The Geometry of Imprecise Inference

Miķelis Bickis
Department of Mathematics and Statistics
University of Saskatchewan
bickis@snoopy.usask.ca

Abstract

A statistical model can be constructed from a null probability measure by defining a set of statistics representing log-likelihood ratios of alternative measures to the null measure. Conversely, any model consisting of equivalent measures can be so expressed. A linear combination of statistics will also define a log-likelihood ratio if the normalizing constant is finite. In this way, any such model can be naturally extended to a convex subset of the linear span of these statistics. A finite dimensional subset defines an exponential family with the canonical parameters of a measure defined by coordinates relative to a set of basis functions.

Given a base measure on the parameter space, one can implement a similar structure with a set of parametric functions. The log-likelihood itself being a parametric function, the set of all possible log-likelihoods thus defines a space of measures conjugate to the statistical model. The conjugate space will have one more dimension spanned by the above-mentioned parameter-dependent normalizing constant.

If the base measure is considered a prior distribution, then the translation by the observed log-likelihood defines the posterior. An imprecise prior defined by a set of measures is in the same manner translated to a set of posterior measures. Upper and lower previsions can then be computed as extrema over this posterior set.

Keywords. Information geometry, exponential family, sets of measures.

1 Introduction

Statistical inference deals with observations that are realizations of a random process whose probability law is postulated to be one of a set of probability laws. We call this set the *model space*. Bayesian inference also requires a probability measure defined on the model space indexed by a set of *parameters* such that the distribution of the observations is viewed as being conditional on an unobserved realized parameter. Bayes' rule is then used to combine the *prior distribution* on the model space with the observation to give a *posterior distribution* on the model space, which will hopefully be more informative than the prior. This procedure is called *Bayesian updating*, but in the computer science community it is also known as *learning from data*, a terminology that is more descriptive of what is actually happening.

While Bayesian inference is based on a solid mathematical foundation, its use has been much criticized as being an improper method for scientific investigation (see Mayo [10] for an overview). One of the criticisms relates to the arbitrariness of the prior distribution. The subjectivity reflected in the prior seems out of place in the objectiveness of science. Even if one acknowledges that all inference relies on prior assumptions that are inherently subjective, there remains the practical issue of enunciating these assumptions sufficiently precisely to define a probability distribution on the model space.

These criticisms were addressed in Walley's fundamental treatise [14]. Walley introduces the concepts of lower and upper *previsions* on a set of *gambles*. In more conventional language, gambles are just random variables, and the term prevision (borrowed from de Finetti [6]), is essentially an expectation. Walley's novelty is in allowing the prevision to be defined on only a subset of random variables, thus providing for an incomplete description of a prior probability distribution which is more realistic than the classical Bayesian requirement. Moreover, Walley posits so-called *upper* and *lower previsions* which are merely bounds on the expectations, thereby further providing for incomplete knowldege, freeing one from having to specify a precise number as the prior expectation of any random variable. When applied to indicator variables, upper and lower previsions define upper and lower probabilities. Walley's development however is constrained

by the assumption that gambles are bounded. The case of unbounded gambles is discussed by Troffaes and de Cooman [13].

Walley's lower envelope theorem [14, Section 3.3.3] shows that if the upper and lower previsions satisfy coherence axioms, then they can be expressed in terms of conventional expectations: One can find a set of probability measures (dubbed *credal set* by Levi [9]) with corresponding expectation functionals, such that the lower prevision is the infimum of all expectations over the set, and the upper prevision is the supremum. Thus working with upper and lower previsions is equivalent to replacing probability measures with sets of probability measures.

Inference can now be based on such imprecise prior probabilities. Walley proposed a *generalized Bayes' rule* in which imprecise prior probabilities are updated to imprecise posterior probabilities. The posterior probabilities would then be expected to be more precise than the priors in the sense that the difference between upper and lower probabilities is reduced. Walley [15] also introduced the *imprecise Dirichlet model* (IDM) for learning from multinomial data, in which the priors are defined as a set of Dirichlet distributions with a fixed concentration parameter s, and the posteriors are Dirichlet distributions with s increased by the sample size.

Diaconis and Ylvisaker [5] discussed the process of Bayesian updating in exponential families. When the model space is an exponential family, then one can define a conjugate exponential family of prior distributions (indexed by *hyperparameters*) on the model parameters such that Bayesian updating can be expressed as a data induced change in the hyperparameters. Moreover, under certain regularity conditions, the predictive expectations of the canonical sufficient statistics can be expressed as a weighted average of prior expectation and sample mean.

Since the multinomial and Dirichlet distributions are conjugate in the sense of Diaconis and Ylvisaker, Walley's IDM can be viewed as an imprecise probability version of their setup. Imprecise versions of other exponential families have been proposed by Quaeghebeur and de Cooman [12], Quaeghebeur [11], Bickis [4], Benavoli and Zaffalon [3], Bataineh [2], and Lee [8]. The problematic step in all these situations is determining a set of priors. One wants a set sufficiently large such that previsions are near-vacuous *a priori* but not so large that learning from data is not possible. Such a set of priors will be said to have the *Benavoli-Zaffalon (BZ) property* as discussed in their paper [3].

In this paper, we consider a geometric representation of model and prior probabilities in which the idea of conjugacy is extended beyond that considered by Diaconis and Ylvisaker. Using canonical parameterizations, Bayes' rule can be seen as a data-dependent translation of a point representing the prior distribution. The generalized Bayes rule can similarly be seen as a translation of an entire set. We can thus visualize how various choices of prior set affect the process of learning from data. We present several examples to illustrate various situations that arise in this paradigm. In most of these examples we consider the effect of a single observation. The effect of i.i.d. samples should then be viewed as iterations of the updating paradigm, illustrating the effect of accumulating information.

2 Geometry of Probability Measures

Let $\mathcal{Y}$ be an observation space whose elements represent possible empirical observations. We make few assumptions about the structure of this space; elements may be numeric or nominal, scalar or vector of finite or infinite dimension. All we require is that we are able to specify a probability measure P_0 on some σ-algebra of events defined on $\mathcal{Y}$. We are interested in making an inference about the probabilistic nature of $\mathcal{Y}$ and may think of P_0 as a null model which we wish to compare with some other putative measure P_1. We will assume that no deterministic inference is possible, i.e., that any event that is possible (with positive probability) under one measure is similarly possible under another. In the language of probability theory, P_0 and P_1 are equivalent measures: $P_0 \sim P_1$.

