxing the constraint $y_0 \in \mathscr{Y}_0$ to $y_0 \in \mathbb{R}$ and by taking the limit of the inferences for $\overline{n}_0 \to 0$. In fact, $y_0 \notin \mathscr{Y}_0$ allows us to include at the limit also improper priors in (31), while the limit $\overline{n}_0 \to 0$ guarantees the fulfillment of (A1) as it is shown in Appendix A. Translation invariance is important because it guarantees that the posterior imprecision does not depend on the sample mean. This in general does not hold for (31).

Consider for instance the Poisson case. The lower and upper posterior means obtained from (31) and the Poisson likelihood are [13]:

$$\frac{n\hat{y}_n}{n+\overline{n}_0} \leq E[\lambda\,|n, \hat{y}_n] \leq \frac{n\hat{y}_n + c}{n}. \tag{32}$$

It can be noticed that the difference between the upper and lower means depends on $\hat{y}_n$ for any $\overline{n}_0 > 0$. In fact, it can be verified that the set of priors (31) does not satisfy translation invariance, i.e., $\overline{E}[I_{\{A\}}] \neq \overline{E}[I_{\{A+a\}}]$. Translation invariance is only satisfied if we take the limit $\overline{n}_0 \to 0$ of the posterior inferences. However, in this case, the lower and upper posterior means reduce to:

$$\hat{y}_n \leq E[\lambda\,|n, \hat{y}_n] \leq \frac{n\hat{y}_n + c}{n}, \tag{33}$$

and, thus, the lower and upper means are not symmetric w.r.t. $\hat{y}_n$: the left bound is equal to the sample mean $\hat{y}_n$; this is not desirable because it implies weakening the robustness of the posterior inferences. Conversely, by relaxing $y_0 \notin \mathscr{Y}_0$, the lower and upper posterior means at the limit $\overline{n}_0 \to 0$ become:

$$\max\left(0, \frac{n\hat{y}_n - c}{n}\right) \leq E[\lambda\,|n, \hat{y}_n] \leq \frac{n\hat{y}_n + c}{n}. \tag{34}$$

Now the lower and upper posterior means coincide with $\hat{y}_n$ only for $n \to \infty$ and the posterior imprecision depends only on n and c for a sufficiently large n. In other words, posterior robustness and independence of posterior imprecision to the sample mean can both be guaranteed with conjugate prior models only by allowing finitely additive priors in the set $\mathscr{M}$. This means that they are incompatible with countable additivity.

[1] Since the priors in $\mathscr{M}$ are all countably additive, it also satisfies strong coherence as defined in [6, Ch. 7].

7.2 *Imprecise Dirichlet Model*

The *imprecise Dirichlet model* (IDM) [21], [22] is a model of prior near-ignorance for multinomial observations. The IDM considers the following set of priors:

$$\mathscr{M} = \left\{ \frac{\Gamma(s)}{\prod_{i=1}^{d} \Gamma(st_i)} \prod_{i=1}^{d} \theta_{\cdot i}^{st_i - 1} : \; t_i > 0, \; \sum_{i=1}^{d} t_i = 1, \; s > 0 \right\}, \tag{35}$$

where $\Gamma(\cdot)$ is the Gamma function. Therefore $\mathscr{M}$ includes Dirichlet densities parametrized by t_i and s. It follows that a priori $\underline{E}[\theta_{\cdot i}] = 0$ and $\overline{E}[\theta_{\cdot i}] = 1$, while a-posteriori:

$$\underline{E}[\theta_{\cdot i} | n, \hat{y}_n] = \frac{n\hat{y}_{ni}}{n+s}, \quad \overline{E}[\theta_{\cdot i} | n, \hat{y}_n] = \frac{n\hat{y}_{ni} + s}{n+s}. \tag{36}$$

The IDM is a model of prior ignorance that guarantees conjugacy between likelihood and priors and satisfies (A2)–(A4) but it does not satisfy translation invariance (A1) in the transformed domain $\mathscr{W}$. Observe that although the IDM does not satisfy translation invariance, the posterior imprecision does not depend on the sample mean. Furthermore, the IDM is invariant to permutations of the components of θ and satisfies RIP. Thus the IDM satisfies permutation invariance and independence of the posterior imprecision to the sample mean without the need of finitely additive priors.

The question is what happens if we impose translation invariance to the IDM in $\mathscr{W}$. Fixed $s > 0$, the only way to satisfy also (A1) is not constraining t to lie in $\mathscr{Y}_0 = \{t : \; t_i > 0, \; \sum_{i=1}^{d} t_i = 1\}$. Assuming that $t_{\cdot i} \in \mathbb{R}$ and imposing the constraint $0 < |st_i| \leq c_{\cdot i} = c$ for $i = 1, \ldots, d-1$, the set of priors (35) reduces to

$$\mathscr{M} = \left\{ \prod_{i=1}^{d-1} \theta_{\cdot i}^{\ell_{\cdot i} - 1} \left(1 - \sum_{j=1}^{d-1} \theta_{\cdot j} \right)^{-\sum_{j=1}^{d-1} \ell_{\cdot j} - 1} : \; \ell_{\cdot i} \in [-c, c] \right\}, \tag{37}$$

which is the model in (20). If we impose one-step ahead RIP to (20), we obtain (23).

7.3 *Nonparametric Predictive Inference for Multinomial Data and Known Number of Categories*

In the case of multinomial data, it is worth pointing out also the relationship between the model in Theorem 4.6 and another model for robust inferences, the so-called *nonparametric predictive inference* (NPI) model [15]. The NPI is a nonparametric model based on a post-data assumption about the uncertainty associated to a future observation and can only yield predictive inferences (inferences on the space of the observations). Consider $k \geq 3$ possible categories denoted by $C_1, \ldots, C_k$. Without loss of generality, we assume that the first o of these have already been observed and the last $k - o$ have not yet been observed. There are n observations $\{y_1, \ldots, y_n\}$ such that $y_{ij} \in \{0, 1\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ and $\sum_{i,j} y_{ij} = n$. The event of interest can generally be denoted by

$$Y_{n+1} \in \bigcup_{j \in J} C_j,$$

i.e., the next observation belongs to the subset of categories $J \subseteq \{1, \ldots, k\}$. Let $OJ = J \cap \{C_1, \ldots, C_o\}$ denote the index-set for the categories in the event of interest that have already been observed, and $UJ = J \cap \{C_{o+1}, \ldots, C_k\}$ the corresponding index-set for the categories in the event of interest that have not yet been observed. Let r be the number of elements of OJ and

l the number of elements of UJ, so $0 \leq r \leq o$ and $0 \leq l \leq k-o$. The NPI-based lower and upper probabilities for the event of interest, based on the n observations, are:

$$\begin{array}{l} \underline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \frac{1}{n}\left(\sum_{i=1}^{n} \sum_{j \in J} y_{ij} + \max(r+l-k, -r)\right), \\ \overline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \frac{1}{n}\left(\sum_{i=1}^{n} \sum_{j \in J} y_{ij} + \min(r+l, o-r)\right). \end{array} \tag{38}$$

