Probabilistic models: models for what?

P.G.L. Porta Mana
<piero.mana@ntnu.no>

Draft of 28 September 2021 (first drafted ***)

***

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

> Where do probability models come from?
> To judge by the resounding silence over this
> question on the part of most statisticians,
> it seems highly embarrassing.
> Dawid[a]

[a] Dawid 1982 p. 220.

1 What is a model?

*** model must be about decidable statement. Composite models are not [1]

*A **probability model** is a combination of knowledge and hypotheses that allows You to assign numerical probabilities to a specific set of statements and to all their logical combinations* (1)

Such assignment is usually done by specifying the joint probability distribution over the *constituents* (or *basic conjunctions*, or *fundamental products*) of the set of statements, that is, conjunction of all of these in which some, none, or all are negated[2]. In scientific inferences such constituents statements are typically of the form '$X = x$' where X is a measurable quantity.

It is implicit in this definition that the set of statements includes *all* statements involved in Your inference. If Your inference is about given data , then the model must have assigned the probability $\mathrm{P}(\wedge \mid)$.

This definition also contains an implicit distinction between *model* and *hypothesis*. A scientific hypothesis by itself is often not sufficient

[1] Hailperin 2011 § 1.2. [2] Hailperin 1996; 2011 ch. 0.

to assign probabilities for statements about the quantities it involves. Measurements of such quantities, for example, often require additional knowledge or hypotheses about the measurement process.

For the moment let us neglect this distinction between *model* and *hypothesis* and use the two words interchangeably, as often done in the statistics literature; replace 'hypotheses' with 'conjectures' in the definition above. In view of that definition, some problems discussed in the statistics literature appear in a new light, and some disappear altogether.

Let consider the following specific, common situation: our statements are $_1 \coloneqq$ '$X_1 x'_1$, $_2 \coloneqq$ '$X_2 x'_2$, and so on. They concern a set of one, two, or more similar measurements.

1.1 Parametric models and model comparison

1.2 Composite hypotheses

1.3 Nested models

2 ***

The probability calculus, just like the truth calculus, proceeds purely syntactically rather than semantically. That is, if I tell you that H and HD are true, you can conclude that D is true; similarly, if I tell you that $\mathrm{P}(H \mid) = p$ and $\mathrm{P}(HD \mid) = q$, you can conclude (try it as an exercise) that $\mathrm{P}(H\ D \mid) = p + q - 1$. And in either case you don't need to know what H and D are about – they could be about Donald Duck or parallel universes. Don't we too often abuse of this syntactical property? We often say 'under model H our belief about the value of quantity x is expressed by such and such distribution $\mathrm{p}(x \mid H) = f(x)$', without explaining what θ really is and why it leads to f. Aren't terms such as 'model' and 'hypothesis', as often used in probability and statistics, convenient and respectable-looking carpets under which we can sweep the fact that we don't quite know what we're speaking about? The need to look under the carpet arises, though, the moment we have to specify our pre-data belief, the prior, about the mysterious H.

And yet again, semantics can very well be a by-product of syntax, or the distinction between the two be a chimera[3]. Such important matters are unfortunately rarely discussed in probability and statistics.

***

Only in extremely rare cases does the set of hypotheses $\{H_h\}$ encompass and reflect the extremely complex and fuzzy hypotheses lying in the backs of our minds. The background knowledge is therefore only a simplified picture of our actual knowledge. That's why or the hypotheses $\{H_h\}$ are often called *models*. 'A theory cannot duplicate nature, for if it

[3] Wittgenstein 1999; Girard 2001; 2003.

did so in all respects, it would be isomorphic to nature itself and hence useless, a mere repetition of all the complexity which nature presents to us, that very complexity we frame theories to penetrate and set aside. If a theory were not simpler than the phenomena it was designed to model, it would serve no purpose. Like a portrait, it can represent only a part of the subject it pictures. This part it exaggerates, if only because it leaves out the rest. Its simplicity is its virtue, provided the aspect it portrays be that which we wish to study' (Truesdell[4]).

***

A good portion of recent literature seems to focus on the relation between a hypothesis and *future* data rather than the collected data; but I think it's equally important to see how it relates to the data we already have. This tension between future and past data stems, in my opinion, from a misunderstanding about which model or hypothesis we're actually interested in. I hope to discuss this topic more at length in another work; but let me give a brief sketch of this misunderstanding.