2.1 One-Dimensional Case

The likelihood principle implies that any inference concerning P_1 vs. P_0 is based solely on the likelihood ratio, which is convenient to express in its logarithmic form:

$$\ell = \log \frac{dP_1}{dP_0}, \tag{1}$$

from which it follows that we can write

$$P_1(A) = \int \mathbf{1}_A e^{\ell}\, dP_0 \tag{2}$$

where $\mathbf{1}_A$ represents the indicator function of a measurable subset A of $\mathcal{Y}$. By introducing a scalar parameter θ, we can define one-dimensional exponential family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where

$$P_\theta(A) = \int \mathbf{1}_A \exp(\theta\ell - \phi(\theta))\, dP_0, \tag{3}$$

$$\phi(\theta) = \log \int e^{\theta\ell}\, dP_0, \tag{4}$$

and

$$\Theta = \{\theta \in \mathbb{R} : \phi(\theta) < \infty\}. \tag{5}$$

Theorem 1 *Θ is a convex set.*

Proof: If $\theta_1, \theta_2 \in \Theta$ and $0 < \alpha < 1$ then

$$e^{\phi(\alpha\theta_1+(1-\alpha)\theta_2)} = \int e^{\alpha\theta_1\ell} e^{(1-\alpha)\theta_2\ell}\, dP_0.$$

By Hölder's inequality, this is less than

$$\left(\int \left(e^{\alpha\theta_1\ell}\right)^{1/\alpha} dP_0\right)^{\alpha} \left(\int \left(e^{(1-\alpha)\theta_2\ell}\right)^{1/(1-\alpha)} dP_0\right)^{1-\alpha}$$
$$= \phi(\theta_1)^{\alpha}\phi(\theta_2)^{1-\alpha}.$$

Since $\phi(\theta_1)$ and $\phi(\theta_2)$ are both finite by definition, so is $\phi(\alpha\theta_1 + (1-\alpha)\theta_2)$, and the result follows. ∎

Instead of postulating an alternative probability model P_1, we may start with a random variable (i.e., measurable function) v on $\mathcal{Y}$ that we think encapsulates the inference we are interested in making. In the same fashion we may define a one-dimensional exponential family

$$P_\theta(A) = \int \mathbf{1}_A \exp(\theta v - \phi(\theta))\, dP_0, \quad \theta \in \Theta \tag{6}$$

where ϕ and Θ are defined as before in (4) and (5).

Definition 1 *The family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ defined by ((6)) will be called the family generated by v (over P_0).*

Definition 2 *For any random variable T, $E_\theta(T)$ is defined as*

$$E_\theta(T) = \int T\, dP_\theta = \int T e^{\theta v - \phi(\theta)}\, dP_0.$$

Theorem 2 *If $\theta_1 \neq \theta_2$ then $P_{\theta_1} \neq P_{\theta_2}$ iff v is not almost surely constant.*

Proof:

$$P_{\theta_1} = P_{\theta_2} \iff P_0\left\{e^{\theta_1 v - \phi(\theta_1)} = e^{\theta_2 v - \phi(\theta_2)}\right\} = 1$$

which is equivalent to

$$(\theta_1 - \theta_2)v = \phi(\theta_1) - \phi(\theta_2) \quad \text{a.s.} \tag{7}$$

Since the right side of (7) is constant, this equality can hold only if v is almost surely constant or if $\theta_1 = \theta_2$. On the other hand, if v is almost surely constant, then

$$\phi(\theta) = \log \int e^{\theta v}\, dP_0 = \theta v \quad \text{a.s} \tag{8}$$

and thus (7) holds. ∎

If v is almost surely constant, then $P_\theta = P_0$ for all θ. On the other hand, if

$$\int e^{\theta v}\, dP_0 = \infty \quad \text{for all } \theta \neq 0, \tag{9}$$

then Θ consists of a single point. In either case the family generated by v has but a single probability measure and provides no prospect for inference. In the following we will assume that v is not constant and that $\int \exp(\theta_1 v)\, dP_0 < \infty$ for at least one $\theta_1 \neq 0$. By Theorems 1 and 2, Θ will then include an interval with endpoint θ_1, with distinct θ's corresponding to distinct probability measures.

Theorem 3 *If v_1 and v_2 are random variables on $\mathcal{Y}$ such that $v_1 - v_2$ is almost surely constant, then for any $\theta \in \Theta$, v_1 and v_2 define the same probability measure and hence v_1 and v_2 generate the same family.*

Proof: Let

$$\phi_i(\theta) = \log \int e^{\theta v_i}\, dP_0, \quad i = 1, 2.$$

Then

$$\phi_2(\theta) = \log \int e^{\theta v_1} e^{\theta(v_2 - v_1)}\, dP_0$$
$$= \phi_1(\theta) + \theta(v_2 - v_1) \quad \text{a.s.} \tag{10}$$

For any event A, $\theta \in \Theta$, the probability defined by v_1 is

$$\int \mathbf{1}_A e^{\theta v_1 - \phi_1(\theta)}\, dP_0 = \int \mathbf{1}_A e^{\theta v_2 - (\phi_1(\theta) + \theta(v_2 - v_1))}\, dP_0, \tag{11}$$

$$= \int \mathbf{1}_A e^{\theta v_2 - \phi_2(\theta)}\, dP_0, \tag{12}$$

by (10). ∎

The random variable v may thus differ from a log likelihood ratio by an arbitrary constant. We can make the representation (6) unique by requiring that

$$\int v\, dP_0 = 0. \tag{13}$$

Since

$$\theta v = \log \frac{dP_\theta}{dP_0} + \phi(\theta),$$

the convention (13) implies that

$$\phi(\theta) = \int \log \frac{dP_0}{dP_\theta}\, dP_0. \tag{14}$$

The right side of (14) was described by Kullback [7] as the *mean information for discrimination in favour of P_0 against P_θ* and is one way of quantifying the ease with which a probability measure P_θ can be distinguished from P_0. It is commonly called the *Kullback-Leibler information* or *divergence* [1] and denoted by $I(P_0|P_\theta)$. The divergence may be viewed as the distance from P_0 to P_θ, although it does not satisfy the axioms of a metric.[1] A significant property of

