Open Science Framework DOI:10.31219/osf.io/***

Foundations of inference under symmetry:
A derivation of algorithms for non-parametric density inference

P.G.L. Porta Mana
<pgl@portamana.org>

15 January 2023; updated 24 February 2023 [draft]

***

0 A warning on notation

The probability calculus, being a generalization of the propositional truth-calculus, concerns propositions or statements. In the expression "P($A|B$)", A and B thus stand for statements or compositions of statements. In scientific applications most statements of interest are of the form "Measurement of the quantity X yields outcome x" or "The quantity X is set to the value x", implicitly or explicitly accompanied by other statements expressing contextual information, for instance "The measurement is made at time … in laboratory … under conditions …".

Writing full statements such as these within probability formulae would take an impractical amount of space. This impracticality is solved by an abuse of notation: outside a probability formula, a symbol like X may denote a quantity; but within a probability formula, say $\mathrm{p}(X \mid \dots)$, it instead denotes a full statement such as "Measurement of the quantity X yields outcome x"; sometimes the latter is written "$X = x$" when the value x is not generic or needs to be explicit.

To limit the confusion that can arise from this notation abuse, I use a convention similar to Jaynes's (2003 § 2.5 p. 43): the probability symbol "P" is only used when the symbols in its arguments univocally denote statements; the symbol "p" is used instead to warn that some of the symbols in its arguments are abused. For example, if the quantity N denotes the number of people satisfying some condition, A := "Measurement of N yields 101 people", and B denotes some contextual statement, then these three expressions have the same meaning:

$$\mathrm{P}(A \mid B) = 0.3\,, \qquad \mathrm{p}(N \mid B) = 0.3\,, \qquad \mathrm{p}(N = 101 \mid B) = 0.3\,.$$

We take the logical connectives ¬ (not), ∨ (or), ∧ (and) as a functionally complete (redundant) base set. A comma "," is interchangeably used for

$\wedge$ in probability formulae. The probability-calculus rules for the base connectives are

$$\begin{gathered}\mathrm{P}(\neg A \mid H) + \mathrm{P}(A \mid H) = 1 \\ \mathrm{P}(A \wedge B \mid H) - \mathrm{P}(A \mid B \wedge H)\,\mathrm{P}(B \mid H) = 0 \\ \mathrm{P}(A \vee B \mid H) + \mathrm{P}(A \wedge B \mid H) - \mathrm{P}(A \mid H) - \mathrm{P}(B \mid H) = 0 \\ \text{for any atomic or composite statements } A, B, H.\end{gathered} \tag{0}$$

The rules ca be used in different guises, for example

$$\frac{\mathrm{P}(A|H)=x}{\mathrm{P}(\neg A|H)=1-x} \qquad \frac{\mathrm{P}(\neg A|H)=x}{\mathrm{P}(A|H)=1-x}$$

$$\frac{\mathrm{P}(A|B\wedge H)=x \quad \mathrm{P}(B|H)=y}{\mathrm{P}(A\wedge B|H)=xy} \qquad \frac{\mathrm{P}(A|B\wedge H)=x \quad \mathrm{P}(A\wedge B|H)=y}{\mathrm{P}(B|H)=y/x} \qquad \frac{\mathrm{P}(B|H)=x \quad \mathrm{P}(A\wedge B|H)=y}{\mathrm{P}(A|B\wedge H)=y/x}$$

$$\frac{\mathrm{P}(A|H)=x \quad \mathrm{P}(B|H)=y \quad \mathrm{P}(A\wedge B|H)=z}{\mathrm{P}(A\vee B|H)=x+y-z} \qquad \frac{\mathrm{P}(A|H)=x \quad \mathrm{P}(B|H)=y \quad \mathrm{P}(A\vee B|H)=z}{\mathrm{P}(A\wedge B|H)=x+y-z} \qquad \frac{\mathrm{P}(A|H)=x \quad \mathrm{P}(A\wedge B|H)=y \quad \mathrm{P}(A\vee B|H)=z}{\mathrm{P}(B|H)=y+z-x}$$

Compare with the inference rules for sequent calculi.[1]

1 Inference problem

The general context of our inference problem is the following. There is a collection of "units", each of which has an associated set of measurable quantities $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, which we call "variates" according to statistics terminology. The collection of variates $X_1, X_2, \ldots$ is denoted X, and similarly for Y. Examples: These "units" could be mechanical gadgets coming out of a production line; the X variates, some of their mechanical properties; the Y variates, other hidden properties or future events such as mechanical failure. Or the units could be clinical patients; the X variates, outcomes of clinical tests performed on these patients; the Y variates, medical conditions or future events such as life length or occurrence of some disease. Our inference context is thus quite general indeed. The units can be labelled with some index i, but the values of this index and their ordering have no informational relevance. The variate X_1 specific to unit i is denoted by X_1^i, and similarly for the other variates.

Our general goal is to make inferences about some of the variates for some of the units, given knowledge about other variates units.

One specific and frequently occurring inference is about the values of the variates Y^0 for some unit, call it $i = 0$, given:

[1] e.g. Huth & Ryan 2004 ch. 1; Prawitz 1965 App. A; see also Boričić 2020.

- the values of the variates X^0 for the same unit;
- the values of the X, Y variates for other units $i = 1, 2, \ldots$;
- some auxiliary information I^0 for the unit 0;
- all remaining relevant information, denoted I.

The most general inference gives probabilities to the possible values of Y^0. Completely certain inference (deduction) is included as a special case with 0 or 1 probabilities.

Thus in mathematical notation we want to determine the distribution of probability

$$\mathrm{p}(Y^0 \mid X^0,\ Y^1, X^1, Y^2, X^2, \ldots,\ I^0, I) \tag{1}$$

over the possible values of Y^0 (more precisely, over statements such as "Measurement of Y for unit 0 yields y").

Variations of this specific inference are also of interest. For instance, some of the X or Y variate values maybe be unknown for some units; or we want to infer about more than one unit; and similar variations. In the following we show the steps to specifically calculate the probabilities (1); variations of this calculation will be discussed later.

