What is a probability model?

Luca
<piero.mana@ntnu.no>

Draft of 10 January 2019 (first drafted 21 October 2017)

Some notes or model comparison and selection, hypothesis testing, parameter estimation, Bayes factors, and so on.
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 1982 p. 220)

Summary:

Problems are reported in the literature about Bayes factors and model averaging. The discussions suggest that we ask: (1) Is this an inference problem – that is, quantification of uncertainty – or a decision problem? (2) What is our uncertainty, ultimately, about? (3) What must we decide about?

When the problem is purely inferential, the focus is usually on some kind of parameters. But is our uncertainty about the parameters what we really care about? It seems to me that what we want to quantify is, in the end, our uncertainty about some concrete situation, for example whether an individual has a particular disease, or the amount of a particular resource in a particular place. Parameters may enter the mathematical formulation of our uncertainty, but they aren't the end goal.

Focus on the parameters may be important in particular kinds of decision problems. For example, a software engineer must choose the specific form of a function that will be used in some application – say, to convert handwriting to character codes – or the specific form of a plausibility model to be used in some automated inferential task; or a physicists wants to choose a particular theory (a problem especially considered by Jeffreys 1973; 1983). In such situations the engineer or physicist has to choose from a manifold of possible functions or models, and the parameters are just coordinates on that manifold. Choosing a parameter corresponds to choosing a function or model. This choice requires a quantified uncertainty about some 'true' model (a notion to

be thoroughly discussed) but also a quantified gain or loss function for the choice. In this kind of problems a lot of attention is usually given to the quantification of uncertainty – choosing priors, likelihoods, and so on – but very little to a reasonable quantification of the gain function. Thus the final quality of the solution, which involves both, is poor.

Sometimes there is a more general problem upstream: the initial inferential or decision problem is poorly formulated. For example, many 'null hypothesis' questions force an unnatural dichotomous view on situations better viewed as a continuum of cases. And questions like 'is this probability model true?' baffle comprehension if we sit down and try to explain – without circularly using technical jargon – what they're actually asking.

The bottom problem is, in my opinion, that even though formal Bayesian methods are becoming more common, the way of thinking about and formulating inference and decision problems is still antiquated. In particular the 'null hypothesis' or 'what is the true model' ways of thinking.

Here's an example. In genome-wide association studies we observe that some diseases are more frequent among individuals that have particular variants of pieces of DNA, called single-nucleotide polymorphisms. This means that we could examine a person's SNPs to make better guesses on whether the person has or will develop a particular disease. Our uncertainty come from the small size of our samples: we'd like to guess the conditional frequencies of the disease given the SNP variants in the full population. Many studies face this kind of association in a dichotomous way: 'is this particular SNP associated with that disease? yes or no?' But the association isn't dichotomous: there is a continuum of differences between the frequencies of the disease conditional on the SNP variants. The question is: how do we want to use this association? what choices do we need to make? Suppose that the idea is to examine a person's SNPs to diagnose the disease. The issue could be that the cost and time involved allow us to check, say, only one SNP. Then our decision problem is to choose the SNP to be used in an automated diagnostics. For simplicity assume that the candidate SNPs have only two variants each. The gain in examining a SNP for diagnostics is proportional to the absolute difference $|\triangle f|$ between the frequencies, in the full population, of the disease conditional on the two variants. Let's say that a good gain function is $|\triangle f|^2$, and our degree of belief about $|\triangle f|$ given some sample

study with data D and background information I is $\mathrm{p}(|\triangle f|\,|\,D, I)$. Then according to decision theory we should choose the SNP with highest expectation $\mathrm{E}(|\triangle f|^2\,|\,D, I)$. This simple example shows that no 'null hypothesis' or dichotomous questions where necessary for our inference and decision. I believe that this is the case for many problems that are today stated in 'null hypothesis' terms.