[1] While $I(P_0|P_\theta) > 0$ iff $P_0 \neq P_\theta$, it is not symmetric, does not satisfy the triangle inequality and may even be infinite. However, it can be shown that $I(P_0|P_\theta) < \infty$ when θ is in the *interior* of the set (5).

divergence is additivity over independent observations. Let $P_\theta^{(1,2)} = P_\theta^{(1)} \times P_\theta^{(2)}$ be the joint distribution of two independent observations with distributions $P_\theta^{(1)}$ and $P_\theta^{(2)}$. Then

$$I(P_0^{(1,2)}|P_\theta^{(1,2)}) = I(P_0^{(1)}|P_\theta^{(1)}) + I(P_0^{(2)}|P_\theta^{(2)}). \quad (15)$$

The requirement (13) makes the representation unique, and relates the normalizing constant ϕ to the divergence.

The set of random variables forms a vector space, and the representation (6) identifies a family of probability measures with a convex subset of a one-dimensional subspace, the origin representing the null measure P_0. The function v is a basis vector such that all probability measures in the family can be represented as scalar multiples of v, the scalar being the parameter θ. Because of the need of a normalizing constant $\phi(\theta)$, the log-likelihood ratios actually do not lie in a one-dimensional subspace, but in a two-dimensional subspace spanned by v and the constant function $\mathbf{1}$ equal to 1 everywhere. A probability measure P_θ actually corresponds to an equivalence class of vectors differing by a multiple of $\mathbf{1}$. The convention (13) picks a particular representative of the equivalence class.

We illustrate these ideas with an almost trivial example.

Example 1. Let $\mathcal{Y} = \{0, 1\}$ with $P_0\{0\} = P_0\{1\} = \frac{1}{2}$ and $P_1\{0\} = 1 - P_1\{1\} = 1 - p$ for some $p \in (0, 1)$. Then

$$\frac{dP_1}{dP_0}(0) = (1-p)/\tfrac{1}{2} \qquad \frac{dP_1}{dP_0}(1) = p/\tfrac{1}{2}$$

so that

$$\begin{aligned}\frac{dP_1}{dP_0}(y) &= 2(1-p)^{1-y}p^y \\ \log\frac{dP_1}{dP_0} &= \log 2 + (1-y)\log(1-p) + y\log p \\ &= \log\frac{p}{1-p}\,y + \log 2 + \log(1-p).\end{aligned}$$

putting $v(y) = y - \frac{1}{2}$ and $\theta = \log(p/(1-p))$ we have that

$$\begin{aligned}\log\frac{dP_1}{dP_0} &= \theta v + \theta/2 + \log 2 - \log(1+e^\theta) \\ &= \theta v - \log\left(\frac{1+e^\theta}{2e^{\theta/2}}\right) \\ &= \theta v - \log\cosh(\theta/2). \quad (16)\end{aligned}$$

The family of binary distributions is thus displayed in the form (6) parametrized by the log-odds $\theta = \log(p/(1-p))$ with

$$\phi(\theta) = I(P_0|P_1) = \log\cosh(\theta/2).$$

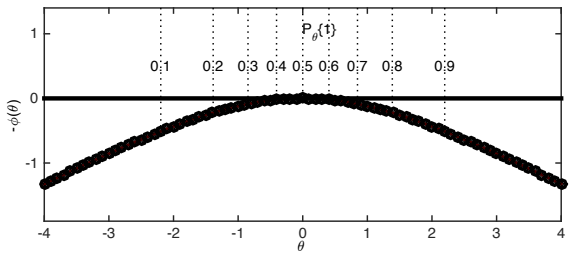

Figure 1: Probability manifold for binary distributions. The set of measures forms a one-dimensional manifold in the plane. The distance in the vertical directions represents the divergence from the uniform distribution. The points in the manifold can be projected in this direction onto the tangent plane. Location along the plane is linear in the canonical parameter θ, but non-linear in the success probability p.

The set of *all* functions on $\{0, 1\}$ is two-dimensional, being isomorphic to $\mathbb{R}^2$. The representation (16) identifies those functions that are log-likelihood ratios relative to uniform probabilities. Using a basis consisting of the functions $v_0(y) = 1$ and $v_1(y) = y - \frac{1}{2}$, the set of log-likelihood ratios (equivalently, probability measures) can be visualized as in Figure 1. In this figure, the equivalence classes correspond to vertical lines.

Example 2. Suppose now that we have n i.i.d. observations $y_1, \ldots, y_n$ from Example 1. Let $P_{0,n}$ (resp. $P_{1,n}$) represent the joint distribution of n i.i.d. binary observations with success probability $\frac{1}{2}$ (esp. p). Again, let $\theta = \log(p/(1-p))$. The joint log likelihood of independent observations is the sum of the log likelihoods, and by (15) the same will be true for the divergences. Thus adding terms of the form (16) we get

$$\log\frac{dP_{1,n}}{dP_{0,n}}(y_1, \ldots, y_n) = \theta\left(\sum_{i=1}^n y_i - \frac{n}{2}\right) - n\log\cosh\frac{\theta}{2}, \quad (17)$$

which is the canonical form of the binomial family. Alternatively, let $\mathcal{Y}$ be the set of all 2^n binary sequences and P_0 be the uniform measure on this set. Then if we decide that inferences are to made solely on the basis of the function $v(y_1, \ldots, y_n) = \sum_i y_i$, the family generated by v is again binomial. The picture of this family is just a rescaling of Figure 1 and thus has the same intrinsic geometry. This geometric equivariance under repeated sampling is characteristic of exponential families.