As mentioned in the introduction, H sometimes denotes some unspecified knowledge that leads to a specific probabilistic model $\mathrm{P}(\dots \mid H)$ (see the remark about syntax and semantics on p. 2). We must remember that the conjunction of H and some data D defines a new and *different* probabilistic model $H' \coloneqq D\,H$. In the context of exchangeable models[5] there's an important distinction between two kinds of models:

(a) models that lead to the *same* belief about new data, independently of the data already observed:

$$\mathrm{P}(D' \mid D\,H) = \mathrm{P}(D' \mid H) \qquad (\text{if } D \not\Rightarrow D'); \tag{2}$$

these are usually called 'simple' hypotheses or models[6], but I prefer to call them *non-learning* models, because when we assume such a model, observed data become irrelevant for our beliefs about unobserved data (although observed data are of course relevant for our belief about the model);

(b) models that lead to different beliefs about new data, depending on the data already observed:

$$\mathrm{P}(D' \mid D\,H) \neq \mathrm{P}(D' \mid H) \qquad (\text{if } D \not\Rightarrow D'); \tag{3}$$

[4] Truesdell et al. 1980 Prologue p. xvi. [5] Bernardo et al. 2000 § 4.2; Dawid 2013; de Finetti 1937; Kingman 1978; Koch et al. 1982; see also Diaconis et al. 1980. [6] Bernardo et al. 2000 § 6.1.4.

these are usually called 'composite' hypotheses or models,[6] but I prefer to call them *learning* models, for reasons analogous to the ones in the previous point.

Geometrically, learning models constitute a convex set, of which the non-learning models are the extreme points.[7] In fact, a learning model can be considered as a state of uncertainty about which specific non-learning model holds, within a certain set of non-learning models. Our belief about new data is therefore give, by the theorem of total probability, by a mixture of our beliefs conditional on the contemplated non-learning models entertained; the weights of this mixture being our distribution of belief among these models themselves.

From this point of view the question 'how well does the *learning* model forecast new data?' appears somewhat strange. Because it's the *non-learning* models that, so to speak, forecast new data, while the learning model only expresses our beliefs in those models. And it's the *observed* data that help us sharpen such beliefs (future data can't help us: we don't have them) – precisely by quantifying how much the competing individual non-learning models have been able to forecast the observed data so far. And this is exactly what the likelihood or log-likelihood (**??**) does, whereas the log-score (**??**) only does so incompletely.

Learning models have the important, factual property that *their conjunction with observed data tends to a non-learning model* as the number of observed data increases. Geometrically, as data accumulate, our model moves in the convex set mentioned above, approaching an extreme point.[8] I say 'factual' because this is not a mathematical theorem: we may imagine and do encounter sequences of data for which no convergence occurs;[9] but in practice this usually makes us replace our simplified set of hypotheses with a less simplified one and start anew. There's a sort of Bayes theorem operating on a large hypothesis space which we keep on the back of our minds, of which the space represented on paper is just a necessarily simplified portrait or caricature (see Truesdell's quote in § **??**). Compare Jaynes's[10] discussion about the 'resurrection of dead hypotheses'.

The hypotheses typically investigated in the sciences are non-learning models.

[7] Lauritzen 1974; 1988; 1984; Barndorff-Nielsen 2014; Dynkin 1978; Dawid 2013 gives a terse summary. [8] compare Hjort in Barron et al. 1986 p. 50; Hjort 1986 ch. 3. [9] see Bruno 1964; Berk 1966; 1970; also Diaconis 1988 § 3. [10] Jaynes 2003 § 4.4.1.

the degrees of belief about data given some hypothesis only *after* some data have been collected, but I think it's equally important to see how it relates to the data we already have. (compare for example[11], although their motivation stems from o

If we see the symbol H as just a placeholder for a probabilistic model, rather than an empirical hypothesis (see the remark about syntax and semantics on p. 2), we must remember that the conjunction of H and some data D simply defines a new probabilistic model $H' \coloneqq H\,D$.

***

This approximation is reasonable if the amount of data is large with respect to the dimension of the space of a single datum, because *** (ref to geisser, stone, gelfandetal)

***

*** remark that $D\,H$ is a *new* probability model: which of the two are we assessing? Connection with learning vs non-learning models which I hope to take up in another work.

***[12] show the approximation only valid if number of data large enough – not very interesting situation in today's problems.

*** problems calculation with time-relevant hypotheses

'we cannot give a universal rule for them beyond the common-sense one, that if anybody does not know what his suggested value is, or whether there is one, he does not know what question he is asking and consequently does not know what his answer means'[13].

[11] O'Hagan 1995; Berger et al. 1996; 1998. [12] Bernardo et al. 1994 § 6.1.6. [13] Jeffreys 1983 § 3.1 p. 124.

Bibliography

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

Barndorff-Nielsen, O. E. (2014): *Information and Exponential Families: In Statistical Theory*, repr. (Wiley, New York). First publ. 1978.