2 Initial probabilities

The propositional truth-calculus allows us to calculate the truth-values of non-tautological statements only if we give the truth-values of some other, related, non-tautological statements, usually called "premises". Likewise, the probability calculus allows us to calculate probability-values of non-tautological statements only if we give the probability-values of some other, related, non-tautological statements. These are usually called "prior" or "initial" probabilities (these adjectives do *not* imply that these probabilities concern statements about chronologically earlier events).

Therefore we must first assign the probability values of some statements which can lead us to the probabilities (1) through the rules (0).

There are several prior-probability assignments appropriate to our inference problem. All of them express observations of informational

symmetry. Their analysis goes under the name of "exchangeability"[2]. Let us start with the most important.

Consider the set of units $i = 1, 2, \ldots$, excluding $i = 0$ According to our information I, we have no reason a priori to assign special probabilities to particular units, before we know their variate values $Y^1, X^1, Y^2, X^2, \ldots$. This is reflected in the irrelevance of any possible permutation of their indices. This informational symmetry is expressed by the equality of all joint probabilities for sets of values that can be obtained from one another by permutations of the index i:

$$\begin{aligned}&\mathrm{p}(Y^1 = y^1, X^1 = x^1,\ Y^2 = y^2, X^2 = x^2,\ Y^3 = y^3, X^3 = x^3, \ldots \mid I) =\\ &\mathrm{p}(Y^1 = y^{r(1)}, X^1 = x^{r(1)},\ Y^2 = y^{r(2)}, X^2 = x^{r(2)},\ Y^3 = y^{r(3)}, X^3 = x^{r(3)}, \ldots \mid I)\\ &\qquad\qquad \text{for all permutations } r \text{ and all allowed values } y, x. \qquad (2)\end{aligned}$$

If the number of units is large enough, de Finetti's theorem[2] proves that such a joint probability distribution must then have the following mathematical form:

$$\begin{aligned}&\mathrm{p}(Y^1 = y^1, X^1 = x^1,\ Y^2 = y^2, X^2 = x^2, \ldots \mid I) =\\ &\qquad\qquad \int F(y^1, x^1)\, F(y^2, x^2) \cdots \mathrm{p}(\mathrm{d}F \mid I) \qquad (3)\end{aligned}$$

the integral being over all normalized, positive weight distributions $F(y, x)$ over the variables y, x. The term $\mathrm{p}(\mathrm{d}F \mid I)$ is a positive measure over such distributions.

This mathematical representation has an intuitive interpretation. The weight distribution $F(y, x)$ can be interpreted as the frequency distribution of the values y, x among all units. If we knew this frequency distribution, then by symmetry we would assign the probability $F(y, x)$ to measuring y, x on an unsystematically chosen unit, exactly as in sampling from an urn[3]. Observations on further units would have probabilities according to "sampling without replacement"; however, if the total number of units is large and the observed units are few, these probabilities

[2] For a detailed account with references see Bernardo & Smith 2000; for a shorter summary, Dawid 2013; important references and variations are in Johnson 1924 (Appendix on eduction); 1932 §§ 4.2–4.3; de Finetti 1930; 1937; Hewitt & Savage 1955; Ericson 1969a; Heath & Sudderth 1976; Diaconis 1977; Diaconis & Freedman 1980a,b; Lindley & Novick 1981; Lindley 2014 especially around §§ 7.3, 8.6. [3] Jaynes 2003 ch. 3.

would approximately equal those for "sampling with replacement"[4]. These probabilities correspond to the term $F(y^1, x^1)\, F(y^2, x^2) \cdots$ in the integral. If the frequency distribution F is unknown and we assign a probability density $\mathrm{p}(\mathrm{d}F \mid I)$ to it, then the resulting probability for the observations is a convex mixture of the sampling-with-replacement probabilities, weighted by this probability density. This last step follows from the law of total probability, a consequence of the rules (0). The integral above corresponds to this convex mixture.

The symmetry or "exchangeability" requirement does not determine specific numeric values for the joint probability (3), but greatly restrict our freedom in assigning such values. The only freedom we have left is in the specification of the density $\mathrm{p}(\mathrm{d}F \mid I)$. This specification will depend on the particular inference problem and on the nature of the units and variates. The intuitive interpretation of $\mathrm{p}(\mathrm{d}F \mid I)$ as our uncertainty about the "long-run frequencies" can greatly help in this specification. De Finetti's theorem has moreover an important consequence in this respect: as the values of more and more units become available, the initial numerical choice of density $\mathrm{p}(\mathrm{d}F \mid I)$ affects further inferences less and less. Let us discuss these further inferences.

Once the joint probability distribution (3) has been numerically assigned, through specification of $\mathrm{p}(\mathrm{d}F \mid I)$, all other marginal and conditional distributions concerning units and variates are also numerically fully determined by repeated application of the rules (0). Some examples:

$$\mathrm{p}(Y^2 = y^2 \mid I) = \int \sum_x F(y^2, x)\, \mathrm{p}(\mathrm{d}F \mid I)\,, \tag{4a}$$

$$\mathrm{p}(Y^2 = y^2 \mid X^2 = x^2, I) = \frac{\int F(y^2, x^2)\, \mathrm{p}(\mathrm{d}F \mid I)}{\int \sum_x F(y^2, x)\, \mathrm{p}(\mathrm{d}F \mid I)}\,, \tag{4b}$$

$$\mathrm{p}(Y^2 = y^2 \mid X^1 = x^1, I) = \frac{\int \sum_x F(y^2, x) \sum_y F(y, x^1)\, \mathrm{p}(\mathrm{d}F \mid I)}{\int \sum_y F(y, x^1)\, \mathrm{p}(\mathrm{d}F \mid I)}\,. \tag{4c}$$

3 Inferences about unit 0

Any inferences about any variates of any units for which the joint probability distribution (3) is defined can now be probabilistically answered.

[4] Heath & Sudderth 1976; for error bounds see Diaconis & Freedman 1980a.