Example 3. Let $\mathcal{Y} = \mathbb{R}^+$, and define P_0 by the cumulative distribution function

$$P_0((0, y]) = 1 - e^{-y}, \quad y > 0,$$

then with $v_1(y) = y - 1$ the one-dimensional exponential family is

$$\log \frac{dP_\theta}{dP_0} = \theta v_1 - \phi(\theta) \tag{18}$$

where

$$\phi(\theta) = I(P_0|P_\theta) = -\theta - \log(1-\theta).$$

The natural parameter space is $\Theta = (-\infty, 1)$ which defines the family of exponential distributions with expectation $(1-\theta)^{-1}$.

2.2 Multidimensional Case

The inference of interest may not be expressible in terms of a single function v; we may require a family of functions $\mathcal{L}_0$, in which case a construction as in (3) is possible for any $v \in \mathcal{L}_0$. Indeed, for any finite number of functions $v_1, \ldots, v_k \in \mathcal{L}_0$ and scalar parameters $\theta_1, \ldots, \theta_k$ we can construct a probability measure

$$\begin{aligned} P_{\boldsymbol{\theta}}(A) &= P_{\theta_1,\ldots,\theta_k}(A) \\ &= \int \mathbf{1}_A \exp\left(\sum_{i=1}^{k} \theta_i v_i - \phi(\boldsymbol{\theta})\right) dP_0 \end{aligned} \tag{19}$$

provided that

$$\phi(\boldsymbol{\theta}) = \log \int \exp\left(\sum_i \theta_i v_i\right) dP_0 < \infty. \tag{20}$$

Thus a given set $\mathcal{L}_0$ of functions can be augmented by their linear combinations, the set $\mathcal{L}$ of all such linear combinations forming a vector space. In that case we have a generalization of Definition 1:

Definition 3 *Given a set $\mathcal{L}_0$ of random variables, the set of probability measures defined by* (19) *and* (20) *will be called the family generated by $\mathcal{L}_0$.*

If for a fixed set of functions $v_1, \ldots, v_k$ every function in $\mathcal{L}$ can be *uniquely* expressed as a linear combination of $v_1, \ldots, v_k$, then $\mathcal{L}$ will be a k-dimensional vector space and $v_1, \ldots, v_k$ will be basis vectors. The vector space will be infinite dimensional if no such finite basis can be found.[2] We focus on the finite-dimensional case. Here it is convenient to fix a basis $v_1, \ldots, v_k$ and consider $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_k)$ representing the measure $P_{\boldsymbol{\theta}}$ as a row vector and the values $v_1(y), \ldots, v_k(y)$ as a column vector $\mathbf{v}$. Then the vectors of parameters and statistics act on each other via matrix multiplication. Thus, the family generated by $v_1, \ldots, v_k$ can be represented as

$$\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \quad \text{where} \tag{21}$$

$$P_\theta(A) = \int \mathbf{1}_A e^{\boldsymbol{\theta}^\top \mathbf{v} - \phi(\boldsymbol{\theta})}\, dP_0 \tag{22}$$

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^k : \phi(\boldsymbol{\theta}) = \int e^{\boldsymbol{\theta}^\top \mathbf{v}}\, dP_0 < \infty\}. \tag{23}$$

As in the one-dimensional case, the log likelihood ratio may differ by a constant from a function in $\mathcal{L}$. Thus if $\mathcal{L}$ is k-dimensional with basis $v_1, \ldots, v_k$, the set of log-likelihood ratios lies in a $k+1$-dimensional space spanned by $v_0, v_1, \ldots, v_k$, where $v_0 = \mathbf{1}$. Again, two functions that differ by a scalar multiple of $\mathbf{1}$ will define the same probability measure, and we can consider probability measures to correspond to equivalence classes of functions. To make the representation (22) unique we add the additional constraint that $E_0(v_i) = 0$ for every $i \geq 1$, which again will specify a representative of the equivalence class. In that case the normalizing constant $\phi(\boldsymbol{\theta}) = I(P_0|P_{\boldsymbol{\theta}})$ as discussed before. Uniqueness also requires that the functions $v_0, v_1, \ldots, v_k$ are linearly independent *when restricted to the support of* P_0.

With these additional conditions, for each $P_{\boldsymbol{\theta}} \in \mathcal{P}$, $\log dP_{\boldsymbol{\theta}}/dP_0$ corresponds to a unique point $(-I(P_0|P_{\boldsymbol{\theta}}), \theta_1, \ldots, \theta_k)$ in $\mathbb{R}^{k+1}$. The set of probability measures thus defines a k-dimensional manifold

$$\mathcal{M} = \{(-I(P_0|P_{\boldsymbol{\theta}}), \theta_1, \ldots, \theta_k) : I(P_0|P_{\boldsymbol{\theta}}) < \infty\} \tag{24}$$

embedded in $\mathbb{R}^{k+1}$. This manifold can be projected one-to-one onto its tangent plane at the origin, giving the *natural parameter space*[3]

$$\Theta = \{\boldsymbol{\theta} : I(P_0|P_{\boldsymbol{\theta}}) < \infty\} \tag{25}$$

which is a convex subset of $\mathbb{R}^k$.

The family of normal distributions is a well-known example:

Example 4. Let $\mathcal{Y} = \mathbb{R}$, and let P_0 be the standard normal distribution.

$$P_0(A) = \frac{1}{\sqrt{2\pi}} \int \mathbf{1}_A e^{-y^2/2}\, dy,$$

and define $v_1(y) = y$, $v_2(y) = y^2 - 1$. The representation (19) gives

$$\log \frac{dP_{\boldsymbol{\theta}}}{dP_0} = \theta_1 y + \theta_2(y^2 - 1) - I(P_0|P_{\boldsymbol{\theta}})$$

[2] If $\mathcal{L}$ spans an infinite-dimensional space, then a basis might be impossible to find, even if its existence is implied by the axiom of choice.

[3] This is slightly more restrictive than the usual definition, which only requires the finiteness of $\phi(\theta)$ and not of its particular version $I(P_0|P_{\boldsymbol{\theta}})$.