Many inference problems boil down to guessing what the next outcome in a possibly unlimited sequence of observations will be, given the outcomes of some of the observations in this sequence. If we believe that the precise order of these outcomes is irrelevant for our guesses (an exchangeability assumption), then we can break our guess as follows: (a) what would our guess be if we knew the relative frequencies of the outcomes in the whole (observed an unobserved) sequence? (b) what is our guess about such relative frequencies? If the number of observations is very large, by symmetry our degree of belief (a) about the next outcome would be its known relative frequency. So what remains is guess (b). Most inferences about 'true models' can be interpreted as inferences about the unknown final relative frequencies. This point of view calls for a leap of imagination about future observations. But the combined lifespan of the humans involved in our original inference problem is surely finite, so the number of observations is finite although large. The problem thus becomes analogous to drawing from a finite urn without replacement, with the complication that the number of balls in the urn is unknown: a mixture of generalized hypergeometric distributions. It would be interesting to prove that the resulting degree of belief (a) from such a mixture can be well approximated by a final relative frequencies, if our degree of belief about the number of balls is peaked at large values. I like this pragmatic point of view because it gives concrete sense of statements about 'true models' and it doesn't call for more metaphysics than planning what to do in case of sunshine or of rain tomorrow. And it doesn't bring 'propensities' or other doubtful physical notions into the picture.

1 Motivation

The Bayesian literature on probabilistic modelling, including model comparison, hypothesis testing, and parameter estimation, often considers

exchangeable parametric models for our degrees of belief, of this form:

$$\mathrm{p}(D|M,I) = \int \mathrm{d}\theta\; \mathrm{p}(D|\theta,M,I)\,\mathrm{p}(\theta|M,I), \tag{1}$$

where M represents the model, I other background information and assumptions, D a sequence of data, and θ a parameter with values in some manifold. This is a particular case of an exchangeable model (Bernardo et al. 2000 ch. 4). *Bayes factors* and *evidence* (Good 1985; MacKay 1992; Kass 1993; Kass et al. 1995), and model averaging (Draper 2005; Chatfield 1995; Draper 1995; Hoeting et al. 1999) are often discussed in this kind of literature. If we have several mutually exclusive and exhaustive models M_1, M_2, ... our degree of belief about each given some data is

$$\mathrm{p}(M_j|D,I) = \frac{\mathrm{p}(D|M_j,I)\,\mathrm{p}(M_j|I)}{\sum_k \mathrm{p}(D|M_k,I)\,\mathrm{p}(M_k|I)}; \tag{2}$$

this degree of belief involves the *evidence*, which is just $\mathrm{p}(D|M_j,I)$, eq. (1). The Bayes factor between two models is just the ratio of their evidences.

The literature reports several open issues in 'model comparison' and the evaluation of evidences and Bayes factors. For example, to calculate a model's evidence we must solve the integral in eq. (1), and this requires specifying a distribution for the parameters $\mathrm{p}(\theta|M,I)$. When this distribution is improper the evidence vanishes, so the ratio of two evidences becomes undetermined. Several ways of fixing this issue have been proposed in the literature (Kass et al. 1995; O'Hagan 1995; Berger et al. 1996; De Santis et al. 1997; Berger et al. 1998). It has also been pointed out if we choose just one out of several possible models, basing our choice on their evidence, and then we this model to make inferences, we can end up misrepresenting our degree of belief (Draper 2005; Chatfield 1995; Draper 1995; Hoeting et al. 1999). Our degree of belief is in fact the weighted average of those given by each model. Yet, in some concrete applications – think about the use of a neural net – we have to choose only one model. ✠ some reference justly stated that this is the domain of decision theory; can't find it

These issues invite us to re-examine what we're actually doing in 'model comparison' or 'model selection'. This is the purpose of the present note.

From a Bayesian point of view model comparison doesn't really exist or is superfluous, because if we are unsure about several models M_j, then our degree of belief, by the theorem of total probability, is just

$$\mathrm{p}(D|\, I) = \sum_j \mathrm{p}(D|\, M_j, I)\, \mathrm{p}(M_j|\, I). \tag{3}$$