Let us consider inferences about a unit, with index $i = 0$, about which we have some additional information I^0.

3.1 Full symmetry

If I^0 amounts to saying that the symmetry underlying the probabilities (3) also includes unit 0, then any probabilities are uniquely determined, as in the examples (4). For instance, our main probability (1) is given by

$$\mathrm{p}(Y^0 = y^0 \mid X^0 = x^0,\ Y^1 = y^1, X^1 = x^1, \dots,\ I^0, I) = \frac{\int F(y^0, x^0)\, F(y^1, x^1) \cdots \mathrm{p}(\mathrm{d}F \mid I)}{\int \sum_y F(y, x^0)\, F(y^1, x^1) \cdots \mathrm{p}(\mathrm{d}F \mid I)} \,. \tag{5}$$

3.2 Symmetry conditional on X variates

A more interesting kind of additional information I^0_{x} amounts to saying that there is an informational symmetry involving unit 0 an the other units, but only with regard to the Y variates, and only with units sharing the same values x^* of the X variates. Using a different terminology, this would be a symmetry restricted to a subpopulation of units which is defined by particular X values. Mathematically,

$$\begin{aligned}&\mathrm{p}(Y^0 = y^0, Y^1 = y^1, Y^2 = y^2 \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{x}}, I) = \\ &\mathrm{p}(Y^0 = y^{r(0)}, Y^1 = y^{r(1)}, Y^2 = y^{r(2)} \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{x}}, I) \\ &\qquad\qquad \text{for all permutations } r \text{ and all allowed values } y, x. \end{aligned} \tag{6}$$

If the number of units is large enough, a variant of de Finetti's theorem[5] proves that such a joint probability distribution must then have the following mathematical form:

$$\mathrm{p}(Y^0 = y^0, Y^1 = y^1, Y^2 = y^2 \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{x}}, I) = \int \frac{F(y^0, x^*)}{\sum_y F(y, x^*)} \frac{F(y^1, x^*)}{\sum_y F(y, x^*)} \frac{F(y^2, x^*)}{\sum_y F(y, x^*)} \cdots \mathrm{p}(\mathrm{d}F \mid I) \,. \tag{7a}$$

This representation has an intuitive interpretation similar to (3); the difference is that a term like $\frac{F(y^0, x^*)}{\sum_y F(y, x^*)}$ now corresponds to the long-run

[5] Bernardo & Smith 2000 § 4.6; their presentation is in terms of the integral (7b).

conditional frequency $F(y^0 \mid x^*)$. Our conditional probabilities would be equal to these conditional frequencies if they were known. The uncertainty about the latter leads again to a convex mixture of these conditional probabilities, from the rules (0). With a change of variables we could represent the joint frequency $F(y, x)$ as the pair of conditional and marginal frequencies $\big(F_{Y|X}(y \mid x), F_X(x)\big)$. The formula above then simplifies, by marginalization over $\mathrm{d}F_X$, to

$$\mathrm{p}(Y^0 = y^0, Y^1 = y^1, Y^2 = y^2 \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{x}}, I) = \int F(y^0 \mid x^*)\, F(y^1 \mid x^*)\, F(y^2 \mid x^*) \cdots \mathrm{p}(\mathrm{d}F_{Y|X} \mid I)\,. \quad (7b)$$

Under this kind of auxiliary information I^0_{x}, our main probability (1) is given by

$$\mathrm{p}(Y^0 = y^0 \mid X^0 = x^0,\ Y^1 = y^1, X^1 = x^1, \dots,\ I^0_{\mathrm{x}}, I) = \frac{\int \frac{F(y^0, x^0)}{\sum_y F(y, x^0)}\, F(y^1, x^1) \cdots \mathrm{p}(\mathrm{d}F \mid I)}{\int F(y^1, x^1) \cdots \mathrm{p}(\mathrm{d}F \mid I)}\,. \quad (8)$$

3.3 Symmetry conditional on Y variates

Yet another kind of auxiliary information, I^0_{y}, is the one similar to I^0_{x} but with the variates X and Y swapped; that is, symmetry under permutation of X values for a subpopulation sharing the same Y values. The discussion proceeds analogously to the previous section up to the representation

$$\mathrm{p}(X^0 = x^0, X^1 = x^1, X^2 = x^2 \mid Y^0 = y^*, Y^1 = y^*, Y^2 = y^*, \dots,\ I^0_{\mathrm{y}}, I) = \int \frac{F(x^0, y^*)}{\sum_x F(x, y^*)}\, \frac{F(x^1, y^*)}{\sum_x F(x, y^*)}\, \frac{F(x^2, y^*)}{\sum_x F(x, y^*)} \cdots \mathrm{p}(\mathrm{d}F \mid I)\,. \quad (9)$$

In this case, however, the main probability (1) is *not* determined by the probability rules once $\mathrm{p}(\mathrm{d}F \mid I)$ has been assigned. We arrive at the following relation, where the proportionality constant can be found by

normalizing over Y^0:

$$\mathrm{p}(Y^0 = y^0 \mid X^0 = x^0,\ Y^1 = y^1, X^1 = x^1, \dots,\ I^0_{\mathrm{y}}, I) \propto$$
$$\mathrm{p}(Y^0 = y^0 \mid Y^1 = y^1, X^1 = x^1, \dots,\ I^0_{\mathrm{y}}, I) \times \frac{\int \frac{F(x^0,y^0)}{\sum_x F(x,y^0)}\, F(x^1,y^1)\cdots \mathrm{p}(\mathrm{d}F \mid I)}{\int F(x^1,y^1)\cdots \mathrm{p}(\mathrm{d}F \mid I)}\,. \tag{10}$$

The probability in the second row, $\mathrm{p}(Y^0 = y^0 \mid Y^1 = y^1, X^1 = x^1, \dots,\ I^0_{\mathrm{y}}, I)$, must still be assigned[6]. This assignment depends strongly on the circumstances. In clinical applications, for example, it is typically further assumed that the units $i = 1, 2, \dots$ are irrelevant for this probability:

$$\mathrm{p}(Y^0 = y^0 \mid Y^1 = y^1, X^1 = x^1, \dots,\ I^0_{\mathrm{y}}, I) = \mathrm{p}(Y^0 = y^0 \mid I^0_{\mathrm{y}}, I) \tag{11}$$

and $\mathrm{p}(Y^0 = y^0 \mid I^0_{\mathrm{y}}, I)$ is taken to be equal to a *base rate*[7], through exchangeability arguments analogous to those of § 3 but without X variates and referred to a *different* sets of units having Y variates.