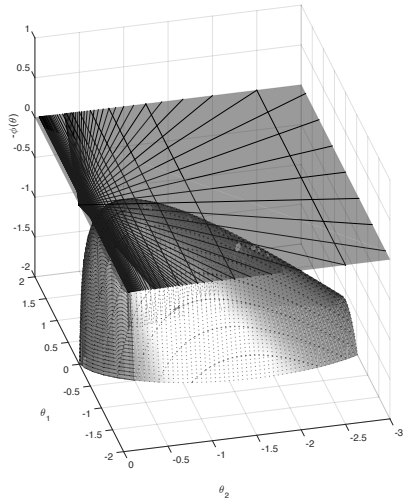

Figure 2: Probability manifold for the Gaussian family, with tangent plane at $P_0 = N(0,1)$. The tangent plane is ruled with coordinate lines corresponding to mean and variance.

where $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : \theta_2 < 0\}$. This can be seen to be a Gaussian distribution with mean $\mu = -\theta_1/(2\theta_2)$ and variance $\sigma^2 = -1/(2\theta_2)$.

Example 5. Consider now the setup of Example 3 but with the observation is right-censored at T. This means that if $y > T$ then one actually observes $y = T$. Now

$$P_0((0,y]) = \begin{cases} 1 - e^{-y} & 0 < y < T, \\ 1 & y \geq T. \end{cases}$$

so that the distribution is no longer continuous but has an atom, i.e., point of positive probability, at T.

Now let P_ϑ be an exponential distribution with mean $(1-\vartheta)^{-1}$ also censored at T. Then the log likelihood ratio is

$$\ell = \log \frac{dP_{\theta'}}{dP_0} = \begin{cases} \theta' y + \log(1-\theta') & 0 < y < T, \\ \theta' T & y \geq T. \end{cases}$$

The one-dimensional family generated by ℓ now has natural parameter space $(-\infty, \infty)$, but only P_0 and P_φ represent censored exponential distributions.[4] To model a family of censored exponential distributions, we need to introduce a second function $\delta = \mathbf{1}_{y<T}$. For any exponential distribution censored at T we can now write

$$\log \frac{dP_\theta}{dP_0} = \theta_1 v_1 + \theta_2 v_2 - \phi(\theta_1, \theta_2). \tag{26}$$

[4]Each of the members of the family is a mixture of a truncated exponential distribution and a point mass at T, but the probability of the point mass in most cases is different from that given by censoring.

The canonical representation with $\phi(\theta_1, \theta_2) = I(P_0|P_{\theta_1,\theta_2})$ would require that

$$v_1(y) = y - E_0(y) = y - (1 - e^{-T}) \tag{27}$$
$$v_2(y) = \delta - E_0(\delta) = \delta - (1 - e^{-T}). \tag{28}$$

$$\begin{aligned} \phi(\theta_1, \theta_2) &= I(P_0|P_{\theta_1,\theta_2}) \\ &= \log\left(e^{(\theta_1-1)T} + (\theta_1 - 1)e^{(\theta_2-1)T-\theta_2}\right) \\ &\quad - \log(\theta_1 - 1) + \theta_2 \\ &\quad - (\theta_1 + \theta_2)\left(1 - e^{-T}\right). \end{aligned}$$

Exponential distributions censored at T form a one-dimensional non-linear manifold, defined by

$$\{(\theta_1, \theta_2) : \theta_2 = \log(1 - \theta_1)\},$$

in this two-dimensional exponential family. Such a family is called a *curved exponential family*[1]. Restricted to this submanifold, we have

$$I(P_0|P_{\theta_1,\theta_2}) = -(\theta_1 + \theta_2)(1 - e^{-T}),$$

which in this instance is a *linear* function of the canonical parameters.

3 Geometry of Inference

3.1 Precise Priors

Suppose now that we express our prior uncertainty about the model by a probability measure Π_0 defined on a suitable σ-algebra of subsets of $\mathcal{P}$.

Denote by π_0 the density of Π_0 (considered as a measure on $\mathcal{P}$) with respect to some dominating measure λ. Then if the likelihood is given by (21) and an observation y is observed, Bayes' rule will give the posterior density

$$\pi_y(v) = \frac{\pi_0(\boldsymbol{\theta}) \exp(\boldsymbol{\theta}^\top \mathbf{v}(y) - I(P_0|P_{\boldsymbol{\theta}}))}{\int \pi_0(\vartheta) \exp(\vartheta^\top \mathbf{v}(y) - I(P_0|P_\vartheta))\, d\lambda(\vartheta)}, \tag{29}$$

where $\mathbf{v}(y)$ is the vector $(v_1(y), \ldots, v_k(y))^\top$. If we take the log ratio of posterior to prior, we get

$$\log \frac{d\Pi_y}{d\Pi_0}(v) = \boldsymbol{\theta}^\top \mathbf{v} - I(P_0|P_{\boldsymbol{\theta}}) - \psi(y) \tag{30}$$

where

$$\psi(y) = \log \int \exp\left(\boldsymbol{\theta}^\top \mathbf{v} - I(P_0|P_{\boldsymbol{\theta}})\right)\, d\Pi_0(v). \tag{31}$$

The set of possible posteriors (30) is of the same exponential form as (19) where the roles of parameter and function are reversed.

Let $\mathcal{L}^*$ be the vector space of functions $v^* : \mathcal{P} \to \mathbb{R}$ spanned by

$$v^0 : P \mapsto -I(P_0|P) \quad \text{and} \tag{32}$$

$$v^i : (P_{\theta_1,\ldots,\theta_k}) \mapsto \theta_i \quad i = 1, \ldots, k. \tag{33}$$

For brevity, denote by $\mathbf{v}^*$ the *row* vector $(v^0(P), v^1(P), \ldots, v^k(P))^\top$.

Now given a vector $\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_k)$ of *hyperparameters* we can now define analogously to (19) for any measurable set W of measures in $\mathcal{P}$

$$\Pi_{\boldsymbol{\eta}}(W) = \int \mathbf{1}_W \exp\left(\sum_{i=0}^{k} v^i \eta_i - \psi(\boldsymbol{\eta})\right) d\Pi_0. \tag{34}$$

This will define a probability measure on $\mathcal{P}$ provided that

$$\psi(\boldsymbol{\eta}) = \int e^{\mathbf{v}^* \boldsymbol{\eta}}\, d\Pi_0 < \infty. \tag{35}$$

Definition 4 *The conjugate hyperparameter space* Θ^* *is the set of all* $\boldsymbol{\eta} \in \mathbb{R}^{k+1}$ *such that* (35) *holds.*

Definition 5 *The space of measures* $\mathcal{P}^*$ *conjugate* [5] *to the family* $\mathcal{P}$ *is the set* $\{\Pi_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \Theta^*\}$.