As we gather new data D' our degree of belief is updated to

$$\mathrm{p}(D|\, D', I) = \frac{\mathrm{p}(D, D'|\, I)}{\mathrm{p}(D'|\, I)} \equiv \frac{\sum_j \mathrm{p}(D, D'|\, M_j, I)\, \mathrm{p}(M_j|\, I)}{\sum_k \mathrm{p}(D'|\, M_k, I)\, \mathrm{p}(M_k|\, I)}, \tag{4}$$

which can be suggestively rearranged as follows, multiplying and dividing by $\mathrm{p}(D'|\, M_j, I)$:

$$\mathrm{p}(D|\, D', I) = \sum_j \underbrace{\frac{\mathrm{p}(D, D'|\, M_j, I)}{\mathrm{p}(D'|\, M_j, I)}}_{=:\mathrm{p}(D|D', M_j, I)} \underbrace{\frac{\mathrm{p}(D'|\, M_j, I)\, \mathrm{p}(M_j|\, I)}{\sum_k \mathrm{p}(D'|\, M_k, I)\, \mathrm{p}(M_k|\, I)}}_{=:\mathrm{p}(M_j|D', I)}. \tag{5}$$

This is a weighted sum of our degrees of belief conditional on each model, the weights being our updated degrees of belief about the model themselves. The models which have very low updated degrees of belief effectively drop out of the sum, so the probability calculus is automatically doing 'model selection' for us. From this point of view, the practice of model selection can be viewed as an approximation of the formula above. The problem with improper distributions for the parameters still remains, though: models involving such distributions assign a vanishing degree of belief to the data and therefore disappear from eqs (4) and (5). We'll discuss this issue later.

If we compare the model-averaging formula (3) with the parametric formula (1) for a specific model, we see that the two have the same structure. The second is just the continuum limit of the first. In fact many of the issues about model comparison just discussed appear also when we consider just one model of parametric form (1). Given new gathered data D', our degree of belief is updated to

$$\mathrm{p}(D|\, D', M, I) = \frac{\mathrm{p}(D, D'|\, M, I)}{\mathrm{p}(D'|\, M, I)} \equiv \frac{\int \mathrm{d}\theta\ \mathrm{p}(D, D'|\, \theta, M, I)\, \mathrm{p}(\theta|\, M, I)}{\int \mathrm{d}\theta'\ \mathrm{p}(D'|\, \theta', M, I)\, \mathrm{p}(\theta'|\, M, I)}, \tag{6}$$

which can be suggestively rearranged as follows, multiplying and dividing by $\mathrm{p}(D'|\,\theta, M, I)$:

$$\mathrm{p}(D|\,D', M, I) = \int \mathrm{d}\theta\; \underbrace{\frac{\mathrm{p}(D, D'|\,\theta, M, I)}{\mathrm{p}(D'|\,\theta, M, I)}}_{=:\mathrm{p}(D|D',\theta,M,I)} \underbrace{\frac{\mathrm{p}(D'|\,\theta, M, I)\,\mathrm{p}(\theta|\,M, I)}{\int \mathrm{d}\theta'\,\mathrm{p}(D'|\,\theta', M, I)\,\mathrm{p}(\theta'|\,M, I)}}_{=:\mathrm{p}(\theta|D',M,I)}. \tag{7}$$

This is again equivalent to the model-average case (5) but for one important difference: in the first fraction in the product above, the dependence on data D' disappear:

$$\mathrm{p}(D|\,D', \theta, M, I) = \mathrm{p}(D|\,\theta, M, I). \tag{8}$$

This is a typical property of parametric models: the presence of the parameters in the conditional makes all other new data in the conditional *irrelevant*. Use of an improper distribution $\mathrm{p}(\theta|\,M, I)$ for the parameter usually doesn't lead to problems in this case if correctly used – that is, taking the limit only at the very end of all calculations (Jaynes 2003 ch. 15). Analogously to the model-average case, parameters with low updated density $\mathrm{p}(\theta|\,D', M, I)$ give a vanishing contribution to the continuum average (7), so the probability calculus is automatically doing a 'parameter selection' for us. From this point of view, the practice of selecting just one parameter can be viewed as an approximation of the formula above (though such approximation can be bad if the density for θ has many modes).