3.4 Other symmetries

It is possible to consider also symmetries for only some of the X variates or some of the Y variates or both, with analyses similar to those of the previous subsections. Finally, it is also possible to consider situations with no symmetries; then the theorems based on exchangeability are not available and one have to assign initial probabilities by different arguments.

3.5 Common mathematical term

In all examples of additional information I^0 considered above, the resulting expression for our main probability, eqs (5), (8), (10), involve an integral containing a common term: the density

$$\mathrm{p}(\mathrm{d}F \mid D, I) \coloneqq \frac{F(x^1,y^1)\,F(x^2,y^2)\cdots}{\int F(x^1,y^1)\,F(x^2,y^2)\cdots \mathrm{p}(\mathrm{d}F \mid I)}\,, \tag{12}$$

[6] Lindley & Novick 1981. [7] Bar-Hillel 1980; Jenny et al. 2018; Sprenger & Weinberger 2021; Matthews 1996.

where D is the conjunction of all statements "$Y^i = y^i$", "$X^i = x^i$" for $i = 1, 2, \dots$. Indeed this probability density can be considered as the updated probability for $\mathrm{d}F$, given the information about all units, D.

Our next task is to calculate this term.

4 A specific representation of F and its densities

4.1 Mathematical representation of F

In order to specify the probability densities $\mathrm{p}(\mathrm{d}F \mid I)$ and $\mathrm{p}(\mathrm{d}F \mid D, I)$ we need a mathematical representation of the possible joint distributions $F(y, x)$ and of their space; from a differential-geometric point of view[8], we want a coordinate system on the manifold of distributions $\{F\}$. It would be an advantage if this mathematical representation would also let us easily compute conditional and marginal distributions such as $F(y \mid x)$ and $F(x)$, as we need these to calculate our main probability from formulae (8) or (11).

A very clever representation is discussed by Dunson & Bhattacharya (2011)[9]; it is an evolution of representations based on Dirichlet-process mixtures[10]. It expresses a particular $F(y, x)$ as an in-principle infinite convex mixture of products of parametric distributions $K(\dots \mid \dots)$:

$$F(y_1, y_2, \dots, x_1, x_2, \dots) = \sum_c w^c \, K(y_1 \mid \upsilon_1^c) \, K(y_2 \mid \upsilon_1^c) \cdots K(x_1 \mid \xi_1^c) \, K(x_2 \mid \xi_1^c) \cdots \qquad (13)$$

where $w \coloneqq (w^c)$ are normalized weights, and $\upsilon \coloneqq (\upsilon_j^c)$ and $\xi \coloneqq (\xi_j^c)$ are parameters, possibly multidimensional. The functional forms of the distributions $K(y \mid \dots)$, $K(x \mid \dots)$ depend on the nature of the variates y, x. For example K could be a Gaussian distribution for a continuous unbounded variate. The choice of the K depends on the properties expected in the distributions F and on computational requirements.

Any F is thus represented by an in-principle infinite tuple of variables:

$$F \mathrel{\hat{=}} (w, \upsilon, \xi) \equiv (w^c, \upsilon_j^c, \xi_j^c) \,. \qquad (14)$$

The relation between the manifolds $\{F\}$ and $\{(w, \upsilon, \xi)\}$ is typically not one-to-one: different tuples of variables can correspond to the same F.

[8] Choquet-Bruhat et al. 1996 III.A, VII.A. [9] see also Bhattacharya & Dunson 2012; Rossi 2014. [10] e.g. Antoniak 1974; Ferguson 1983; Escobar & West 1995; Müller et al. 1996; Rasmussen 1999.

We can interpret this as a non-regular embedding $\{F\} \to \{(w, \upsilon, \xi)\}$; or as the existence of different coordinate charts on $\{F\}$, with overlapping domains; or as the expression of F through a frame[11], rather than through a set of basis functions . This is not a problem for us, because a probability density on the manifold $\{(w, \upsilon, \xi)\}$ also defines, by pull-back, a probability density on the manifold $\{F\}$.

The representation (13) is extremely convenient because it leads to simple expressions for conditional and marginal frequencies. For instance in can be easily calculated that

$$F(y_1 \mid x_2) = \sum_c \frac{w^c \; K(x_2 \mid \xi_1^c)}{\sum_{c'} w^{c'} \; K(x_2 \mid \xi_1^{c'})} \; K(y_1 \mid \upsilon_1^c) \,, \tag{15}$$

$$F(y_1) = \sum_c w^c \; K(y_1 \mid \upsilon_1^c) \,. \tag{16}$$

Although the sum in the representation (13) is in principle infinite, a particular choice of density for the (w, υ, ξ) variables allows us to approximate it with a finite number of terms, as shown by Ishwaran & Zarepour (2002). This possibility is essential for computational purposes.

A probability density $\mathrm{p}(\mathrm{d}F)$ is finally specified by specifying a density $\mathrm{p}(\mathrm{d}w, \mathrm{d}\upsilon, \mathrm{d}\xi)$. This specification is determined partly by the information underlying the density, partly by computational demands.

These densities can be calculated by a variety of Monte Carlo methods. We discuss the computation in detail in §***.