By definition, $\mathcal{P}^*$ includes the prior distribution and all possible posteriors (but is generally much larger). Moreover, if a posterior distribution is in $\mathcal{P}^*$, then a proper prior from which it was updated must also be in $\mathcal{P}^*$.

Theorem 4 *If a prior distribution* $\Pi_{\boldsymbol{\eta}}$ *in* $\mathcal{P}^*$ *has hyperparameters*

$$\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots, \eta_k)$$

then after observing y *the posterior distribution will have hyperparameters*

$$(\eta_0 + 1, \eta_1 + v_i(y), \ldots, \eta_k + v_k(y)).$$

Proof: The density of the prior Π is $d\Pi/d\Pi_0 = \exp(\mathbf{v}^*\boldsymbol{\eta} - \psi(\boldsymbol{\eta}))$. By Bayes' theorem, the posterior density is

$$\frac{d\Pi_y}{d\Pi_0} = \frac{e^{\mathbf{v}^*\boldsymbol{\eta} - \psi(\boldsymbol{\eta})} e^{\boldsymbol{\theta}^\top \mathbf{v}(y) - I(P_0|P_{\boldsymbol{\theta}})}}{\int e^{\mathbf{v}^*\boldsymbol{\eta} - \psi(\boldsymbol{\eta})} e^{\boldsymbol{\theta}^\top \mathbf{v}(y) - I(P_0|P_{\boldsymbol{\theta}})}\, dP_{\boldsymbol{\theta}}} \tag{36}$$

By definition,

$$\mathbf{v}^*(P_{\boldsymbol{\theta}}) = (-I(P_0|P_{\boldsymbol{\theta}}), \theta_1, \ldots, \theta_k)$$

so the numerator of (36) becomes $\exp\big(\mathbf{v}^*(\boldsymbol{\eta} + (1, \mathbf{v}(y))\big)$, and the denominator becomes $\psi(\boldsymbol{\eta} + (1, \mathbf{v}(y))$. ∎

[5] This definition is more general than that of Diaconis and Ylvisaker. Their construction would follow from using a (possibly improper) Lebesgue prior for Π_0.

The transformation from prior to posterior by an observation y can be represented in Θ^* as a translation by the vector $(1, v_1(y), \ldots, v_k(y))$ is a translation by the vector $y^* - v_0$. Note that the translation is the same for all priors. Even improper priors can be accommodated by going outside Θ^*.

3.2 Imprecise Priors

Because the translation in Θ^* is the same for all priors (proper or improper), one can update a *set* of priors simply by translating the whole set. This provides a convenient way of representing updating of imprecise priors, as the set of hyperparameters for the posteriors is congruent to the set of prior hyperparameters.

It is often of interest to predict the value of some future observation, by the posterior expectation of a random variable $v \in \mathcal{L}$. With a precise prior distribution Π_0, this would be computed as

$$\hat{v} = \int \int v(z)\, dP_\theta(z)\, d\Pi_y(\theta).$$

If instead of a precise prior, we have a set of priors $\boldsymbol{\Pi}_0$ leading to a set of posteriors $\boldsymbol{\Pi}_y$, then we compute *lower and upper previsions* as

$$\underline{v} = \inf_{\Pi \in \boldsymbol{\Pi}_y} \int \int v(z)\, dP_\theta(z)\, d\Pi(\theta) \tag{37}$$

$$\overline{v} = \sup_{\Pi \in \boldsymbol{\Pi}_y} \int \int v(z)\, dP_\theta(z)\, d\Pi(\theta). \tag{38}$$

If the conjugate family is of the type discussed by Diaconis and Ylvisaker, and if $v \in \mathcal{L}$, then the sets of constant predictive expectation $\hat{v}$ form hyperplanes in $\mathcal{L}^*$ that intersect in a subspace containing the improper Lebesgue prior. In this case the lower and upper previsions (37) and (38) are given by the supporting hyperplanes of the convex hull of the posterior set, which thus can, without loss of generality, be taken to be convex. If the prior set intersects all of these diverging hyperplanes, then the prior prediction is vacuous. As data are observed, the prior set is shifted such that it no longer intersects all the hyperplanes, and non-vacuous prediction can be made.

Definition 6 *A set of priors will be said to have the Benavoli-Zaffalon (BZ) property relative to the function* v *if* $\underline{v} > \inf v$ *and* $\overline{v} < \sup v$ *in* (37) *and* (38) *for some observation* y*, but* $\underline{v} = \inf v$ *and* $\overline{v} = \sup v$ *when* $\boldsymbol{\Pi}_y$ *is replaced by the prior set* $\boldsymbol{\Pi}_0$.

Example 6. Consider the setup in Example 1. For Π_0 take a logistic distribution of θ (which is equivalent to a uniform on $p = (1 + \exp(-\theta))^{-1}$):

$$\frac{d\Pi_0}{d\lambda}(\theta) = \frac{e^\theta}{(1 + e^{-\theta})^2}. \tag{39}$$

Define $v_0^*, vii_1^* \in \mathcal{L}^*$ by

$$v_0^*(\theta) = -\log\cosh(\theta/2)$$
$$v_1^*(\theta) = \theta/2$$

It can be shown that

$$\Theta^* = \{\eta_1 f_1^* + \eta_0 f_2^* : \ |\eta_1| < 1 + \eta_2/2\}.$$

Plotting the basis vector v_0^* horizontally and v_1^* vertically, the set of proper priors and posteriors is defined by the wedge-shaped region in Figure 3. The update rule for a single binary observation y can be expressed as

$$\eta_0 \mapsto \eta_0 + 1 \tag{40}$$
$$\eta_1 \mapsto \eta_1 + y - \tfrac{1}{2} \tag{41}$$

Given any point representing a prior, the posterior after a single observation is obtained by moving one step to the right, a half-step up for a success, a half-step down for a failure. A sequence of independent observations then traces a path in the hyperparameter space.