In some situations, however, we can't or don't want to consider all models in the model-average case (5), even when no model dominates our updated degree of belief over the others. We want just one model. Similarly, in some situations we can't or don't want to consider all parameters in the single-model case (7), even when the updated density for the parameter has no dominating peaks. We want just one parameter. These analogous situations involve *decision theory*, and therefore require us to assign gain/loss functions. But what is our decision actually about? is it really about a model or a parameter? or rather about possible data outcomes? Let's examine these question after making some remarks on models in general.

2 Remarks on models

I consider exchangeable models, but the following remarks may apply to more general cases.

Let's use this restricted definition: given a set Y of possible data, and possibly a set X of conditional data, a *probability model* is a conjunction M of assumptions or hypotheses that allows us to assign a definite, numerical plausibility

$$\mathrm{p}(y_1, y_2, \dots \mid x_1, x_2, \dots, M, I) \tag{9}$$

for every meaningful combination of $y_i \in Y$ and $x_i \in X$. For the moment I consider finite sets.

This definition applies in particular to exchangeable models, where X is empty; to partially exchangeable models, where X is a set of labels for the exchangeable categories; and to models used in machine learning and neural nets. This definition also include functions or maps $f\colon X \to Y$ as special cases, when the plausibility is unity for a particular y only, dependent on x: $\mathrm{p}(y \mid x, M, I) = \delta[y, f(x)]$.

An important distinction can be made between two kinds of model: *learning* models and *non-learning* models, which can also be called *extremal* for reasons explained later.

A learning model M is one that yields different plausibilities about some data $(y_1, y_2, \dots) =: \boldsymbol{y}$ conditional on $(x_1, x_2, \dots) =: \boldsymbol{x}$ if we condition on knowledge about other data $(y'_1, y'_2, \dots) =: \boldsymbol{y}'$, $(x'_1, x'_2, \dots) =: \boldsymbol{x}'$:

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x},\ \boldsymbol{y}', \boldsymbol{x}',\ M, I) \neq \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x},\ M, I). \tag{10}$$

A non-learning model is one for which these plausibilities are not affected:

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x},\ \boldsymbol{y}', \boldsymbol{x}',\ M, I) = \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x},\ M, I). \tag{11}$$

This means that

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x},\ M, I) = \prod_i \mathrm{p}(y_i \mid x_i,\ M, I). \tag{12}$$

From the above formulae we see that when we update a learning model we obtain a new learning model; when we update a non-learning model we obtain the same non-learning model.

If we look at eq. (8) we see that the conjunction (θ, M) is defining a non-learning model. So a model like M, eq. (1), which is clearly a

learning model, is given as a continuous weighted mixture of non-learning models. This is a property of most parametric models typically considered in the literature, and is a corollary of de Finetti's theorem for partially exchangeable models (de Finetti 1938; Bernardo et al. 2000 § 4.6).

The parametric form (1) can also be viewed in the following way. Consider the space of all possible exchangeable models. This space is equivalent to the space of all possible limit frequency distributions for the possible outcomes of our data. The choice of $\mathrm{p}(D|\,\theta, M, I)$ corresponds to the selection of a submanifold in this space, and the choice of $\mathrm{p}(\theta|\,M, I)$ corresponds to the choice of a plausibility density on such submanifold. Thus a model M could also be divided into two assumptions, the first more general, selecting a set, and the second more specific, leading to specific degrees of belief for the elements of that set. When we consider several models we are selecting several distinct submanifolds (possibly intersecting in sub-submanifolds of lower dimension), with a density on each.

My impression is that in model-comparison studies 'model' often seems to be mistakenly identified with the specification of a submanifold only, as if the specification of the density were not part of the model. Just as an example, see Kass et al. (1995)'s extensive review: even if they say 'the prediction rule is derived from the model H_k *(i.e., likelihood and prior)*' (p. 777, my emphasis), they also say 'Bayes factors require priors on the parameters appearing in the models that represent the competing hypotheses' (p. 773), 'the prior distributions $\pi(\theta_k|\,H_k)$ on the parameters of each model must be specified' (p. 781), 'Sensitivity analysis concerns distributional forms for models $\mathrm{pr}(\mathbf{D}|\,\theta_k, H_k)$ as well as priors', 'In choosing priors, just as in choosing models for data distributions, simplifications are often made' (p. 781). The last sentences seem to set the choice of a parameter prior apart from that of a model. But without parameter prior there is no model at all.