4.2 Choice of parametric distributions K

The choice of parametric distribution K for each variate depends on the mathematical type of variate, on the kind of long-run distribution we expect for it, and on computational convenience. Here are some specific choices. What is said about a variate x and parameters ξ also holds for any variate y and parameters υ. In the following, Φ is the function

$$\Phi \colon \{-\infty\} \cup \mathbf{R} \cup \{+\infty\} \to [0,1], \qquad \Phi(z) \coloneqq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\!\left(-\frac{t^2}{2}\right) \mathrm{d}t \,, \tag{17}$$

and $\Phi^{-1} \colon [0,1] \to \{-\infty\} \cup \mathbf{R} \cup \{+\infty\}$ its inverse ("probit" function).

[11] Balan et al. 2006; Heil 2011 ch. 8; Daubechies 1999 ch. 3.

Binary variates For a binary variate $x \in \{0, 1\}$ we choose a Bernoulli distribution with weight parameter $\xi = q$:

$$x \in \{0, 1\} :$$
$$K(x \mid q) \coloneqq x\, q + (1 - x)\,(1 - q) \equiv \begin{cases} q & \text{if } x = 1 \\ 1 - q & \text{if } x = 0 \end{cases} \tag{18}$$

Obviously any other binary values $\{a, b\}$ can be mapped into $\{0, 1\}$.

Categorical variates For a categorical variate with n possible values, $x \in \{a_1, \dots, a_n\}$, we choose a categorical distribution with normalized weight parameters $q \coloneqq (q_1, \dots, q_n)$:

$$x \in \{a_1, \dots, a_n\} :$$
$$K(x \mid q) \coloneqq \sum_{i=1}^{n} q_i\, \delta(x = a_i) \equiv \begin{cases} q_1 & \text{if } x = a_1 \\ \dots & \\ q_n & \text{if } x = a_n \end{cases} \tag{19}$$

Obviously the values $\{a_1, \dots, a_n\}$ can be set in one-one correspondence with $\{1, \dots, n\}$.

Integer or discrete ordinal variates By "discrete ordinal" variate it is meant a variate x assuming a discrete number m of values which have a natural ordering, for example integers $x \in \{1, \dots, m\}$. The fact that they have a natural ordering has implications for smoothness, in a loose sense of this word. That is, we expect the probabilities for two consecutive values not to have excessive jumps. This is the difference between a discrete-ordinal variate and the categorical one discussed above.

For such a variate we can use a "discretized-Gaussian" distribution with mean and variance parameters (μ, σ^2):

$$x \in \{1, \dots, m\} :$$
$$\begin{aligned} K(x \mid \mu, \sigma^2) &\coloneqq \Phi\left[\frac{\Phi^{-1}\left(\frac{x}{m}\right) - \mu}{\sigma}\right] - \Phi\left[\frac{\Phi^{-1}\left(\frac{x-1}{m}\right) - \mu}{\sigma}\right] \\ &\equiv \frac{1}{\sqrt{2\pi\,\sigma^2}} \int_{\Phi^{-1}\left(\frac{x-1}{m}\right)}^{\Phi^{-1}\left(\frac{x}{m}\right)} \exp\left[-\frac{(z-\mu)^2}{2\,\sigma^2}\right] \mathrm{d}z\,. \end{aligned} \tag{20}$$

The discretization points are chosen so that for $\mu = 0$, $\sigma = 1$ the distribution K is uniform with constant value $1/m$.

Discrete ordinal variates with other domains can be mapped onto $\{1, \dots, m\}$.

Continuous real variates For a continuous real variate $x \in \mathbf{R}$ we choose a Gaussian distribution with mean and variance as the parameters $\xi = (\mu, \sigma^2)$:

$$x \in \mathbf{R}:$$
$$K(\mathrm{d}x \mid \mu, \sigma^2) \coloneqq \Phi'\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}x \equiv \frac{1}{\sqrt{2\pi\,\sigma^2}}\,\exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right]\mathrm{d}x\,. \tag{21}$$

Censored continuous variates A "censored continuous" variate is meant here, more generally, to be a continuous variate with values in a *closed* or semi-closed bounded set, for example $x \in [a, b]$ or $x \in [a, +\infty[$. Owing to the topology of its domain, such variate cannot be mapped onto a continuous unbounded one, previously discussed.

The boundary value or values that such a variable can take often have a singular probability mass (delta distribution), which may arise from the fact that x is censored, that is, it could in principle assume continuous values outside the boundaries, but such values are truncated to the boundary ones.

For such a variate we can use a Gaussian distribution with mean and variance parameters (μ, σ^2), where the $x \leqslant a$ and $x \geqslant b$ tails represent the probability mass given to the boundary points:

$$x \in [a, b]:$$
$$K(\mathrm{d}x \mid \mu, \sigma^2) \coloneqq \begin{cases} \Phi\left(\dfrac{a-\mu}{\sigma}\right) & \text{if } x = a \\ \Phi'\left(\dfrac{x-\mu}{\sigma}\right)\mathrm{d}x & \text{if } a < x < b \\ \Phi\left(-\dfrac{b-\mu}{\sigma}\right) & \text{if } x = b \end{cases} \tag{22}$$

The semi-bounded case $x \in [a, +\infty[$ can be obtained with obvious adjustments. The continuous non-bounded case previously discussed can be seen as a limit case of the present one for $a \to -\infty$, $b \to +\infty$.

Other continuous variates Other continuous variates with open or semi-open connected domains in the real line can be transformed into the

continuous real or censored ones discussed last. For example a variate $z \in]0, +\infty[$ can be transformed into $x = \ln z$, a variate $z \in]a, b[$ into $x = \Phi^{-1}\left(\frac{z-a}{b-a}\right)$, and a variate $z \in [0, +\infty[$ into $x = \ln(z + q)$ for some $q > 0$ to be treated as a censoring value. These kinds of transformation are often suggested by the nature of the original variate.

4.3 Density for F

Once the density F is represented by the variables (w, υ, ξ) through eq. (13), any density $\mathrm{p}(\mathrm{d}w, \mathrm{d}\upsilon, \mathrm{d}\xi \mid I)$ for the latter variables also determines (possibly by pull-back) a density $\mathrm{p}(\mathrm{d}F \mid I)$ for F.