It can be noticed that in case $l = 0$, i.e., all the categories have been observed, then $r = |J|$ and $\sum_{i=1}^{n} \sum_{j \in J} y_{ij} \geq |J|$ and, thus, one has:

$$\begin{array}{l} \underline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \frac{1}{n}\left(\sum_{i=1}^{n} \sum_{j \in J} y_{ij} + \max(|J|-k, -|J|)\right), \\ \overline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \frac{1}{n}\left(\sum_{i=1}^{n} \sum_{j \in J} y_{ij} + \min(|J|, k-|J|)\right), \end{array} \tag{39}$$

which coincides exactly with (22) in the case $c = 1$. It is thus extremely interesting that in this case the same inferences of the NPI can be obtained by our set of conjugate parametric priors which includes, at the limit, improper densities. In the case $l \neq 0$, the two models still coincide provided that $\sum_{i=1}^{n} \sum_{j \in J} y_{ij} = r$, i.e., each category in OJ has been observed at most once. In the other cases, the NPI is in general more precise. This can be easily proven by comparing (38) and (39) and by noticing that $-r \geq -|J|$, $r+l-k = |J|-k$ and $r+l = |J|$, $o-r \leq k-|J| = o-r+k-o-l$. For instance, consider the case $k = 4$, $J = \{1,2\}$, $n = 2$ and only the category 1 has been observed. Then, since $l = 1$, the lower and upper probability of the event of interest obtained with the NPI are:

$$\begin{array}{l} \underline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \frac{1}{2}(2 + \max(1+1-4, -1)) = \frac{1}{2}, \\ \overline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \frac{1}{2}(2 + \min(1+1, 0)) = 1, \end{array}$$

while the inferences obtained with the model in (22) are:

$$\begin{array}{l} \underline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \max\left(0, \frac{1}{2}(2 + \max(2-4, -2))\right) = 0, \\ \overline{E}[Y_{n+1} \in \cup_{j \in J} C_j | n, \hat{y}_n] = \min\left(1, \frac{1}{2}(2 + \min(2, 2))\right) = 1, \end{array}$$

and it remains vacuous up to 3 observations. In [15] the authors compare the NPI with the IDM highlighting the following differences between them. First, the IDM lower probability for the second observation to be equal to the first, is $1/(1+s)$. Thus, small values of s, e.g., $s = 1$ or $s = 2$, lead to surprisingly high values for this lower probability. In the NPI, this lower probability is 0 and the same holds for our model (22).

Second, the IDM predictive lower and upper probabilities depend only on the observed frequency of that category and the total number of observations (RIP). This is not the case for NPI-based lower and upper probabilities and also for our model (21). However, in Section 4.2, we have pointed out that one-step ahead RIP can be satisfied by our model by adding an additional constraint w.r.t. the NPI, see (23).

Third, the IDM upper probabilities for events that the next observation is in an as yet unseen category do not depend on the number of categories seen so far. This is not the case for the NPI, while this is also the case for our model. The difference is the term l in the computation of the lower and upper probabilities, which makes NPI inferences depend on the unobserved categories. This behavior could be included in our model only by letting the set of priors depend on the data.

The main advantage of the NPI w.r.t. thet IDM and our models is that it is a nonparametric model. In fact, the NPI can also be used if one does not know the total number of possible categories, and wishes to distinguish in the event of interest between fully defined categories that have not yet been observed, and any new category occurring at the next observation [15]. In the IDM and in our model, the lower and upper probabilities of these two cases are the same.

7.4 *Bounded Derivative Model*

Another comparison is with the so-called *bounded derivative model* (BDM) [14]. The BDM is a prior near-ignorance model for a scalar variable $w \in \mathbb{R}$ in which the set of priors $\mathscr{M}$ includes all continuous proper probability density functions for which the derivative of the log-density is bounded by a positive constant:

$$\left| \frac{d}{dw} \ln p(w) \right| \leq c. \tag{40}$$

Observe that the BDM satisfies (A1)–(A4) as stated in Theorem 4.2. It is clear that the set of priors in Theorem 4.2 also satisfies (40), since

$$\left| \frac{d}{dw} \ln \exp(\ell w) \right| = |\ell| \leq c.$$

However, the BDM set of priors is larger, since it does not restrict the set of priors to be $\{\exp(\ell w),\ |\ell| \leq c\}$. We have imposed this set of priors in order to exploit conjugacy. This is not the case of the BDM in which the extreme priors that obtain the lower and upper expectation of a RVBF cannot always be expressed as simple limits of conjugate priors. Walley in fact proves in [14] that in the BDM the extreme priors which obtain the upper and lower expectations of a generic RVBF g are piecewise exponential, i.e., the real line can be divided into intervals on which $p(w) = exp(cw)$ or $p(w) = exp(-cw)$. These piecewise priors do not belong to $\{\exp(\ell w),\ |\ell| \leq c\}$ in general. It is however worth to observe that in the case $g = b'$ or g is an indicator over a subset of $\mathscr{W} = \mathbb{R}$ (in general if g is monotone), both the BDM and our model produce the same lower and upper expectations. This holds since the lower and upper expectations are obtained in correspondence of either the kernel $p(w) = exp(cw)$ or $p(w) = exp(-cw)$. For a generic RVBF g, we expect that the inferences obtained with the BDM are more conservative, i.e., it has a larger posterior imprecision. The price to be paid for this gain of robustness and generality is the increase of the computational cost for the inferences.

8. Some Practical Examples

In this section we give some insight of the possible applications of the proposed models in statistical inference.

8.1 *One-Sample Location Test*

Consider i.i.d. observations $y_1, \ldots, y_n$ from a real variable Y with unknown mean x and unknown precision λ ($\lambda = 1/\sigma^2$). Our goal is to test the hypotheses:

$$H_0 :\ x \leq 0, \quad H_1 :\ x > 0.$$

A way to tackle this problem is by means of a frequentist t-test. The t-test considers the statistic:

$$t = \frac{\hat{y}_{p1} - \Delta_0}{\sqrt{\frac{\hat{y}_{p2} - \hat{y}_{p1}^2}{n-1}}},$$

where Δ_0 is the value of the mean under the null hypothesis, $\hat{y}_{p1} = \frac{1}{n}\sum_i y_i$, $\hat{y}_{p2} = \frac{1}{n}\sum_i y_i^2$ and, thus, $\hat{y}_{p2} - \hat{y}_{p1}^2 = \frac{1}{n}\sum_i (y_i - \hat{y}_{p1})^2$. Then it computes the p-value of t under the null hypothesis,

i.e., t is Student-t distributed with mean $\Delta_0 = 0$ and $n-1$ degrees of freedom. Hypothesis H_0 is rejected when the p-value is less than the given threshold $\alpha = 0.05$. An alternative approach is to perform a Bayesian test, using a Normal-Gamma model based on the improper prior $p(x,\lambda) = \lambda^{\frac{-\nu+1}{2}}$, with parameter $\nu > 0$. In this case we accept hypothesis H_1 if the posterior probability $P(x > 0|y_1,\ldots,y_n) > 1-\alpha$, with

$$P(x > 0|y_1,\ldots,y_n) = 1 - tcdf\left(\frac{0-\hat{y}_{p1}}{\sqrt{\frac{\hat{y}_{p2}-\hat{y}_{p1}^2}{n-\nu+1}}}, n-\nu+1\right),$$

where *tcdf* is the cumulative distribution function of Student distribution with $n-\nu+1$ degrees of freedom computed in 0. Note that for $\nu = 3$ we obtain the left Haar invariant prior, while for $\nu = 2$ we obtain the right Haar invariant prior [2, Sec. 3.3.3]. In the experiments we have selected $\nu = 2$ because the Bayesian test and the frequentist test coincide for this value. We can also devise a Bayesian test based on a hierarchical Normal model [23, Sec. 11.7]:

$$\textit{likelihood: } N(y;x,\sigma^2), \quad \textit{prior: } N(x;\mu,\tau^2), \quad \textit{hyper-prior: } p(\mu,log(\sigma),log(\tau)) \propto \tau.$$

Note that the hyperprior has been defined on $log(\sigma), log(\tau)$ to avoid that the posterior distribution is improper. However several other choices are possible. The posterior probability $P(x > 0|y_1,\ldots,y_n)$ (necessary for the hypothesis test) is computed numerically by Gibbs sampling.