At the same time, the possible tendency of seeing a prior as something apart from a model may point to an unconscious goal: the statistician wants to find one specific *non-learning* model.

3 Decision-theoretic point of view

Let's consider the case when we want to choose one model (that is, one full parametric model like (1) or one specific parameter). This brings decision theory (Raiffa et al. 2000; Berger 1985; Jaynes 2003 chs 13–14) into the picture.

Decision theory concerns the choice of an action a, from a set of possible actions A, whose success depends on a state of affairs we're uncertain about. The gain in choosing action a if the state of affairs is y is given by a gain function $G(a|y)$. In our case the state of affairs is the value $y \in Y$ of some quantity, and our degree of belief about it is $\mathrm{p}(y|x, \boldsymbol{y}', \boldsymbol{x}', M, I)$, as in the previous section. Decision theory says that we should choose the action that minimizes the expected gain:

$$\text{choose} \quad \arg\sup_{a\in A} \int \mathrm{d}y\; G(a|y)\, \mathrm{p}(y|x, \boldsymbol{y}', \boldsymbol{x}', M, I). \tag{13}$$

It's important to stress that we're choosing an *action* $a \in A$, which needs not be a quantity like y or a parameter like θ.

In fact, our first question is this: in a model-selection problem, what is our choice actually about? Let's examine several possible scenarios and answers.

First scenario: we are at the user end of the problem, in the sense that we have a specific input datum x and we need to choose an action a. Example: translating a hand-written digit into a definite decimal digit; or choosing a treatment subsequently to medical tests. If we insert the parametric form (1) into the decision rule (13) we find

$$\text{choose} \quad \arg\sup_{a\in A} \iint \mathrm{d}y\, \mathrm{d}\theta\; G(a|y)\, \mathrm{p}(y|x, \theta, M, I)\, \mathrm{p}(\theta|\boldsymbol{y}', \boldsymbol{x}', M, I). \tag{14}$$

As discussed for example by Jaynes (2003 § 13.12.1), our choice depends on the combination of likelihood and gain function

$$\int \mathrm{d}y\; G(a|y)\, \mathrm{p}(y|x, \theta, M, I). \tag{15}$$

Different combinations may lead to the same decision. From this point of view, it is silly to put much care in the specification of likelihood, if we don't put as much care in the specification of the gain function.

Second scenario: we are at the engineering end of the problem, in the sense that, for example, we're building some automated software that will be used with different inputs. We don't know in advance which sequence of inputs will be given to the software, except that they belong to X. The engineer's choice is about a full model; that is, $a \in A$ denotes possible models. This scenario has two sub-scenarios: we may be building software that automatically learns from its applications, or software that gives the same answer for the same input. Let's consider the second sub-scenario, which currently seems most common.

Using 'H' instead of 'a' for our choice, to remind ourselves that we're choosing a model, we have

$$\text{choose} \quad \arg\sup_{H\in A} \sum_j \int \mathrm{d}\theta \; G(H|\,\theta, M_j)\, \mathrm{p}(\theta, M_j|\,\boldsymbol{y}', \boldsymbol{x}',\, I). \tag{16}$$

Note that the model H doesn't need to belong to the set of non-learning models $\{(\theta, M_j)\}$.

Old text

When we ask about the probability of a 'model' given the data, we're asking if a given region of limit relative frequencies is more probable than another. This may not be what we want to ask, because one region as a whole can have higher probability than another region, and yet a particular frequency in the second region may be more probable than any single frequency in the first region.

I a way, our real goal is to guess the limit frequency, not guess a region in frequency space, which may be quite arbitrary and whose shape has nothing to do with our predictions.

A solution to this is to reparameterize every 'model' with a coordinate that has the same meaning across them, and then work with the union of these models, forgetting about their individualities.

But does it make sense to ask whether the limit frequency distribution belongs to parametric family rather than another? It is like asking whether the limit frequency belongs to a submanifold (for example a curve) rather than another in the simplex, in the case with finite number of outcomes. The limit frequency belongs to many submanifolds at once.