The choice of this density is equivalent, by de Finetti's theorem (3), to a choice of exchangeable joint distribution $\mathrm{p}(Y, X \mid I)$. This choice is thus essentially dependent on the meaning of the variates Y, X and on the specific inference context.

A general-purpose specific mathematical choice is presented here. It can be useful as a starting point, and it can be reasonable as a definitive choice if we have a large number of units with known X, Y. This choice has two main motivations:

- represent a state of uncertainty that is "diffuse" but still realistic,
- possibility of allowing fast Markov-chain Monte Carlo computations using Gibbs sampling with conditionally conjugate distributions[12].

We consider a density that factorizes as follows:

$$\mathrm{p}(\mathrm{d}w, \mathrm{d}\upsilon, \mathrm{d}\xi \mid I) = \mathrm{p}(\mathrm{d}w \mid I) \prod_{c,j} \left[\mathrm{p}(\mathrm{d}\upsilon_j^c \mid I)\, \mathrm{p}(\mathrm{d}\xi_j^c \mid I)\right] . \tag{23}$$

Furthermore, the densities $\mathrm{p}(\mathrm{d}\xi_j^c \mid I)$ are taken to be identical for each component c, but potentially different for different variates j:

$$\mathrm{p}(\mathrm{d}\xi_j^c \mid I) = \mathrm{p}(\mathrm{d}\xi_j^1 \mid I) , \quad \forall c ; \tag{24}$$

and likewise for $\mathrm{p}(\mathrm{d}\upsilon_j^c \mid I)$.

In the following the specific factor densities are described. The generic parameter ξ is used to include both υ and ξ.

[12] cf. Görür & Rasmussen 2010.

Density for w The density for the weights w is a discrete mixture of Dirichlet distributions with different concentration parameters α:

$$\mathrm{p}(\mathrm{d}w \mid I) = \frac{1}{7} \sum_{\alpha=2^{-3}}^{2^3} (\alpha-1)! \left[\prod_{c=1}^{M} \frac{(w^c)^{\alpha/M-1}}{(\alpha/M-1)!}\right] \delta\Bigl(1-\textstyle\sum_{c=1}^{M} w^c\Bigr)\, \mathrm{d}w \tag{25}$$

with α running through the values $\{2^{-3}, 2^{-2}, 2^{-1}, 1, 2^1, 2^2, 2^3\}$, and where M is the number of components in the convex mixture (13). The possible α values are given equal weights.

This density is a discretized version of a mixture of Dirichlet distributions with a continuous range of concentration parameters[13]. The discretized version seems to lead to the same general predictive properties as the continuous one, but is computationally more convenient.

Density for parameters of Gaussian components The density for the mean μ of components with a Gaussian distribution is a broad Gaussian distribution:

$$\mathrm{p}(\mathrm{d}\mu \mid I) = \Phi'\left(\frac{\mu-\bar{\mu}}{2\,\bar{\sigma}}\right)\mathrm{d}\mu \tag{26}$$

where $\bar{\mu}$ and $\bar{\sigma}$ are location and scale hyperparameters for the variate of interest. The values of these hyperparameters are determined by the nature of the variate; otherwise they can approximately – although slightly inconsistently – be determined by the data, if the latter are numerous. For instance, $\bar{\mu}$ can be taken to be the median of the data, and $\bar{\sigma}$ the interquartile range or the median absolute deviation of the data, possibly rescaled to the standard deviation of a standard Gaussian.

The density for the variance σ^2 is a discrete mixture of beta-prime distributions (or beta distributions of the second kind)[14] with unit shape hyperparameters and different scale hyperparameters τ^2. The beta-prime can in turn be written as a mixture of inverse-gamma distributions[15] with different rate parameters λ:

$$\begin{aligned}\mathrm{p}(\mathrm{d}\sigma^2 \mid I) &= \frac{1}{5} \sum_{l=2^{-2}}^{2^2} (l\,\bar{\sigma})^{-2} \left[1+\left(\frac{\sigma}{l\,\bar{\sigma}}\right)^2\right]^{-2} \mathrm{d}\sigma^2 \\ &\equiv \mathrm{d}\sigma^2\, \frac{1}{5} \sum_{l=2^{-2}}^{2^2} \int_0^\infty \frac{\mathrm{e}^{-\frac{1}{\lambda\,\sigma^2}}}{\lambda\,\sigma^4}\, \frac{\mathrm{e}^{-\frac{1}{(l\,\bar{\sigma})^2\,\lambda}}}{(l\,\bar{\sigma})^2\,\lambda^2}\, \mathrm{d}\lambda\ .\end{aligned} \tag{27}$$

[13] e.g. Rossi 2014 § 2.5; Rasmussen 1999 § 2.1. [14] Johnson et al. 1995 § 25.7, 27.2; McDonald 1984. [15] Dubey 1970.

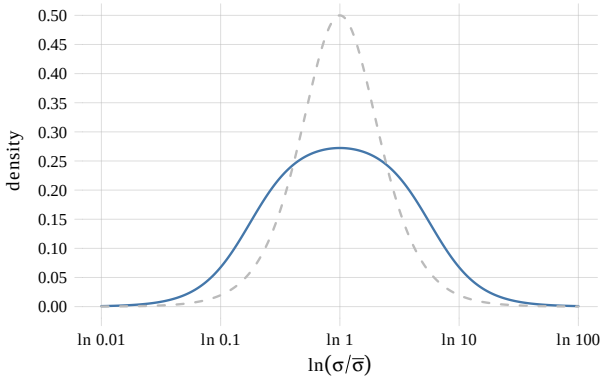

Figure 1 Solid blue: equal mixture of beta-prime (28). Dashed grey: unmixed beta-prime with unit shape hyperparameters

The expression in terms of inverse-gammas is useful for conditionally conjugate Gibbs sampling. The beta-prime distribution has the appealing property of being symmetric with respect to the relative log-scale parameter $\ln(\sigma/\bar{\sigma})$. We can in fact rewrite the density above as

$$\mathrm{p}(\mathrm{d}\ln\sigma \mid I) = \frac{1}{5}\sum_{l=2^{-2}}^{2^2} 2\left(\mathrm{e}^{\ln\frac{\sigma}{l\bar{\sigma}}} + \mathrm{e}^{-\ln\frac{\sigma}{l\bar{\sigma}}}\right)^{-2}\mathrm{d}\ln\sigma \tag{28}$$

The mixture of several scale hyperparameters makes it somewhat broader without excessively fattening the tails, as shown in fig. 1.

