ARTICLE IN PRESS
IJA:7983

International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Towards a geometry of imprecise inference

Miķelis Bickis

Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Road, Saskatoon SK S7J 2T7, Canada

A R T I C L E   I N F O

Article history:
Received 1 December 2015
Received in revised form 23 October 2016
Accepted 23 November 2016
Available online xxxx

Keywords:
Information geometry
Exponential family
Sets of measures

A B S T R A C T

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 normalising 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 normalising 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.

1. Introduction

Statistical inference deals with observations that are realisations 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 realised 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 criticised as being an improper method for scientific investigation (see Mayo [15] 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.

E-mail address: bickis@snoopy.usask.ca.

http://dx.doi.org/10.1016/j.ijar.2016.11.015

Please cite this article in press as: M. Bickis, Towards a geometry of imprecise inference, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2016.11.015

These criticisms were addressed in Walley's fundamental treatise [20]. 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 [9]), is essentially an expectation.[1] 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 knowledge, 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 [19].

Walley's lower envelope theorem [20, 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 [14]) 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.[2]

Inference can now be based on such imprecise prior probabilities. Walley proposed a *generalised 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 [21] 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 [7] 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 [17], Quaeghebeur [16], Bickis [6], Benavoli and Zaffalon [4], Bataineh [3], and Lee [13]. 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 [4].

It is the aim of this paper to demonstrate how both classical Bayesian learning as well as its imprecise analogue can be described in geometric terms. The author's hope is that this approach to visualisation will provide insight about the nature of statistical inference with imprecise probabilities and lead to more transparency about some of its surprises and paradoxes. We will 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 parametrisations, Bayes' rule can be seen as a data-dependent translation of a point representing the prior distribution. The generalised Bayes rule can similarly be seen as a translation of an entire set. We can thus visualise how various choices of prior set affect the process of learning from data. The theory will be illustrated with a number of instructive examples.

2. Geometry and statistics

The question may be naturally raised of why one would be interested in geometrical ideas for inference. Geometrical ideas have a long history in mathematics that transcend their original purpose of "measuring the earth". The human mind responds well to visual ideas, and being able to visualise mathematical concepts enhances the understanding of abstraction and provides insight for truths that may escape a purely analytical approach.

Geometrical approaches to statistics also have a long history. The theory of linear models has benefited from its presentation as projections in Euclidean space. A similar approach to more general models has required the tools of differential geometry which was to a great extent developed for the visualisation of problems in mathematical physics. The later part of the twentieth century saw the development of geometric ideas applied to the theory of inference [10,1] and today what has come be known as *information geometry* is a well-established discipline [11].

Geometric visualisation has had less penetration in the literature using the Bayesian paradigm, although de Finetti [8] has used geometric arguments to justify Bayes rule.

Distance is the central concept of geometry. When using geometric analogies to discuss anything other than physical space, it is important that one define a meaningful measure of distance which is not merely an artefact of an arbitrary representation. A probability measure can be viewed as a linear functional on random variables, and hence a set of probability

[1] More precisely, the concept of prevision is more general than that of expectation, as a prevision is not required to satisfy the monotone convergence theorem.

[2] Again, this is an oversimplification, as the credal set under Walley's treatment may include measures that are only finitely additive.

Please cite this article in press as: M. Bickis, Towards a geometry of imprecise inference, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2016.11.015