If we really want to ask that question we should first choose a probability distribution in the whole limit-frequency space, a metric, and then determine the induced probability on the submanifold.

This kind of question may be useful if we are asking about several *experiments*, not just one. In this case it may make sense to ask whether their different limit frequencies belong to some common submanifold.

'Model' often seems to be mistakenly identified with the specification of likelihoods only, as if the specification of the parameter prior were not part of the model. Compare Kass et al. (1995): 'Bayes factors require priors on the parameters appearing in the models that represent the competing hypotheses' (p. 773) 'the prior distributions $\pi(\theta_k|\, H_k)$ on the parameters of each model must be specified' (p. 781), 'Sensitivity analysis concerns distributional forms for models $\mathrm{pr}(\mathbf{D}|\, \theta_k, H_k)$ as well as priors', 'In choosing priors, just as in choosing models for data distributions, simplifications are often made' (p. 781). But see also 'the prediction rule is derived from the model H_k *(i.e., likelihood and prior)*' (p. 777, emphasis added).

The probability of hypotheses like those – concerning whole regions of limit-frequency space – cannot be computed.

The problem of 'model dimensionality' is also misplaced because we identify models with likelihoods only. In reality the dimensionality of a model is determined by the parameter prior. In fact, the very choice of likelihood can be interpreted as the choice of a particular prior from a 'non-parametric' point of view. (Compare Kass (1995), end of § 6.1.)

Also the idea of model *selection* can be dangerous, because we may be discarding the model than contains the frequency with highest likelihood.

The evidence is just an average of cross-validations (or splitting, see Kass § 6.5). Naive cross-validation is testing the wrong hypothesis.

4 Prediction and forecast

Some notation: We assume to have a possibly infinite set of observations, each of which can yield one of N outcomes, labelled by integers i. The proposition $O_i^{(a)}$ denotes that outcome i is observed at the ath observation. Such propositions also contain information about the time or place where the outcome was observed, so that from a proposition like $O_{i_2}^{(2)} \wedge O_{i_4}^{(4)}$ we can for example infer the time interval $t^{(4)} - t^{(2)}$ between observations number 4 and 2.

A statistical model is a set of assumptions M that jointly allow us to consistently assign numerical values to the probabilities

$$\mathrm{P}\bigl(O_{i_{n+1}}^{(a_{n+1})} \bigm| O_{i_1}^{(a_1)} \wedge \dots \wedge O_{i_n}^{(a_n)} \wedge M\bigr), \tag{17}$$

for any legitimate n and any sets of observations $\{a_1, \dots, a_{n+1}\}$ and outcomes $\{i_1, \dots, i_{n+1}\}$; 'consistently' means that these assignments are properly related by operations like marginalization.

This definition is very general; in fact it amounts to say that a model is an assignment of the probabilities for all possible conjunctions of n outcomes, for all legitimate n.

it includes exchangeable models of various kinds, models for time series and forecasts.

Now I'd like to make a distinction between two main classes of statistical models: those that 'learn' and those what 'don't learn'.

This distinction is clear within the subclass of infinitely exchangeable models: for any such model the probability above has the form

$$\int \mathrm{p}(i_{n+1}|\,\theta, M)\,\mathrm{p}(\theta|\,i_1,\dots,i_n,M)\,\mathrm{d}\theta, \tag{18a}$$

$$\mathrm{p}(\theta|\,i_1,\dots,i_n,M) \propto \Bigl[\prod_{k=1}^{n}\mathrm{p}(i_k|\,\theta,M)\Bigr]\,\mathrm{p}(\theta|\,M), \tag{18b}$$

where the specific form of $\mathrm{p}(i|\,\theta,M)$ and $\mathrm{p}(\theta|\,M)$ are determined by the model. Within this subclass, models that don't learn are characterized by $\mathrm{p}(\theta|\,M) = \delta(\theta-\theta^*)$, so that the probability for an outcome does not depend on knowledge of other outcomes:

$$\mathrm{P}\bigl(O_{i_{n+1}}^{(a_{n+1})}\bigr|\; O_{i_1}^{(a_1)}\wedge\dots\wedge O_{i_n}^{(a_n)}\wedge M\bigr) = \mathrm{P}\bigl(O_{i_{n+1}}^{(a_{n+1})}\bigr|\; M\bigr) \equiv \mathrm{p}(i_{n+1}|\,\theta^*,M). \tag{19}$$

Such a model doesn't 'learn' because it makes all knowledge about other observations irrelevant for the prediction of each observation.

Among all statistical models for a particular set of

**models (e.g. exchangeability for which accumulation of data leads to stable probabilities, and models (e.g. Markov) for which this doesn't happen.

Thanks

. . . to Mari & Miri for continuous encouragement and affection, and to Buster Keaton and Saitama for filling life with awe and inspiration. To the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Bibliography

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

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. (Springer, New York). First publ. 1980.

Berger, J. O. et al. (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.

Bernardo, J.-M. et al., eds. (1985): *Bayesian Statistics 2*. (Elsevier and Valencia University Press, Amsterdam and Valencia). https://www.uv.es/~bernardo/valenciam.html.

Bernardo, J.-M. et al. (2000): *Bayesian Theory*, reprint. (Wiley, New York). First publ. 1994.

Chatfield, C. (1995): *Model uncertainty, data mining and statistical inference*. J. Roy. Stat. Soc. A **158**[3], 419–444. See also discussion in Copas et al. (1995).

Clyde, M. et al. (1999): *[Bayesian model averaging: a tutorial:] Comments and rejoinder*. Stat. Sci. 412–417. See Hoeting et al. (1999).

Copas, J. B. et al. (1995): *Discussion of the paper by Chatfield [Model uncertainty, data mining and statistical inference]*. J. Roy. Stat. Soc. A **158**[3], 444–466. See Chatfield (1995).

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

De Santis, F. et al. (1997): *Alternative Bayes factors for model selection*. Can. J. Stat. **25**[4], 503–515.

de Finetti, B. (1938): *Sur la condition d'équivalence partielle*. In: *Colloque consacré à la théorie des probabilités. VI : Conceptions diverses*. Ed. by B. de Finetti et al. (Hermann, Paris), 5–18. Transl. in Jeffrey (1980), pp. 193–205, by P. Benacerraf and R. Jeffrey.

Draper, D. (1995): *Assessment and propagation of model uncertainty*. J. Roy. Stat. Soc. B **57**[1], 45–70. See also discussion and reply in Spiegelhalter et al. (1995). https://classes.soe.ucsc.edu/ams206/Winter05/draper.pdf.

— (2005): *On the relationship between model uncertainty and inferential/predictive uncertainty*. https://users.soe.ucsc.edu/~draper/writings.html, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.9402. First written 1993.

Good, I. J. (1985): *Weight of evidence: a brief survey*. In: Bernardo et al. (1985), 249–270. With discussion by H. Rubin, T. Seidenfeld, and reply.

Hoeting, J. A. et al. (1999): *Bayesian model averaging: a tutorial*. Stat. Sci. **14**[4], 382–412. See also Clyde et al. (1999).

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

Jeffreys, H. (1973): *Scientific Inference*, 3rd ed. (Cambridge University Press, Cambridge). First publ. 1931.

— (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.

Kass, R. E. (1993): *Bayes factors in practice*. The Statistician **42**[5], 551–560. http://ecologia.ib.usp.br/bie5782/lib/exe/fetch.php?media=bie5782:00_curso_avancado:uriarte:kass_statistician_1993_bayesfactors.pdf.

Kass, R. E. et al. (1995): *Bayes factors*. J. Am. Stat. Assoc. **90**[430], 773–795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf; https://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf.

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

MacKay, D. J. C. (1992): *Bayesian interpolation*. Neural Comp. **4**[3], 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html.

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

Raiffa, H. et al. (2000): *Applied Statistical Decision Theory*, reprint. (Wiley, New York). First publ. 1961.

Spiegelhalter, D. J. et al. (1995): *Discussion of the paper by Draper [Assessment and propagation of model uncertainty]*. J. Roy. Stat. Soc. B **57**[1], 71–97. See Draper (1995). https://classes.soe.ucsc.edu/ams206/Winter05/draper.pdf.