Finally we can perform a test based on the *imprecise exponential family model* (IEM) presented in Section 6.2. In this case we compute the lower and upper probability:[1]

$$\begin{aligned}
\underline{P}(x > 0|y_1,\ldots,y_n) &= \min_{\ell_{\cdot 1}\in[-c_{\cdot 1},c_{\cdot 1}],\ell_{\cdot 2}\in[-c_{\cdot 2},c_{\cdot 2}]} 1 - tcdf\left(\frac{0-\hat{y}_{p1}+\frac{\ell_{\cdot 1}}{n}}{\sqrt{\frac{\hat{y}_{p2}+\frac{\ell_{\cdot 2}}{n}-(\hat{y}_{p1}+\frac{\ell_{\cdot 1}}{n})^2}{n-\nu+1}}}, n-\nu+1\right), \\
\overline{P}(x > 0|y_1,\ldots,y_n) &= \max_{\ell_{\cdot 1}\in[-c_{\cdot 1},c_{\cdot 1}],\ell_{\cdot 2}\in[-c_{\cdot 2},c_{\cdot 2}]} 1 - tcdf\left(\frac{0-\hat{y}_{p1}+\frac{\ell_{\cdot 1}}{n}}{\sqrt{\frac{\hat{y}_{p2}+\frac{\ell_{\cdot 2}}{n}-(\hat{y}_{p1}+\frac{\ell_{\cdot 1}}{n})^2}{n-\nu+1}}}, n-\nu+1\right),
\end{aligned} \tag{41}$$

and we perform the hypothesis test as follows:

- return H_1 if $\underline{P}(x > 0|y_1,\ldots,y_n) > 1-\alpha$;
- return H_0 if $\overline{P}(x > 0|y_1,\ldots,y_n) < 1-\alpha$;
- otherwise issue an indeterminate answer, i.e., $\{H_0, H_1\}$.

In the latter case we cannot decide: the result of the hypothesis test depends on the choice of the prior. To compare the frequentist/Bayesian and IEM-based test we have considered a Monte Carlo experiment in which n observations Y are generated based on $Y \sim N(\Delta, 1)$, with Δ ranging from 0 to 1.5; for $\Delta = 0$, we are under the null hypothesis of the frequentist/Bayesian tests. In particular, for each value of Δ we have performed 20000 Monte Carlo runs by generating in each run $n = 15$ and $n = 30$ observations for Y and computed the percentage of cases in which the alternative hypothesis is returned, using $\alpha = 0.05$ for both the tests. This means that for $\Delta = 0$

[1] In the formulation we need to impose the additional constraint that the argument of the square root is positive. This implies that the parameters $\ell_{\cdot 1}$ and $\ell_{\cdot 2}$ cannot vary independently in $[-c_{\cdot 1}, c_{\cdot 1}]$ and, respectively, $[-c_{\cdot 2}, c_{\cdot 2}]$. However, for suitably large n and non-degenerate distributions, the argument of the square root is usually positive for any $\ell_{\cdot 1}$ and $\ell_{\cdot 2}$ in the intervals.

we are computing the Type-I error of the test and for $\Delta > 0$ its power. The IEM test has been implemented by choosing $\nu = 2$ (equal to the frequentist/Bayesian tests). About the choice of the c parameters, we have selected $c_{\cdot 1}$ equal to 0.75 so that the posterior imprecision for the estimate of x is reduced to 1.5 after one observation, while $c_{\cdot 2} \approx 0.28$, so that $c_{\cdot 2} = (c_{\cdot 1})^2/n$ for $n = 2$. The expression $(c_{\cdot 1})^2/2$ gives the sample second non-central moment of the pseudo-observations $\pm c_{\cdot 1}$; remember in fact that $\ell_{\cdot 2}$ can be interpreted as the square of a pseudo-observation and we need two observations to estimate both x and λ. This is the meaning of $c_{\cdot 2} = (c_{\cdot 1})^2/2$. Note that with this choice our model has only one free parameter $c_{\cdot 1}$. We will empirically show that this choice is a good compromise between robustness and conservativeness (indeterminacy).

The average power[1] is shown in Figure 1 as a function of Δ for the two cases. In particular,

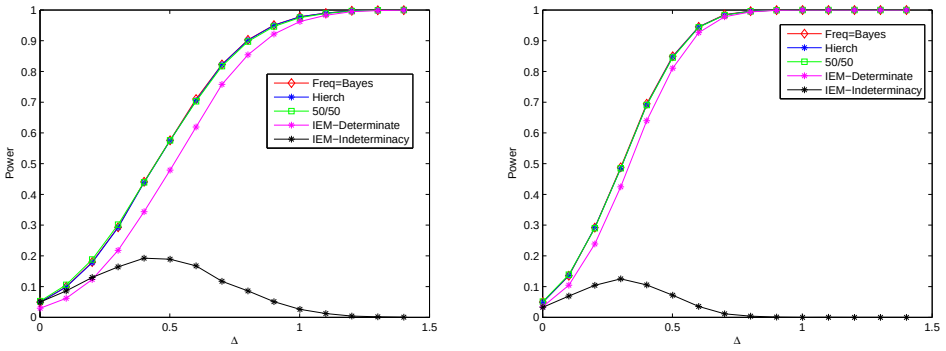


Figure 1.: Power as a function of Δ for the case $n = 15$ (left) and $n = 30$ (right).

Figure 1 reports (i) the power of the frequentist/Bayesian tests; (ii) the power of the hierarchical Bayesian test; (iii) the power of the IEM test when it is determinate; (iv) the indeterminacy of the IEM test, i.e., the number of times it returns an indeterminate response divided by the total number of Monte Carlo runs; (v) the power of a new test (called "50/50") which returns the same response as IEM when IEM is determinate, and issues a random answer (with 50/50 chance) when IEM is indeterminate. We have introduced this new artificial test to facilitate the comparison of the frequentist/Bayesian, hierarchical Bayesian and IEM tests. From Figure 1 (left and right), it is evident that the performance of the frequentist/Bayesian, hierarchical Bayesian and 50/50 tests practically coincide. Furthermore, since in all cases in which IEM is determinate, the frequentist/Bayesian and hierarchical Bayesian return the same response as IEM (i.e., in the determinate cases the accuracies of the frequentist/Bayesian and hierarchical Bayesian tests are also equal to the (magenta) star curve in Figure 1), the difference between the three tests is only in the runs where IEM is indeterminate. In these runs, the frequentist/Bayesian and hierarchical Bayesian tests have essentially the same accuracy as the 50/50 test. Therefore, the IEM is able to isolate some instances in which the frequentist/Bayesian and hierarchical Bayesian tests are virtually guessing at random, in the sense that they have the same accuracy of the 50/50 test. Assume for instance that we are trying to evaluate the effects of a medical treatments ("X is greater than zero") and that, given the available data, the IEM is indeterminate. In such a situation the frequentist/Bayesian and hierarchical tests always issue a determinate response (I can tell if "X is greater than zero"), but it turns out that their response is completely random (like if we were tossing a coin). On the other side, the IEM acknowledges the impossibility of making a decision (I do not know whether "X is greater than zero") and thus, although the frequentist/Bayesian and hierarchical tests and the IEM (more precisely the 50/50 test derived by IEM) have the same accuracy, the IEM provides more information.

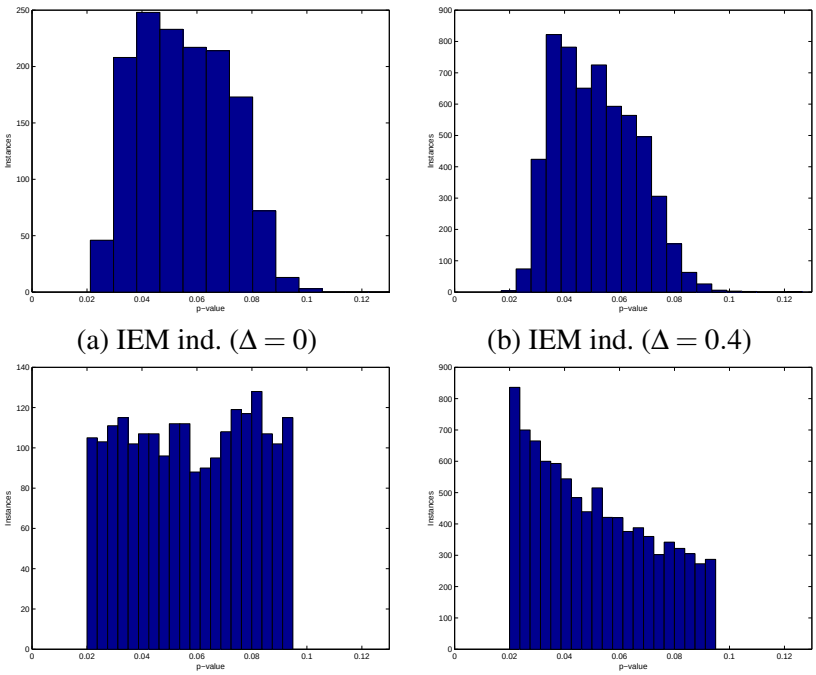
[1] For $\Delta = 0$ the plot actually reports the Type I error.