Barron, A. R., Berger, J., Clayton, M. K., Dawid, A. P., Doksum, K. A., Lo, A. Y., Doss, H., Hartigan, J. A., et al. (1986): *Discussion and rejoinder: On the consistency of Bayes estimates*. Ann. Stat. **14**[1], 26–67. 10.1214/aos/1176349831–10.1214/aos/1176349843. See **diaconisetal1986**.

Berger, J. O., Pericchi, L. R. (1996): *The intrinsic Bayes factor for model selection and prediction*. J. Am. Stat. Assoc. **91**[433], 109–122.

— (1998): *Accurate and stable Bayesian model selection: the median intrinsic Bayes factor*. Sankhyā A **60**[1], 1–18. https://www2.stat.duke.edu/~berger/papers/medianibf.html.

Berk, R. H. (1966): *Limiting behavior of posterior distributions when the model is incorrect*. Ann. Math. Stat. **37**[1], 51–58.

— (1970): *Consistency a posteriori*. Ann. Math. Stat. **41**[3], 894–906.

Bernardo, J.-M., Smith, A. F. (1994): *Bayesian Theory*. (Wiley, Chichester).

— (2000): *Bayesian Theory*, repr. (Wiley, New York). 10.1002/9780470316870. First publ. 1994.

Bruno, A. (1964): *On the notion of partial exchangeability*. Giorn. Ist. Ital. Att. **27**, 174–196. Transl. in **definetti1972**, ch. 10, pp. 229–246.

Dawid, A. P. (1982): *Intersubjective statistical models*. In: Koch, Spizzichino (1982): 217–232.

— (2013): *Exchangeability and its ramifications*. In: **damienetal2013**: ch. 2:19–29.

de Finetti, B. (1937): *La prévision: ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**[1], 1–68. Transl. in **kyburgetal1964_r1980**, pp. 53–118, by Henry E. Kyburg, Jr.

Diaconis, P. (1988): *Recent progress on de Finetti's notions of exchangeability*. In: **bernardoetal1988**: 111–125. With discussion by D. Blackwell, Simon French, and author's reply. http://statweb.stanford.edu/~cgates/PERSI/year.html, https://statistics.stanford.edu/research/recent-progress-de-finettis-notions-exchangeability.

Diaconis, P., Freedman, D. (1980): *Finite exchangeable sequences*. Ann. Prob. **8**[4], 745–764. 10.1214/aop/1176994663.

Dynkin, E. B. (1978): *Sufficient statistics and extreme points*. Ann. Prob. **6**[5], 705–730.

Girard, J.-Y. (2001): *Locus solum: From the rules of logic to the logic of rules*. Math. Struct. in Comput. Science **11**[3], 301–506. http://iml.univ-mrs.fr/~girard/Articles.html. See also **curien2001**.

— (2003): *From foundations to ludics*. Bull. Symbolic Logic **9**[2], 131–168. http://iml.univ-mrs.fr/~girard/Articles.html.

Hailperin, T. (1996): *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. (Associated University Presses, London).

— (2011): *Logic with a Probability Semantics: Including Solutions to Some Philosophical Problems*. (Lehigh University Press, Plymouth, UK).

Hjort, N. L. (1986): *Notes on the theory of statistical symbol recognition*. Tech. rep. 778/1986. (Norwegian Computing Center, Oslo).

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). 10.1017/CBO9780511790423. Ed. by G. Larry Bretthorst. First publ. 1994.

https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffreys, H. (1983): *Theory of Probability*, 3rd ed. with corrections. (Oxford University Press, London). First publ. 1939.

Kingman, J. F. C. (1978): *Uses of exchangeability*. Ann. Prob. **6**[2], 183–197.

Koch, G., Spizzichino, F., eds. (1982): *Exchangeability in Probability and Statistics*. (North-Holland, Amsterdam).

Lauritzen, S. L. (1974): *Sufficiency, prediction and extreme models*. In: **barndorffnielsenetal1974**: 249–269. With discussion. Repr. without discussion in **lauritzen1974_r1974**.

— (1984): *Extreme point models in statistics*. Scand. J. Statist. **11**[2], 65–83. See also discussion and reply in **barndorffnielsenetal1984**.

— (1988): *Extremal Families and Systems of Sufficient Statistics*. (Springer, Berlin). First publ. 1982.

O'Hagan, A. (1995): *Fractional Bayes factors for model comparison*. J. Roy. Stat. Soc. B **57**[1], 99–138.

Truesdell III, C. A., Muncaster, R. G. (1980): *Fundamentals of Maxwell's Kinetic Theory of a Simple Monatomic Gas: Treated as a Branch of Rational Mechanics*. (Academic Press, New York).

Wittgenstein, L. (1999): *Philosophical Investigations / Philosophische Untersuchungen*, re-issued 2nd ed. (Blackwell, Oxford). German text, with English transl. by G. E. M. Anscombe. Written 1945–49; first publ. 1953.