Sets of constant prediction of $v = y - \frac{1}{2}$ form rays emanating from $\eta_0 = -2, \eta_1 = 0$ (Figure 3). (The intersection of these rays is not in Θ^* but represents an improper prior.) From this picture, one can visualize which sets of priors will have the Benavoli-Zaffalon property. For example, Walley's imprecise beta model (IBM) gives a prior set corresponding to

$$\{(\eta_0, \eta_1) : \eta_0 = s, |\eta_1| < s/2\}, \tag{42}$$

where s is taken to be 1 or 2. The prior predictions are thus $\underline{v} = 0$ and $\overline{v} = 1$. After taking observations, the prior set has moved such that it is contained in a narrow cone of rays, leading to informative upper and lower previsions.

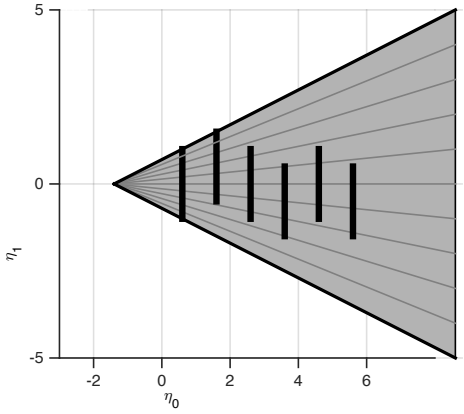

Figure 3: Path of sets of posteriors from IDM

Example 7. If the data are $N(\mu, 1)$, then the conjugate prior family would be $N(\nu, \sigma^2)$ which can be reparametrized in canonical exponential form by $\eta_0 = 1/2\sigma^2$ and $\eta_1 = \nu/\sigma^2$. If we choose $\Pi_0 \sim N(0, 1)$ then $\Theta^* = (-1, \infty) \times (-\infty, \infty)$. Sets of constant predictive expectation are again rays emanating from $(-1, 0)$ (Figure 4). Note that η_0 again represents a concentration parameter. Unlike the case of the IBM, fixing the set of priors by fixing the concentration parameter does not allow for learning from data, as the interval of posterior predictions remains infinite. Benavoli and Zaffalon [3] suggested using a set of priors which in the present parametrization is the rectangular region in Figure 4 which satisfies the BZ-property.

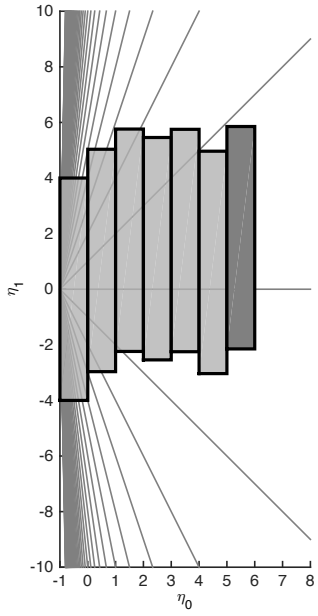

Figure 4: Set of posteriors from Normal distribution, using prior set suggested by Benavoli

Example 8. Let the model space be as in Example 6 but define Π_0 as a Gaussian distribution on Θ.

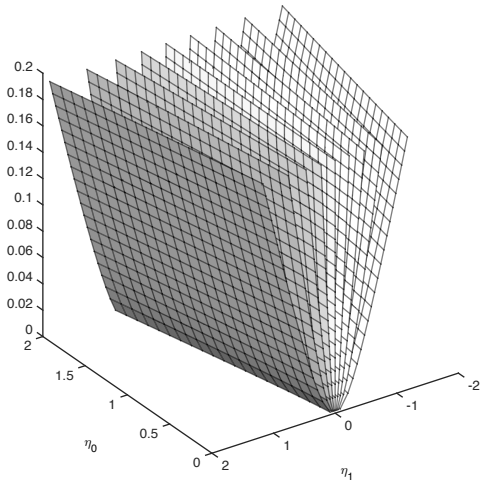

Figure 5: Contour sheets showing sets of constant predictions for the logit-normal model.

The same construction applies, but in this case the conjugate family is not conjugate in the sense of Diaconis and Ylvisaker. The 2-dimensional exponential family created from a $N(\mu, \sigma^2)$ prior is spanned by θ and $\cosh(\theta/2)$. Nonetheless, similar arguments still allow for learning from data. If we start with a set of $N(\mu, \sigma^2)$ priors for various σ^2, we obtain a 3-dimensional family of posteriors spanned by $\cosh(\theta/2)$, θ and $-\theta^2$. The update rules for η_0 and η_1 are the same as in (41), but η_2, the coefficient of $-\theta^2$, is not changed.

There seems to be no explicit formula for the normalizing constant ψ, nor for the predictive expectations. Nonetheless, such quantities can be computed numerically. As shown in Figure 8, the level sets of predictive expectations appear as a set of almost flat sheets pinched together at the origin. The limiting case $\eta_2 = 0$ is equivalent to the conjugate family in Example 6. The path traced by a sequence of observations is as in Figure 3, raised by $\eta_2 = 1/(2\sigma^2)$ in the prior distribution. A set of priors with the Benavoli-Zaffalon property can be obtained by including in its boundary a set of the type in (42).

Example 9. Consider now the censored exponential model of example (5). While the "natural" parameter space is two-dimensional, we are only concerned with models on the one-dimensional manifold $\theta_2 = \log(1 + \theta_1)$. We thus take as Π_0 the singular distribution concentrated on this manifold such that that θ_1 has an exponential distribution with mean 1. (Note that in this case the dominating measure λ is not Lebesgue measure.) The conjugate space of posteriors then takes the form

$$\log \frac{dP_0}{d\Pi_{\eta_1,\eta_2}} = \theta_1\eta_1 + \theta_2\eta_2 - \psi(\eta_1, \eta_2)$$

where

$$\psi(\eta_1, \eta_2) = (\eta_2 + 1)\log(\eta_1 + 1) - \Gamma(\eta_2 + 1)$$

The natural hyperparameter space is $\{\eta_1 > -1, \eta_2 > -1\}$. In this case the family is only two-dimensional because of the linear dependence between ϕ and (θ_1, θ_2).

The Bayesian updating rule is

$$\eta_1 \mapsto \eta_1 + y$$
$$\eta_2 \mapsto \eta_2 + \delta,$$

moving to the right by the observed lifetime and one step up if the lifetime is not censored. This setup still works if we allow T itself to vary with time. The hyperparameter keeps moving right while the individual is alive (i.e., censored) and then jumps up one step once a death is observed.

The posterior predictive expectation of the *uncensored* lifetime is $(\eta_1 + 1)/\eta_2$. To create an imprecise inference, we can start with the hyperparameter set $\{\eta_2 > 0, \eta_1 + \eta_2\} = 0$. Initially, the predictive lower prevision is 0, and the predictive upper prevision is ∞. If an individual is observed to be alive at time y, the lower prevision rises to y, but the upper prevision remains at ∞. Once the individual is observed to die at y, the upper prevision drops to $1 + y$ and the lower prevision drops to $y/2$. If one observes a set of independent lifetimes, then this process compounds. If t is the total of observed lifetimes and d is the total number of observed deaths, then the lower prevision is $t/(d + 1)$ and the upper prevision is $(t + 1)/d$. This set of priors again has the Benavoli-Zaffalon property (Figure 3.2).

4 Conclusions

In this paper we have shown how an exponential family of probability measures is generated by postulating a null distribution and a set of inferential functions. If the set of functions is k-dimensional, then the family of probability measures forms a k-dimensional manifold embedded in $k + 1$-dimensional Euclidean space. This manifold can be uniquely projected onto a tangent plane whose coordinates parametrize the model. If a prior distribution is defined on the set of probability distribution, then the above development can be repeated with the parametric functions, thus giving an exponential family that includes all possible posteriors.

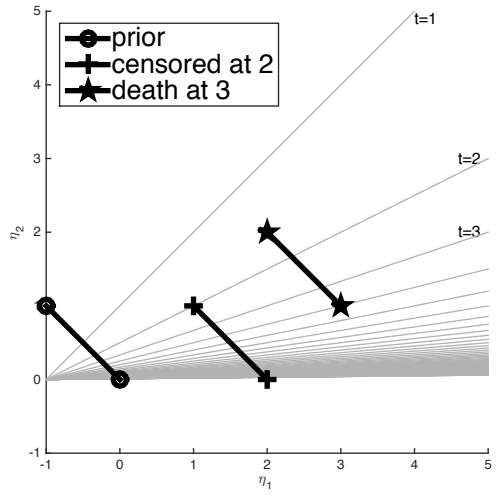

Figure 6: Imprecise updating of censored exponential survival times, showing the prior set, the set after an observation censored at time 2, and the set after observing a death after another time unit. The rays are level sets for predicted uncensored lifetimes.

This family can again be projected onto a tangent space of hyperparameters.

In this representation, Bayesian updating of a hyperparameter is expressed as a translation by a data-dependent vector. This same translation can be applied to a set of hyperparameters, demonstrating the updating of imprecise priors to imprecise posteriors. The geometric perspective allows one to see when a set of priors would enjoy the Benavoli-Zaffalon property of near vacuous priors that allow for learning from data.

This paper concentrates on the linear aspects of the space of measures, and does not further explore the metric aspects of the geometry implied by the Kullback-Leibler information measure. These topics will be examined in future papers.

Acknowledgments

The author is grateful to several anonymous referees for their detailed comments that helped to improve the paper. This research has been supported by grants from the Natural Science and Engineering Research Council of Canada and from the Office of Vice-President Research at the University of Saskatchewan. Part of this research was done while the author was the Alan Richards Mathematics Fellow at Grey College, Durham University.

References

[1] Shun-ichi Amari and Hiroshi Nagaoka, *Methods of Information Geometry*, American Mathematical Society, 2000

[2] Osama Bataineh, *Imprecise Probability Models for Logistic Regression.* PhD Thesis, University of Saskatchewan, 2012.

[3] Alessio Benavoli and Marco Zaffalon, A model of prior ignorance for inferences in the one-parameter exponential family, *Journal of Statistical Planning and Inference*, 2012, 1960–1979.

[4] M.G. Bickis, The imprecise logit-normal model and its application to estimating hazard functions, *Journal of Statistical Theory and Practice* **3** (2009), 183–195.

[5] P. Diaconis and D. Ylvisaker, Conjugate priors for exponential families. *Ann. Statist.* **7** (1979), 269–281.

[6] Bruno de Finetti, *Theory of probability*, Wiley, New York, 1974.

[7] Solomon Kullback, *Information Theory and Statistics*, Wiley,1959.

[8] Chel Hee Lee, *Imprecise Prior for Imprecise Inference on Poisson Sampling Model.* PhD Thesis, University of Saskatchewan, 2014.

[9] Isaac Levi, *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Change*, MIT Press, 1980.

[10] Deborah Mayo, *Error and the Growth of Experimental Knowledge*, University of Chicago Press, 1996.

[11] Erik Quaeghebeur, *Learning from samples using coherent lower previsions.* PhD thesis, University of Ghent, 2009.

[12] Erik Quaeghebeur and Gert de Cooman, *Imprecise probability models for inference in exponential families*, ISIPTA '05: Proc. 4th Int. Symp. on Imprecise Probabilities and Their Applications (Fabio G. Cozman, Robert Nau, and Teddy Seidenfeld, eds.), July 2005, pp. 287–296.

[13] Matthias C. M. Troffaes and Gert de Cooman, *Lower Previsions*, Wiley, 2014.

[14] Peter Walley, *Statistical reasoning with imprecise probabilities*, Chapman and Hall, London, 1991.

[15] Peter Walley, *Inferences from multinomial data: Learning about a bag of marbles*, Journal of the Royal Statistical Society, Series B **58** (1996), no. 1, 3–34.