Note that in Figure 1 (left) for $\Delta \approx 0.4$, the percentage of runs in which IEM is indeterminate is about 18% ; this means that the frequentist/Bayesian and hierarchical tests are issuing a random answer in 18% of the cases, which is a large percentage. Looking at the curve of the IEM indeterminacy, it can be observed that it reaches its maximum value at $\Delta = 0.4$ and then goes to zero. For $|\Delta| > 1.2$ the power of the frequentist/Bayesian, hierarchical Bayesian and IEM-determinate is practically one, which means that when the hypothesis test is easy (large Δ), there are not indeterminate instances and both tests have power one. The conclusions for Figure 1 (right) are similar.

For the frequentist/Bayesian test, we have also evaluated the distribution of the p-values in the instances where the IEM is indeterminate. The histogram of the p-values is shown in Figures 2 (a) and (b) for $\Delta = 0$ and, respectively, $\Delta = 0.4$. It can be observed that the IEM test is indeterminate when the p-value is close to the threshold 0.05, but also in instances where it is quite far (e.g., $p = 0.02$ or $p = 0.1$). Moreover, it can be noticed that the distribution of the p-values in the IEM indeterminate instances is roughly symmetric around 0.05 (more precisely, the area of the distribution of the p-values to the left of 0.05 is (approximatively) equal to the area to the right of 0.05), no matter the value of Δ. This explains why the "50/50" test and the frequentist/Bayesian test have the same power. This behavior cannot be reproduced by simply adding a "no decision zone" to the frequentist/Bayesian test, i.e., by suspending the decision when $a < p < b$. For instance, Figures 2 (c) and (d) show the distribution of the p-values of the frequentist/Bayesian test in the interval with boundary $a = 0.02$ and $b = 0.095$.[1] By comparing (c) and (d), it can be observed that the distribution of the p-values in the "no decision zone" $0.02 < p < 0.095$ changes with Δ. It is uniform in the case $\Delta = 0$ (it is in fact well known that the distribution of the p-values is uniform under the null hypothesis), but is strongly skewed in the case $\Delta = 0.4$. This is not true for the distribution of the p-values in the IEM indeterminate instances, which instead does not change much with Δ, see Figures 2 (a) and (b). Therefore, since the distribution of the p-values in the "no decision zone" is not symmetric around 0.05 (the case $\Delta = 0.4$ in Figure 2 (d)), a "50/50" test designed over the "no decision zone" $0.02 < p < 0.095$ would not have the same behavior of IEM "50/50" test, e.g., it would not have the same power as the frequentist/Bayesian test. In other words, the "no decision zone" $0.02 < p < 0.095$ is not able to isolate the instances that are critical for the frequentist/Bayesian test, i.e., it is both classifying as indeterminate the instances that are easy and those that are difficult. This is evident comparing Figures 2 (b) and (d) for $p = 0.02$: the IEM is classifying as indeterminate only few instances with p-value close to $p = 0.02$, while the "no decision zone" $0.02 < p < 0.095$ is classifying as indeterminate all the instances with p-value close to $p = 0.02$. The only way a frequentist test with a "no decision zone" can mimic the behavior of the IEM test is with $a(data) < p < b(data)$. This means that the boundary must depend on the observations. Mathematically, this is evident from (41). The indeterminacy of the IEM test cannot be reproduced by $a < P(x > 0|y_1, \ldots, y_n) < b$, since the values $\ell_{.1}$ and $\ell_{.2}$ that give the maximum and the minimum depend on the data (through the nonlinear function $tcdf$).

Finally, in case $n = 15$ and $\Delta = 0$, we have computed the error for the IEM-based test as a function of $c_{.1}$ (with $c_{.2} = (c_{.1})^2/2$). The error of the frequentist test is in this case equal to 0.05 (we are under the null hypothesis). From Figure 3, it can be observed that the error of the IEM test when determinate decreases with $c_{.1}$, because of the increase of the indeterminacy (note that for $c_{.1} = 0$, the error of IEM when determinate and the error of the 50/50 test coincide, since there are not indeterminate instances in this case). Looking at the plot relative to the 50/50 test, it can be noticed that it is almost constant for $c_{.1} < 0.75$ and then it increases. This (together with the other results of this section) may be seen as an empirical confirmation that the choice of $c_{.1} = 0.75$ is appropriate. It guarantees the maximum robustness at the minimum of conservativeness (indeterminacy). We have also evaluated frequentist properties of this choice.

[1] We have chosen this interval in analogy with that of the IEM test. The distribution of the p-values for a different boundary of the "no decision zone" can easily be deduced from Figures 2 (c) and (d).



(a) IEM ind. ($\Delta = 0$) (b) IEM ind. ($\Delta = 0.4$)

(c) Freq. Ind ($\Delta = 0$) (d) Freq. Ind ($\Delta = 0.4$)

Figure 2.: Histogram of the p-values ($n = 15$).

	Calibration
$n = 15$	0.9686
$n = 30$	0.9593
$n = 150$	0.9552

Table 1.: Calibration of the IEM robust (symmetric) 95% high posterior density interval.

In particular, we have evaluated the calibration of the robust (symmetric) high posterior density interval (HDI) of the IEM, i.e., the minimum length interval that has lower probability equal to 0.95. The results are shown in Table 1 as a function of n. It can be observed that the robust HDI includes the true mean with a probability always greater than 0.95, which means that it is calibrated. The probability does not excessively exceed the prescribed value 0.95, which is another empirical confirmation that the choice $c_{.1} = 0.75$ is a good compromise between robustness and conservativeness.

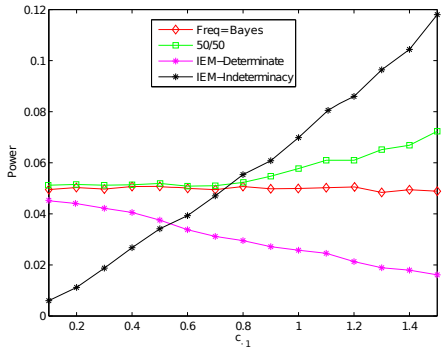


Figure 3.: Power as a function of $c_{.1}$ for the case $n = 15$.

It should be stressed again that the test "50/50" has been introduced only for the sake of comparison. We are not suggesting that when IEM is indeterminate we should toss a coin to take the decision. On the contrary we claim that the indeterminacy of IEM is an additional useful information that our approach gives to the analyst. In these cases she/he knows that (i) her/his posterior decisions would depend on the choice of the prior; (ii) the hypothesis test is difficult as shown by the comparison with the frequentist test. Based on this additional information, the analyst can for example decide to collect additional measurements to eliminate the indeterminacy (in fact we have seen that when the number of observations goes to infinity the indeterminacy goes to zero).

