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 5 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 "$\mathrm{P}(A|B)$", 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 $\neg$ (not), $\lor$ (or), $\land$ (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

$$\begin{gathered}
\mathrm{P}(\neg A \mid H) + \mathrm{P}(A \mid H) = 1 \\
\mathrm{P}(A \wedge B \mid H) - \mathrm{P}(A \mid B \wedge H)\,\mathrm{P}(B \mid H) = 0 \\
\mathrm{P}(A \vee B \mid H) + \mathrm{P}(A \wedge B \mid H) - \mathrm{P}(A \mid H) - \mathrm{P}(B \mid H) = 0
\end{gathered} \tag{0}$$

which can be used in different guises, for example[1]

$$\frac{\mathrm{P}(A|H)=x}{\mathrm{P}(\neg A|H)=1-x} \qquad \frac{\mathrm{P}(A|B\wedge H)=x \quad \mathrm{P}(B|H)=y}{\mathrm{P}(A\wedge B|H)=xy} \qquad \frac{\mathrm{P}(A|H)=x \quad \mathrm{P}(B|H)=y \quad \mathrm{P}(A\wedge B|H)=z}{\mathrm{P}(A\vee B|H)=x+y-z}$$

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 $X_1, X_2, \dots, Y_1, Y_2, \dots$, which we call "variates" according to statistics terminology. The collection of variates $X_1, X_2, \dots$ 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:

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

[1] Compare with the inference rules for sequent calculus, e.g. Huth & Ryan 2004 ch. 1; Prawitz 1965 § I.2; see also Boričić 2020.

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

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

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).

We must therefore set the probability values of some statements, which can lead us to the probabilities (1) through the rules of the probability (0).

Bibliography

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

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.

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.

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.

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).