The combination of prior densities (25), (26), (27) leads to a distribution $\mathrm{p}(\mathrm{d}F \mid I)$ of marginal densities $F(\mathrm{d}x)$, for a real variate x, which is illustrated in fig. 2.

Bibliography

("de X" is listed under D, "van X" under V, and so on, regardless of national conventions.)

Alvarez-Melis, D., Broderick, T. (2015): *A translation of "The characteristic function of a random phenomenon" by Bruno de Finetti*. arXiv DOI:10.48550/arXiv.1512.01229. Transl. of de Finetti (1929).

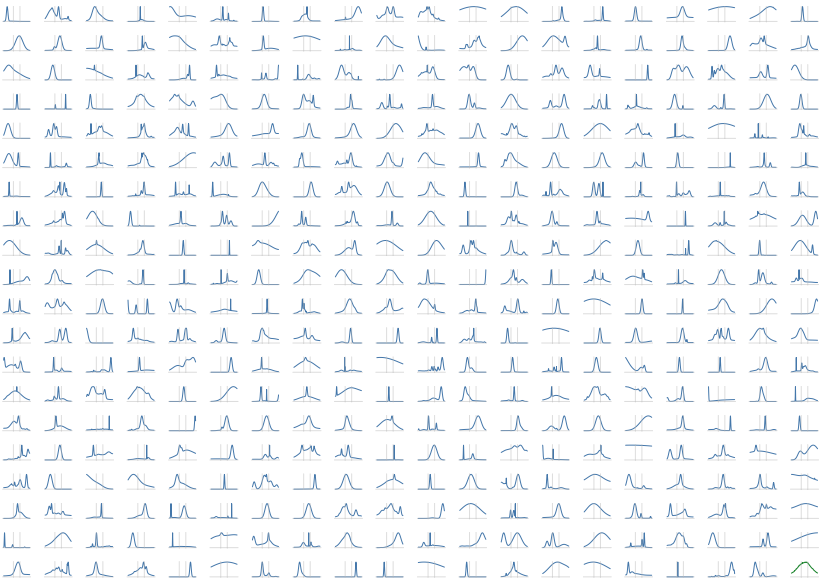

Figure 2 399 samples from the density $\mathrm{p}[\mathrm{d}F \mid I]$ for the marginal density $F(\mathrm{d}x)$ of a continuous real variate x. The vertical grey lines mark the scale $\pm\tilde{\sigma}$. The thicker green density at the bottom-right corner is the expected marginal density $\mathrm{E}[F(\mathrm{d}x) \mid I]$.

Antoniak, C. E. (1974): *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*. Ann. Stat. **2**[6], 1152–1174.

Balan, R., Casazza, P. G., Heil, C., Landau, Z. (2006): *Density, overcompleteness, and localization of frames. I. Theory. II. Gabor systems*. J. Fourier Anal. Appl. **12**[2, 3], 105–143, 307–344. DOI: 10.1007/s00041-006-6022-0, DOI:10.1007/s00041-005-5035-4.

Bar-Hillel, M. (1980): *The base-rate fallacy in probability judgments*. Acta Psychol. **44**[3], 211–233. DOI:10.1016/0001-6918(80)90046-3.

Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M., eds. (2011): *Bayesian Statistics 9*. (Oxford University Press, Oxford). DOI: 10.1093/acprof:oso/9780199694587.001.0001.

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, repr. (Wiley, New York). DOI: 10.1002/9780470316870. First publ. 1994.

Bhattacharya, A., Dunson, D. B. (2012): *Nonparametric Bayes classification and hypothesis testing on manifolds*. J. Multivar. Anal. **111**, 1–19. DOI:10.1016/j.jmva.2012.02.020.

Boričić, M. (2020): *Probabilized sequent calculus and natural deduction system for classical logic*. In: Ognjanović (2020): ch. 7:197–213. DOI:10.1007/978-3-030-52954-3_7.

Choquet-Bruhat, Y., DeWitt-Morette, C., Dillard-Bleick, M. (1996): *Analysis, Manifolds and Physics. Part I: Basics*, rev. ed. (Elsevier, Amsterdam). First publ. 1977.

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford). DOI:10.1093/acprof:oso/9780199695607.001.0001.

Daubechies, I. (1999): *Ten Lectures on Wavelets*, sixth pr. (SIAM, Philadelphia). doi:10.1137/1.9781611970104. First publ. 1992.

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013): ch. 2:19–29. doi:10.1093/acprof:oso/9780199695607.003.0002.

de Finetti, B. (1929): *Funzione caratteristica di un fenomeno aleatorio*. In: *atti del congresso internazionale dei matematici:* ed. by S. Pincherle (Zanichelli, Bologna): 179–190. https://www.mathunion.org/icm/proceedings, http://www.brunodefinetti.it/Opere.htm. Transl. in Alvarez-Melis, Broderick (2015). See also de Finetti (1930).

— (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. **IV**^{5}, 86–133. http://www.brunodefinetti.it/Opere.htm. Summary in de Finetti (1929).

— (1937): *La prévision: ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**^{1}, 1–68. http://www.numdam.org/item/AIHP_1937__7_1_1_0. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

Diaconis, P. (1977): *Finite forms of de Finetti's theorem on exchangeability*. Synthese **36**^{2}, 271–281. doi:10.1007/BF00486116, http://statweb.stanford.edu/~cgates/PERSI/year.html.