8.2 *Election Polls*

A total of n adults are polled to indicate their preference for two candidates A and B. Let $\hat{y}_{n1}$ denote the proportion of the population that supports A, $\hat{y}_{n2}$ denote the proportion of the population that supports B and $1-\hat{y}_{n1}-\hat{y}_{n2}$ denote the proportion of the population that is either undecided or vote for someone else. The counts $n\hat{y}_{n1}$, $n\hat{y}_{n2}$ and $n(1-\hat{y}_{n1}-\hat{y}_{n2})$ are assumed to have a multinomial distribution with sample size n and respectively parameters $\theta_{\cdot 1}$, $\theta_{\cdot 2}$ and $\theta_{\cdot 3}$. Thus, the likelihood model is:

$$p(n,\hat{y}_n|\theta) = \theta_{\cdot 1}^{n\hat{y}_{n1}}\theta_{\cdot 2}^{n\hat{y}_{n2}}\theta_{\cdot 3}^{n\hat{y}_{n3}},$$

where $\theta_{\cdot 1}+\theta_{\cdot 2}+\theta_{\cdot 2}=1$ are the unknown non-negative chances to be estimated. The focus is to compare the proportions of voters for A and B by considering the difference $\theta_{\cdot 1}-\theta_{\cdot 2}$. A Dirichlet conjugate prior can be assumed on $\theta_{\cdot 1}$, $\theta_{\cdot 2}$ and $\theta_{\cdot 3}$:

$$p(\theta) \propto \theta_{\cdot 1}^{\alpha_{\cdot 1}-1}\theta_{\cdot 2}^{\alpha_{\cdot 2}-1}\theta_{\cdot 3}^{\alpha_{\cdot 3}-1},$$

where in the case of lack of prior information the prior parameters are commonly selected as follows: Haldane's prior $\alpha_{\cdot 1}=\alpha_{\cdot 2}=\alpha_{\cdot 3}=0$; Jeffreys' prior $\alpha_{\cdot 1}=\alpha_{\cdot 2}=\alpha_{\cdot 3}=\frac{1}{2}$; uniform prior $\alpha_{\cdot 1}=\alpha_{\cdot 2}=\alpha_{\cdot 3}=1$. The expected value of $E[\theta_{\cdot 1}-\theta_{\cdot 2}]$ is equal to 0 and the prior probability $P(\theta_{\cdot 1}>\theta_{\cdot 2})=0.5$ for both Jeffreys and uniform priors (for Haldane's prior they are not defined). These priors express indifference between candidate A and B, but not prior ignorance. To see that, consider $P[\theta_{\cdot 1}+0.5\theta_{\cdot 3}>\theta_{\cdot 2}+0.4\theta_{\cdot 3}]$, this is the probability that the proportion of votes of A exceeds the votes for B assuming a "swing" scenario in which 50% of the undecideds vote for A and 40% of the undecideds for B. This probability is equal to 0.76 in the case of the uniform prior and 0.66 in the case of Jeffreys' prior. It depends on the choice of the prior and this shows that the uniform and Jeffrey's priors are not really uninformative for this kind of poll.

Combining likelihood and prior, the resulting posterior is

$$p(x|n,\hat{y}_n) \propto \theta_{\cdot 1}^{n\hat{y}_{n1}+\alpha_{\cdot 1}-1}\theta_{\cdot 2}^{n\hat{y}_{n2}+\alpha_{\cdot 2}-1}\theta_{\cdot 3}^{n\hat{y}_{n3}+\alpha_{\cdot 3}-1},$$

which is always proper in the case of Jeffreys' and uniform prior and in the case of Haldane's prior provided that $\hat{y}_{n1},\hat{y}_{n2},\hat{y}_{n3}>0$. The posterior expected value of $\theta_{\cdot 1}-\theta_{\cdot 2}$ is:

$$E[\theta_{\cdot 1}-\theta_{\cdot 2}|n,\hat{y}_n)] = \frac{n\hat{y}_{n1}+\alpha_{\cdot 1}}{n+\alpha_{\cdot 1}+\alpha_{\cdot 2}+\alpha_{\cdot 3}} - \frac{n\hat{y}_{n2}+\alpha_{\cdot 2}}{n+\alpha_{\cdot 1}+\alpha_{\cdot 2}+\alpha_{\cdot 3}},$$

while the posterior probability of the event $\theta_{\cdot 1} - \theta_{\cdot 2} > 0$ is

$$P[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n] = \frac{\int\limits_{\{\theta_{\cdot 1} > \theta_{\cdot 2}\}} \theta_{\cdot 1}^{n\hat{y}_{n1}+\alpha_{\cdot 1}-1} \theta_{\cdot 2}^{n\hat{y}_{n2}+\alpha_{\cdot 2}-1} \theta_{\cdot 3}^{n\hat{y}_{n3}+\alpha_{\cdot 3}-1} \, d\theta}{\int \theta_{\cdot 1}^{n\hat{y}_{n1}+\alpha_{\cdot 1}-1} \theta_{\cdot 2}^{n\hat{y}_{n2}+\alpha_{\cdot 2}-1} \theta_{\cdot 3}^{n\hat{y}_{n3}+\alpha_{\cdot 3}-1} \, d\theta},$$

which can be computed numerically by sampling from the Dirichlet distribution.

Consider now the model of near-ignorance $\mathscr{M}$ defined in Theorem 4.6. From these results it follows that a priori $\underline{E}[\theta_{\cdot 1} - \theta_{\cdot 2}] = -1$, $\overline{E}[\theta_{\cdot 1} - \theta_{\cdot 2}] = 1$. From Corollary 5.1, one has that $\underline{P}(\theta_{\cdot 1} > \theta_{\cdot 2}) = 0, \overline{P}(\theta_{\cdot 1} > \theta_{\cdot 2}) = 1$ and that $\underline{P}[\theta_{\cdot 1} + 0.5\theta_{\cdot 3} > \theta_{\cdot 2} + 0.4\theta_{\cdot 3}] = 0$, and $\overline{P}[\theta_{\cdot 1} + 0.5\theta_{\cdot 3} > \theta_{\cdot 2} + 0.4\theta_{\cdot 3}] = 1$.[1] This is a more correct expression of the lack of prior information on the election result. The set of posteriors produced by $\mathscr{M}$ is:

$$\mathscr{M}_p = \left\{ p(x|n, \hat{y}_n) \propto \theta_{\cdot 1}^{n\hat{y}_{n1}+\ell_{\cdot 1}-1} \theta_{\cdot 2}^{n\hat{y}_{n2}+\ell_{\cdot 2}-1} \theta_{\cdot 3}^{n\hat{y}_{n3}+\ell_{\cdot 3}-1} \right\},$$

with $\ell \in \mathbb{L}$. Assuming that $\mathbb{L}$ is the set defined in (23), two of the extreme densities belonging to this set are:

$$\theta_{\cdot 1}^{n\hat{y}_{n1}-c-1} \theta_{\cdot 2}^{n\hat{y}_{n2}+c-1} \theta_{\cdot 3}^{n\hat{y}_{n3}-1}, \tag{42}$$

$$\theta_{\cdot 1}^{n\hat{y}_{n1}+c-1} \theta_{\cdot 2}^{n\hat{y}_{n2}-c-1} \theta_{\cdot 3}^{n\hat{y}_{n3}-1}. \tag{43}$$

These extreme densities give respectively the lower and upper posterior expected values of $\theta_{\cdot 1} - \theta_{\cdot 2}$:

$$\underline{E}[\theta_{\cdot 1} - \theta_{\cdot 2} | n, \hat{y}_n] = \max\left(-1, \frac{n\hat{y}_{n1}}{n} - \frac{n\hat{y}_{n2}}{n} - \frac{2c}{n}\right), \tag{44}$$

$$\overline{E}[\theta_{\cdot 1} - \theta_{\cdot 2} | n, \hat{y}_n] = \min\left(1, \frac{n\hat{y}_{n1}}{n} - \frac{n\hat{y}_{n2}}{n} + \frac{2c}{n}\right), \tag{45}$$

where it has been assumed that $n\hat{y}_{n1}, n\hat{y}_{n2} > 0$. The lower expectation considers the case in which c votes move from candidate A to B. The upper expectation considers the case in which c votes move from candidate B to A. Clearly this exchange of votes gives also the the lower and upper posterior probabilities of the event $\theta_{\cdot 1} - \theta_{\cdot 2} > 0$:

$$\underline{P}[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n] = \frac{\int\limits_{\{\theta_{\cdot 1} > \theta_{\cdot 2}\}} \theta_{\cdot 1}^{n\hat{y}_{n1}-c-1} \theta_{\cdot 2}^{n\hat{y}_{n2}+c-1} \theta_{\cdot 3}^{n\hat{y}_{n3}-1} \, d\theta}{\int \theta_{\cdot 1}^{n\hat{y}_{n1}-c-1} \theta_{\cdot 2}^{n\hat{y}_{n2}+c-1} \theta_{\cdot 3}^{n\hat{y}_{n3}-1} \, d\theta},$$

$$\overline{P}[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n] = \frac{\int\limits_{\{\theta_{\cdot 1} > \theta_{\cdot 2}\}} \theta_{\cdot 1}^{n\hat{y}_{n1}+c-1} \theta_{\cdot 2}^{n\hat{y}_{n2}-c-1} \theta_{\cdot 3}^{n\hat{y}_{n3}-1} \, d\theta}{\int \theta_{\cdot 1}^{n\hat{y}_{n1}+c-1} \theta_{\cdot 2}^{n\hat{y}_{n2}-c-1} \theta_{\cdot 3}^{n\hat{y}_{n3}-1} \, d\theta}.$$

Above it has been assumed that the resulting posteriors are both proper. Similar calculations can be used to determine $\underline{P}[\theta_{\cdot 1} + 0.5\theta_{\cdot 3} > \theta_{\cdot 2} + 0.4\theta_{\cdot 3} | n, \hat{y}_n]$ and $\overline{P}[\theta_{\cdot 1} + 0.5\theta_{\cdot 3} > \theta_{\cdot 2} + 0.4\theta_{\cdot 3} | n, \hat{y}_n]$.

[1] These lower and upper probabilities are obtained by densities in $\mathscr{M}$ approaching the extreme priors in (42)–(43).

Compare (44)–(45) with the posterior means obtained by the IDM:

$$\underline{E}[\theta_{\cdot 1} - \theta_{\cdot 2} | n, \hat{y}_n] = \frac{n\hat{y}_{n1}}{n+s} - \frac{n\hat{y}_{n2}+s}{n+s}, \tag{46}$$

$$\overline{E}[\theta_{\cdot 1} - \theta_{\cdot 2} | n, \hat{y}_n] = \frac{n\hat{y}_{n1}+s}{n+s} - \frac{n\hat{y}_{n2}}{n+s}. \tag{47}$$

In this case the lower and upper expectations consider the case in which s more adults are polled and assume that these adults vote either for candidate B (lower case) or for candidate A (upper case). Although by choosing an appropriate value of s we can suitably enlarge $\overline{E}[\theta_{\cdot 1} - \theta_{\cdot 2} | n, \hat{y}_n] - \underline{E}[\theta_{\cdot 1} - \theta_{\cdot 2} | n, \hat{y}_n]$, the increase of s decreases the variance of the extreme densities and, thus, makes the inferences sharper.

As numerical example we consider the problem of estimating the number of electoral votes in the 2004 USA Presidential Election using the polling data from `realclearpolitics.com` (last-day poll). We compute the winner prediction of a Bayesian estimator that: (i) using Haldane's prior computes the win probabilities $P[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n]$ for each state; (ii) we use these probabilities to compute the (sampling) distribution of the total electoral votes for the two candidates.

Then we compute the winner prediction for the two near-ignorance prior models: IDM ($s = 2$) and the set of priors $\mathscr{M}$ with $c = 1$ hereafter denoted as IEM (imprecise exponential model), using the same procedure as in the Bayesian estimator but considering the two cases $\underline{P}[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n]$ and $\overline{P}[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n]$, which give the lower and, respectively, upper probability that Bush wins in each state.

A number of 200000 simulated elections are generated. In each simulation, for each State we have first computed the posterior probability of $\theta_{\cdot 1}$ and $\theta_{\cdot 2}$ (using the Bayesian prior or the lower and upper priors). For each State, we have used this posterior to sample $\theta_{\cdot 1}$ and $\theta_{\cdot 2}$ and assigned the electoral vote of the State to Bush whenever $\theta_{\cdot 1} > \theta_{\cdot 2}$ (to Kerry whenever $\theta_{\cdot 2} < \theta_{\cdot 1}$). Finally, we have computed Bush's total electoral votes. We have repeated this process for each of the 200000 runs. Figure 4 reports the histograms of Bush's total electoral votes that we have obtained on the 200000 runs for the three estimator together with the break-even equal to 269 electoral votes. It is evident from the histogram of the Bayesian estimator that Kerry is slightly favorite (for Bush the posterior mean of electoral votes is 267.1 and the posterior median 266) but also that there is a high uncertainty. Because of this uncertainty the contribution of the prior on the final result is crucial. In fact, notice that if we just add two (pseudo-)samples ($s = 2$) in favor of Bush in each state, then Bush becomes slightly favorite (posterior upper mean equal to 271.2 electoral votes and posterior upper median equal to 270 electoral votes for Bush), see IDM light red histogram. This corresponds to the case $\overline{P}[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n]$ in the IDM and the best scenario for Bush. The lower probability $\underline{P}[\theta_{\cdot 1} > \theta_{\cdot 2} | n, \hat{y}_n]$ (IDM light blue histogram) gives instead the worst scenario for Bush. Similar results hold for the IEM. In this case, we assume that in each state one elector ($c = 1$) in the sample changes mind and instead of voting for Kerry he/she votes for Bush. This is enough for Bush to become slightly favorite. Notice that the IDM and IEM give the same result in this case. In fact, being the sample size large (around 600 in each state), $n + s \approx n$ for $s = 2$ and, thus, the inferences of the IDM and IEM almost coincide.

9. Conclusions

In this paper, we have proposed a model of prior ignorance about a multivariate variable based on a set of distributions $\mathscr{M}$. In particular, we have discussed four properties that a prior model should satisfy to model lack of prior information: invariance, near-ignorance, learning and convergence. Near-ignorance and invariance ensure that our prior model behaves as a vacuous model with respect to some statistical inferences (e.g., mean, credible intervals, etc.) and some transfor-

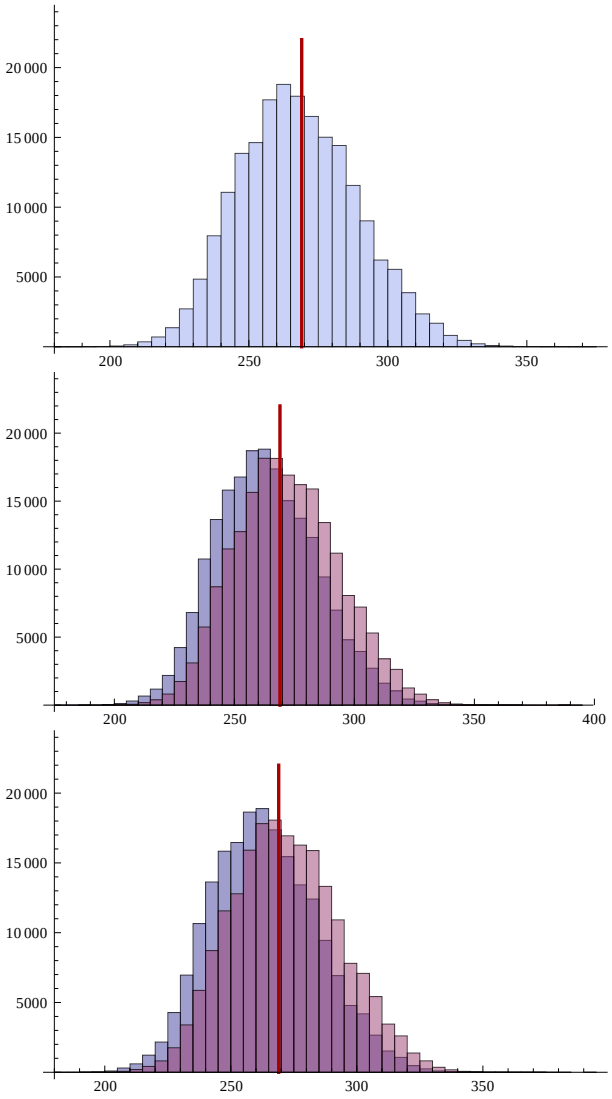


Figure 4.: Histogram of Bush's total electoral votes for the Bayesian estimator (top), IDM (center) and IEM (bottom). The vertical line denotes the break-even 269 electoral votes.

mation of the parameter space. Learning and convergence ensure that our prior model can learn from data and, in particular, that the influence of $\mathscr{M}$ on the posterior inferences vanishes with increasing number of observations. Furthermore, for exponential families, we have shown that translation invariance, near-ignorance, learning and convergence can all be satisfied by a set of conjugate priors if the set of priors $\mathscr{M}$ includes finitely additive probabilities obtained as limits of truncated increasing/decreasing exponential functions.

In future work it would be useful to address the following issues: (1) the extension of this model to more cases for which $\mathscr{W} \neq \mathbb{R}^k$; (2) employing the model to represent prior ignorance in *generalized linear models*.

The extension to the case $\mathscr{W} \neq \mathbb{R}^k$ is important because, for many exponential families of interest $\mathscr{W} \neq \mathbb{R}^k$. Furthermore, with this extension we may employ our approach to model prior ignorance in generalized linear models. Generalized linear models comprise the traditional analyses such as t-tests, analysis of variance, linear regression and logistic regression. Consider for instance a normal linear regression 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 not 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.

Acknowledgements

This work was partly supported by the Swiss NSF grants nos. 200021_146606 / 1 and 200020_137680 / 1. The authors would like to thank the anonymous referees for comments and constructive criticism that helped us to improve the presentation of the paper.

Appendix A. Invariance and Conjugate Improper Priors

In Bayesian statistics, in case of lack of prior information, it is often common to select the prior in order to satisfy some invariance property, see for instance [3, 24, 25]. Consider for instance the group of transformations $\mathscr{F} = \{f_a : a \in \mathbb{R}\}$ with $f_a(w) = w + a$, i.e., a shift of the parameter. From Definition 2.2 with $\underline{E} = \overline{E} = E$, it results that a prior $p(w)$ with support $\mathscr{W}$ is $\mathscr{F}$-invariant if:

$$\int g(w+a)p(w)dw = \int g(w)p(w)dw \quad \forall\, a. \tag{A1}$$

If $g = I_{\{A\}}$, with $A \subseteq \mathbb{R}^k$ bounded and measurable, the above inequality means that the set A and the shifted set $A + a$ should have the same probability for any value of a. By a change of variables and considering the case $g = I_{\{A\}}$, (A1) can be rewritten as:

$$\int_{A-a} p(w-a)dw = \int_{A} p(w)dw \quad \forall\, a. \tag{A2}$$

For a bounded space $\mathscr{W}$, the only prior that satisfies (A2) is the uniform distribution. Conversely, for instance in case $\mathscr{W} = \mathbb{R}$, it is known that there is no countably additive probability measure that is translation invariant, even on the intervals. To see that, note that $\mathbb{R}$ is the countable union of the intervals $[n, n+1)$ with $n \in \mathbb{Z}$ and these intervals must have the same probability, because of (A1), which must be zero by countable additivity.

However, we can define translation-invariant lower and upper expectations as limits of uniform distributions on finite intervals [6, Sec. 2.9.7], [26, Sec. 3]. Let $\mathscr{K}$ be the linear space of Borel-measurable RVBFs on $\mathbb{R}$, and define:

$$\underline{E}[g] = \liminf_{r\to\infty} \frac{1}{2r}\int g(w)I_{[-r,r]}(w)dw, \quad \forall g \in \mathscr{K}, \tag{A3}$$

where $r > 0$ and $\underline{E}$ denotes the lower expectation of a RVBF g belonging to $\mathscr{K}$. This is a lower limit of uniform distributions on intervals centered at zero as their length $2r \to \infty$. By replacing infimum with supremum in the definition of $\underline{E}$, we can obtain the upper expectation $\overline{E}$. Observe that, since $\underline{E}[g] = -\overline{E}[-g]$, the upper expectation in completely determined by the

lower expectation and vice versa. In case $g = I_{\{A\}}$ for some subset A of $\mathscr{W}$, $\underline{E}[I_{\{A\}}]$ represents the lower probability of A and in this case the upper probability of A, i.e., $\overline{E}[I_{\{A\}}]$, is defined as $1 - \underline{E}[I_{\{A^c\}}]$, where A^c is the complementary set of A.

Lemma A.1. *Define the translations f_a by $f_a(w) = w + a$ for any $a \in \mathbb{R}$ and $g(f_a(w)) = g(w + a)$ for any $g \in \mathscr{K}$. The lower expectation model in (A3) satisfies $\underline{E}[g(f_a)] = \underline{E}[g]$ (translation invariance) for any $g \in \mathscr{K}$ and $a \in \mathbb{R}$.*

The proof is given in the first part of the proof of Lemma 4.1. Observe that $\underline{E}$ defined in (A3) satisfies the following properties: (i) for $g = I_{\mathbb{R}}$, $\underline{E}[I_{\mathbb{R}}] = 1$; (ii) for $A = [0, \infty)$, $\underline{E}[I_A] = \frac{1}{2}$; (iii) for $g = I_A$, where A is a bounded subset of $\mathbb{R}$, it holds that $\underline{E}[I_{\{A+a\}}] = \underline{E}[I_A] = 0$ for any $a \in \mathbb{R}$ and, furthermore, by denoting with A^c the complementary set of A it holds that $\underline{E}[I_{A^c}] = \underline{E}[I_{\{A^c+a\}}] = 1$. Hence, since $\overline{E}[I_A] = 1 - \underline{E}[I_{A^c}]$, it follows that $\overline{E}[I_{\{A+a\}}] = \overline{E}[I_A] = 0$. As discussed in the introduction, to each functional $\underline{E}$ (satisfying some regularity properties [6, Ch. 2]), it is possible to associate a closed and convex set $\mathscr{M}$ of probabilities that generates the lower expectation $\underline{E}[g]$ for any g. This is the set of probabilities that dominate $\underline{E}$ (equivalently, that are dominated by $\overline{E}$), i.e., the lower (upper) envelope of the expectation of g with respect to the closed and convex set $\mathscr{M}$ of probabilities is equal to $\underline{E}[g]$ ($\overline{E}[g]$) for any g. In the case of $\underline{E}$ defined in (A3), since $\overline{E}[I_A] = 0$ for any bounded subset A of $\mathbb{R}$, none of the probabilities dominated by $\overline{E}$ can be countably additive. In other words the set $\mathscr{M}$ in this case includes finitely additive probabilities.

In order to be able to interchange the operations of limit and integration even in case of unbounded functions g from $\mathscr{W}$ to $\mathbb{R}$, in the following we take $\mathscr{K}$ to be the linear space of Borel-measurable functions on $\mathbb{R}$ satisfying $\int |g(w)| \exp(n(\hat{y}_n w - b(w)) dw < \infty$ for any $n > 0$ and $\hat{y}_n \in Cl(\mathscr{Y}_0)$ ($Cl(\cdot)$ denotes the closure of the set). This is again a linear space.[1] Now consider the likelihood $p(n, \hat{y}_n | w) \propto \exp(n(\hat{y}_n w - b(w))$; the lower posterior expectation of g can be obtained from (A3):

$$\underline{E}[g|n, \hat{y}_n] = \liminf_{r \to \infty} \frac{\int g(w) \exp(n(\hat{y}_n w - b(w)) I_{[-r,r]}(w) dw}{\int \exp(n(\hat{y}_n w - b(w)) I_{[-r,r]}(w) dw}. \tag{A4}$$

Assuming $\mathscr{W} = \mathbb{R}$, since $\int |g(w)| \exp(n(\hat{y}_n w - b(w)) dw < \infty$ for any $n > 0$ and $\hat{y}_n \in Cl(\mathscr{Y}_0)$, applying Lebesgue's dominated convergence theorem, one has:

$$\underline{E}[g|n, \hat{y}_n] = \frac{\int g(w) \exp\left(n\left(\hat{y}_n w - b(w)\right)\right) dw}{\int \exp\left(n\left(\hat{y}_n w - b(w)\right)\right) dw}, \tag{A5}$$

which is equivalent to the posterior inference obtained from the limit kernel (improper uniform prior) $p(w) = 1 = \lim_{r \to \infty} I_{[-r,r]}(w)$. Observe that $\underline{E}[g|n, \hat{y}_n] = \overline{E}[g|n, \hat{y}_n]$. It is worth to notice that (A5) holds for any likelihood in the exponential families such that $\mathscr{W} = \mathbb{R}$. For instance, in the one-parameter Normal, Beta and Gamma case, $\mathscr{W} = \mathbb{R}$ and thus the limit kernel $p(w) = 1$ guarantees translation invariance. Note that $p(w) = 1$, transformed back to the original parameter space, becomes $p(x) = 1$ in the Normal case, $p(\theta) = \theta^{-1}(1 - \theta)^{-1}$ in the Beta case and $p(\lambda) = \lambda^{-1}$ in the Gamma case. These are the kernels that, when multiplied with the Normal, Binomial and Poisson likelihoods, give a Normal, Beta and, respectively, Gamma posterior density.

Note also that $p(w) = 1$ is not the only limit prior that guarantees translation invariance and preserves conjugacy, i.e., the posterior is the conjugate of the likelihood. Assume again $\mathscr{W} = \mathbb{R}$ and that g satisfies $\int |g(w)| \exp(n(\hat{y}_n w - b(w)) \exp(\ell w) dw < \infty$ for some $\ell > 0$ and any

[1] Notice in fact that if g belongs to $\mathscr{K}$ also λg does for any real λ. If g_1, g_2 belong to $\mathscr{K}$, also $g_1 + g_2$ belongs to $\mathscr{K}$, since $\int |g_1(w) + g_2(w)| \exp(n(\hat{y}_n w - b(w)) dw \leq \int |g_1(w)| \exp(n(\hat{y}_n w - b(w)) dw + \int |g_2(w)| \exp(n(\hat{y}_n w - b(w)) dw < \infty$.

$n > 0$, $\hat{y}_n \in Cl(\mathscr{Y}_0)$. Consider then the posterior limit:

$$\underline{E}[g|n,\hat{y}_n] = \liminf_{r\to\infty} \frac{\int g(w)\exp(n(\hat{y}_n w - b(w)))\exp(\ell w) I_{(-\infty,r]}(w)dw}{\int \exp(n(\hat{y}_n w - b(w)))\exp(\ell w) I_{(-\infty,r]}(w)dw}, \tag{A6}$$

as before, since $\int |g(w)|\exp(n(\hat{y}_n w - b(w)))\exp(\ell w)dw < \infty$, by Lebesgue's dominated convergence theorem, the above limit is equal to:

$$\underline{E}[g|n,\hat{y}_n] = \frac{\int g(w)\exp\left(n\left(\frac{\ell+n\hat{y}_n}{n}w - b(w)\right)\right)dw}{\int \exp\left(n\left(\frac{\ell+n\hat{y}_n}{n}w - b(w)\right)\right)dw}. \tag{A7}$$

Thus, (A6) is equivalent to considering the posterior expectation obtained from the kernel $p(w) = \exp(\ell w) = \lim_{r\to\infty}\exp(\ell w) I_{(-\infty,r]}(w)$. In analogy with (A3), we can define translation invariant lower and upper expectations as limits of exponential priors truncated on finite intervals:

$$\underline{E}[g] = \liminf_{r\to\infty} \frac{\ell}{\exp(\ell r)} \int g(w)\exp(\ell w) I_{(-\infty,r]}(w)dw, \tag{A8}$$

for $r \in \mathbb{R}$, which satisfies translation invariance, $\underline{E}[g(f_a)] = \underline{E}[g]$, for any g in the set of Borel-measurable RVBFs $\mathscr{K}$.

Lemma A.2. *The lower expectation model in (A8) satisfies $\underline{E}[g(f_a)] = \underline{E}[g]$ (translation invariance) for any $g \in \mathscr{K}$ and $a \in \mathbb{R}$.*

The proof is given in the second part of the proof of Lemma 4.1. Observe that $\underline{E}$ defined in (A8) satisfies the following properties: (i) for $g = I_{\mathbb{R}}$, $\underline{E}[I_{\mathbb{R}}] = 1$; (ii) for $A = [0,\infty)$, $\underline{E}[I_A] = 1$; (iii) for $g = I_A$, where A is a bounded subset of $\mathbb{R}$, it holds that $\overline{E}[I_{\{A+a\}}] = \overline{E}[I_A] = 0$ for any $a \in \mathbb{R}$. Thus, because of (iii), none of the probabilities dominated by $\overline{E}$ can be countably additive and the set $\mathscr{M}$ includes finitely additive probabilities. However comparing the lower probabilities in (ii) for (A3) and (A8), it can be noticed that the set of finitely additive probabilities induced by the two improper priors $p(w) = 1$ and, respectively, $p(w) = \exp(\ell w)$ are different. Observe that (A7) holds for any likelihood belonging to the exponential families provided that $\mathscr{W} = \mathbb{R}$. Note that $p(w) = \exp(\ell w)$, transformed back to the original parameter space, becomes $p(x) = \exp(\ell x)$ in the Normal case, $p(\theta) = \theta^{\ell-1}(1-\theta)^{-\ell-1}$ in the Beta case and $p(\lambda) = \lambda^{\ell-1}$ in the Gamma case. It can be noticed that for $\ell = 0$, $\exp(\ell w) = 1$, thus the improper uniform distribution can be seen as a particular case of $\exp(\ell w)$. We are interested in these exponential kernels $\exp(\ell w)$ since they preserve conjugacy with exponential families as shown in (A6). In summary, invariance properties can in general be satisfied by considering the lower expectation obtained as the limit of proper truncated priors.

References

[1] H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1983. 3rd edition.

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

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

[4] Mervyn Stone. Review and analysis of some inconsistencies related to improper priors and finite additivity. *Studies in Logic and the Foundations of Mathematics*, 104:413–426, 1982.

[5] Bruce M Hill and David Lane. Conglomerability and countable additivity. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 366–379, 1985.

[6] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.
[7] 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.
[8] P. J. Huber. The use of Choquet capacities in statistics. *Bull. Internat. Statist. Inst*, 45(4):181–191, 1973.
[9] S. Sivaganesan and J.O. Berger. Ranges of posterior measures for priors with unimodal contaminations. *The Annals of Statistics*, 17(2):868–889, 1989.
[10] L. DeRoberts and J.A. Hartigan. Bayesian inference using intervals of measures. *The Annals of Statistics*, 9(2):235–244, 1981.
[11] L. R. Pericchi and P. Walley. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, pages 1–23, 1991.
[12] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.
[13] A. Benavoli and M. Zaffalon. A model of prior ignorance for inferences in the one-parameter exponential family. *Journal of Statistical Planning and Inference*, 142(7):1960–1979, 2012.
[14] P. Walley. A bounded derivative model for prior ignorance about a real-valued parameter. *Scandinavian Journal of Statistics*, 24(4):463–483, 1997.
[15] F. P. A. Coolen and T. Augustin. A nonparametric predictive alternative to the imprecise dirichlet model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2):217–230, 2009.
[16] Anirban DasGupta. *Asymptotic theory of statistics and probability*. Springer, 2008.
[17] L. D. Brown. *Fundamentals of statistical exponential families: with applications in statistical decision theory*. IMS, Hayward, California, 1986.
[18] O. Barndorff-Nielsen. *Information and exponential families in statistical inference*. Wiley, New York, 1978.
[19] G. Letac. *Lectures on natural exponential families and their variance functions*. Number 50. Conselho Nacional de Desenvolvimento Científico e Tecnológico, Instituto de Matemática Pura e Aplicada, 1992.
[20] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.
[21] 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.
[22] J.-M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2–3):123–150, 2005.
[23] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. CRC press, 2004.
[24] J. Hartigan. Invariant prior distributions. *The Annals of Mathematical Statistics*, pages 836–845, 1964.
[25] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, pages 1343–1370, 1996.
[26] E. Regazzini. De finetti's coherence and statistical inference. *The Annals of Statistics*, pages 845–864, 1987.