Diaconis, P., Freedman, D. (1980a): *Finite exchangeable sequences*. Ann. Probab. **8**^{4}, 745–764. doi:10.1214/aop/1176994663.

— (1980b): *De Finetti's generalizations of exchangeability*. In: Jeffrey (1980): 233–249.

Dubey, S. D. (1970): *Compound gamma, beta and F distributions*. Metrika **16**, 27–31. doi:10.1007/BF02613934.

Dunson, D. B., Bhattacharya, A. (2011): *Nonparametric Bayes regression and classification through mixtures of product kernels*. In: Bernardo, Bayarri, Berger, Dawid, Heckerman, Smith, West (2011): 145–158. doi:10.1093/acprof:oso/9780199694587.003.0005, older version at https://www.researchgate.net/publication/228447342_Nonparametric_Bayes_Regression_and_Classification_Through_Mixtures_of_Product_Kernels.

Ericson, W. A. (1969a): *Subjective Bayesian models in sampling finite populations*. J. R. Stat. Soc. B **31**^{2}, 195–224. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See also discussion in Sampford, Scott, Stone, Lindley, Smith, Kerridge, Godambe, Kish, et al. (1969).

— (1969b): *A note on the posterior mean of a population mean*. J. R. Stat. Soc. B **31**^{2}, 332–334.

Escobar, M. D., West, M. (1995): *Bayesian density estimation and inference using mixtures*. J. Am. Stat. Assoc. **90**^{430}, 577–588. doi:10.1080/01621459.1995.10476550, http://www.stat.duke.edu/~scs/Courses/Stat376/Papers/DirichletProc/EscobarWestJASA1995.pdf.

Ferguson, T. S. (1983): *Bayesian density estimation by mixtures of normal distributions*. In: Rizvi, Rustagi, Siegmund (1983): 287–302.

Görür, D., Rasmussen, C. E. (2010): *Dirichlet process Gaussian mixture models: choice of the base distribution*. J. Comput. Sci. Technol. **25**, 653–664.

Heath, D., Sudderth, W. (1976): *De Finetti's theorem on exchangeable variables*. Am. Stat. **30**^{4}, 188–189.

Heil, C. (2011): *A Basis Theory Primer*, expanded ed. (Birkhäuser). doi:10.1007/978-0-8176-4687-5. First publ. 1998.

Hewitt, E., Savage, L. J. (1955): *Symmetric measures on Cartesian products*. Trans. Am. Math. Soc. **80**^{2}, 470–501. doi:10.1090/S0002-9947-1955-0076206-8.

Huth, M. R. A., Ryan, M. D. (2004): *Logic in Computer Science: Modelling and reasoning about systems*, 2nd ed. (Cambridge University Press, Cambridge). First publ. 2000.

Ishwaran, H., Zarepour, M. (2002): *Dirichlet prior sieves in finite normal mixtures*. Stat. Sinica **12**[3], 941–963. http://www3.stat.sinica.edu.tw/statistica/J12n3/j12n316/j12n316.htm.

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. doi:10.1017/CBO9780511790423, https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffrey, R. C., ed. (1980): *Studies in inductive logic and probability. Vol. II*. (University of California Press, Berkeley).

Jenny, M. A., Keller, N., Gigerenzer, G. (2018): *Assessing minimal medical statistical literacy using the Quick Risk Test: a prospective observational study in Germany*. BMJ Open **8**, e020847, e020847corr2. doi:10.1136/bmjopen-2017-020847, doi:10.1136/bmjopen-2017-020847corr2.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1995): *Continuous Univariate Distributions. Vol. 2*, 2nd ed. (Wiley, New York). First publ. 1970.

Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). https://archive.org/details/logic03johnn.

— (1932): *Probability: the deductive and inductive problems*. Mind **41**[164], 409–423. With some notes and an appendix by R. B. Braithwaite. doi:10.1093/mind/XLI.164.409.

Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.

Lindley, D. V. (2014): *Understanding Uncertainty*, rev. ed. (Wiley, Hoboken, USA). First publ. 2006.

Lindley, D. V., Novick, M. R. (1981): *The role of exchangeability in inference*. Ann. Stat. **9**[1], 45–58. doi:10.1214/aos/1176345331.

Matthews, R. A. J. (1996): *Base-rate errors and rain forecasts*. Nature **382**[6594], 766. doi:10.1038/382766a0.

McDonald, J. B. (1984): *Some generalized functions for the size distribution of income*. Econometrica **52**[3], 647–665. doi:10.2307/1913469.

Müller, P., Erkanli, A., West, M. (1996): *Bayesian curve fitting using multivariate normal mixtures*. Biometrika **83**[1], 67–79. doi:10.1093/biomet/83.1.67.

Ognjanović, Z., ed. (2020): *Probabilistic Extensions of Various Logical Systems*. (Springer, Cham). doi:10.1007/978-3-030-52954-3.

Prawitz, D. (1965): *Natural Deduction: A Proof-Theoretical Study*. (Almqvist & Wiksell, Stockholm).

Rasmussen, C. E. (1999): *The infinite Gaussian mixture model*. Adv. Neural Inf. Process. Syst. (NIPS) **12**, 554–560. https://www.seas.harvard.edu/courses/cs281/papers/rasmussen-1999a.pdf.

Rizvi, M. H., Rustagi, J. S., Siegmund, D., eds. (1983): *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*. (Academic Press, New York).

Rossi, P. E. (2014): *Bayesian Non- and Semi-parametric Methods and Applications*. (Princeton University Press, Princeton). doi:10.1515/9781400850303.

Sampford, M. R., Scott, A., Stone, M., Lindley, D. V., Smith, T. M. F., Kerridge, D. F., Godambe, V. P., Kish, L., et al. (1969): *Discussion on professor Ericson's paper*. J. R. Stat. Soc. B **31**[2], 224–233. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See Ericson (1969b).

Sprenger, J., Weinberger, N. (2021): *Simpson's paradox*. In: *Stanford encyclopedia of philosophy*, ed. by E. N. Zalta (The Metaphysics Research Lab, Stanford). https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson.