

PROBABILITY AND GEOMETRY

John Skilling

Maximum Entropy Data Consultants, Kenmare, Co. Kerry, Ireland

ABSTRACT

There is a unique framework for rational inference. It is the ordinary, simple calculus of probability.

Key words: maximum likelihood, unbiased estimator, frequentist, null hypothesis, chisquared test, F-test, t-test, improper prior, fuzzy logic, quantum logic.

1. INTRODUCTION

I want to learn about the world. I assume you do too. This is how to do it. The keywords above are how *not* to do it.

2. PROBABILITY CALCULUS

Learning means using data to improve our knowledge. It turns out that there is just *one* way of learning in a logically consistent fashion, and that is to use probability calculus. Here's the reason.

Nowadays, we store our knowledge in computer memory, and quantify it in terms of the number of bits of information it needs for its storage or transmission. The question is how to update this knowledge in the light of new data, and how to use it.

In the best traditions of scientific style, we start with a very simple system. Any general procedure must work in special cases, otherwise it's not general, and here we start with a system with only 3 bits, allowing just $2^3 = 8$ states.

3 bits 8 states 000, 001, ..., 111.

I want to manipulate my beliefs $\beta^{000}, \beta^{001}, \dots, \beta^{111}$ about these states. Which bits do I think are ON? Actually, my beliefs are intended to develop, so it's safer to suffix β^j with my developing state of knowledge I and write β_I^j . For this to be useful, I want transitivity: if I believe i more than j , and j more than k , then I believe i more than k — otherwise I find myself arguing in circles.

$$(\beta^i > \beta^j \text{ AND } \beta^j > \beta^k) \text{ IMPLIES } \beta^i > \beta^k \quad \parallel I$$

Here the “ $\parallel I$ ” notation avoids proliferation of symbols by flagging every belief in range to have that subscript. A consequence of this, already anticipated by the “ $>$ ” notation, is that beliefs β may as well be real numbers, albeit rather arbitrary as yet.

Now, if I believe more that a particular bit j is ON than OFF, then I believe less in the converse “ $\sim j$ ” (NOT j). Hence there must be a function f that gives the one from the other.

$$\beta^{\sim j} = f(\beta^j) \quad \parallel I$$

Already, we have some restriction on generality because double negation $f(f(\beta)) = \beta$ brings us back to the start. But we need more. To learn about “ i, j ” (i AND j ; do I believe both bits are ON?), I want to be able to learn whether i is ON, and then (in the light of that extra knowledge) about j . This means that there is another function F which expresses sequential learning.

$$\beta^{i,j} = F(\beta^i, \beta^j) \quad \parallel I$$

F has to have the symmetry $i \leftrightarrow j$, but there's more.

In 1946, Richard Cox [1] proved that the *only* calculus for manipulating beliefs that obeys these desiderata throughout the three-bit system is the ordinary calculus of proportions known as the rules of probability (p);

$$\left. \begin{array}{l} p^j + p^{\sim j} = 1 \quad \text{sum rule for 1 bit,} \\ p^{i,j} = p^i p_j^j \quad \text{product rule for 2 bits.} \end{array} \right\} \parallel I$$

The argument is simply that we could learn about 3 bits by first learning about any two and getting the third later, and that all such ways must give the same answer.

With these rules in place, the two bits i and j could code for any propositions whatever in bigger computers, leading to the general rules;

$$\left. \begin{array}{l} \sum_j p^j = 1 \quad \text{sum rule, over all allowed states,} \\ p^{i,j} = p^i p_j^j \quad \text{product rule, for any joint state.} \end{array} \right\} \parallel I$$

Beliefs β can be and often are identified with these real numbers between 0 (I believe it's false) and 1 (I believe it's true), but they can also be re-scaled into percentages $100p$, logarithms $\log(p)$ and so on. That freedom of presentation, though, is the *only* freedom. Basically, rational inference has a unique, defined calculus. All of inference,

estimation, and decision theory rest upon manipulating these simple rules in different contexts.

Incidentally, the conditional probability p_i^j is more commonly written as $\Pr(j | i)$, but it is crisper to write the operand before the solidus as a superscript, and the conditional after it as a subscript. This notation also keeps in view that all inference problems in our finite world are discrete. By avoiding infinity, we exclude all possibility of paradox. Naturally, we use continuum notation where convenient, but only as a shorthand. We do not abuse the continuum by mistakenly allowing it to introduce paradoxes of the infinite such as the notorious “improper prior” distributions.

3. INFERENCE

Suppose that I have a list of possible states j for the object I plan to investigate. A little thought always enables me to use my existing knowledge base I to place a prior probability distribution

$$\text{Prior} = \pi^j \equiv \Pr(j) \quad || I \quad \left(\sum_j \pi_j = 1 \right)$$

over the states in question. All I have to do is distribute my allotted unit mass of probability over the states in a plausible manner. Sometimes, this can be done by symmetry. Thus I would assign $\frac{1}{6}$ to each face of the standard cubical die if I had no reason to believe in one face being preferred over another. Sometimes, I work by rejecting what seems silly. For example, I would give no credence to the length of a piece of furniture being smaller than an atom or bigger than the planet, and in practice would put most of my probability between 30 cm and 3 m. Some people take a perverse pride in allowing infinite ranges. I do not. Admittedly, my prior assignment may be somewhat wishy-washy. I just do my best. You are always free to try to do better.

The next step in inference is to plan the observations, usually by acquiring equipment that responds somehow to my unknown states j by producing output states $k = 1, 2, \dots$. Let us assume that the equipment has been calibrated properly, so that for any hypothetical state j we know the probabilistic response

$$\text{Likelihood} = L_j^k \equiv \Pr(k | j) \quad \left(\sum_k L_j^k = 1 \text{ for each } j \right)$$

If there are parameters in this that we don't know properly, then such nuisance parameters can be included in our unknowns without any change of principle.

Without even performing the observations, we can use the calculus with prior and likelihood alone. The product rule gives us

$$\text{Joint} \equiv \Pr(k, j) = L_j^k \pi^j \quad || I$$

for any input j and output k . Summing over j yields

$$\text{Evidence} = Z^k \equiv \Pr(k) = \sum_j L_j^k \pi^j \quad || I$$

and another application of the product rule gives

$$\text{Posterior} = P_k^j \equiv \Pr(j | k) = L_j^k \pi^j / Z^k \quad || I$$

Although these distributions are being given different names and symbols, they are all probabilities, and the calculus is really very simple.

Making the observation comes next. Suppose the data say that k takes a particular value D (in practice, usually many values, but here we are considering their bit pattern as one big integer). We can then get a specific numerical value for the evidence

$$\text{Evidence} = Z^D = \sum_j L_j^D \pi^j \quad || I$$

and thence get the posterior

$$\text{Posterior} = P_D^j = L_j^D \pi^j / Z^D \quad || I$$

The latter is what users usually want. It shows how one's prior belief π is modulated by the data to yield an updated distribution P that is called the posterior. It's called Bayes' Theorem. Probability calculus done like this is sometimes called “Bayesian”, supposedly to distinguish it from other usage that breaks the rules.

One step of learning may not be enough. It is common to employ additional data to learn yet more about the object in question. Suppose we get extra data D' from likelihood function L' . We can then use the posterior from the first observations as a prior to be modulated by the second data, arriving at

$$\text{Post-posterior} \propto (L')_j^{D'} \times (L_j^D \times \pi^j) \quad || I$$

Incidentally, this shows terminology changing with context, where what was once called a posterior in one context becomes called a prior in another.

Of course, the brackets were un-necessary because the multiplications on the right could be done in any order. The datasets could have been used in reverse order, or even combined together as a single composite observation. All these methods give the same answer, as indeed is demanded by elementary consistency.

It was Cox's major contribution to show that probability calculus is required by this sort of consistency requirement. If implemented equivalently to probability calculus, an alternative such as fuzzy logic merely adds un-necessary conceptual complication. If implemented non-equivalently, there should be very simple counter-examples because the Cox argument relies on such a small system. As for quantum logic, it bases its structure on the Hilbert space of normalised complex quantum states, while denying the classical logic required for the mathematics that defines Hilbert space in the first place. Enough said.

4. PRESENTATION

The above account of inference blithely assumed that all the sums could be done. However, if a state requires a million bits (as for a simple 1000×1000 image) for storage, the number of states is 2^{million} and it would take too long to do the sums. Good theory is pointless if it can't be used.

Historically, I surmise that this may be why the “orthodox” paradigm of statistics actually went backwards for a century (c. 1850-1950), as documented in [2] (particularly chapter 17 on “principles and pathology of orthodox statistics”). Our forefathers couldn't do the sums, so had no choice but to botch the methodology, and mistakenly presented their compromises as correct. Nowadays, we have far more computational power, and also have Monte Carlo exploration algorithms which simulate the big sums by finding where most of the probability mass lies. We have the luxury of good theory, and of being able to use it.

So, suppose that we have arrived at a posterior distribution P for some million-cell system. How can we present it, given that we can only show $O(\text{million})$ bits, and not $O(2^{\text{million}})$? One idea is to show the maximum of P , making it sound professional (“*Maximum A-Posteriori Probable*”) and important (“*optimal*”). Another common idea is to ignore the prior and show only the maximum likelihood. Yet, especially in a million dimensions, the maximum of a distribution can easily be far off in an obscure little corner of the bulk of the distribution. It may not be well-defined, either, if the data are incomplete. In any case, probability theory tells us to add up, not to maximise. Probability theory is invariant under coordinate transformations, and maximisation isn't. We can move the maximum of a distribution anywhere we want by local compression of the coordinate axes. Actually, given that P represents our current distribution of belief, the honest thing to do is to use coordinates in which that belief is spread uniformly. But then P is flat, and there's no maximum at all.

Well, given 2^{million} states, all equally likely, the only thing we *can* do is select a few *at random*. There's no other symmetrical criterion. That's how Monte Carlo algorithms work, and that is how their results are presented.

5. HYPOTHESIS COMPARISON

Different people can and do hold different prior beliefs. When I am asked to judge different hypotheses H_1 and H_2 (or more), I think about them in the light of my knowledge base I , and assign a prior π representing my judgment of their plausibility:

$$\text{MyPrior (for hypothesis } \#h) = \pi^h \parallel I \quad \left(\sum_h \pi^h = 1 \right)$$

Then, if the hypotheses make different predictions about some data D that's been observed, I can go to the relevant analysts and collect their evidence values

$$Z_h^D \equiv \Pr(D \mid h)$$

These evidence values are now *likelihood* values for the hypotheses. The underlying states j have been summed away, leaving the evidence as the relevant quantity. This, by the way, is why evidence values should always *always* be reported as part of the results of a probability computation. They almost never are, to the serious detriment of professional standards of rational inference.

Anyway, the Z 's modulate the π 's to produce

$$\text{MyPosterior (for hypothesis } h) \propto Z_h^D \pi^h \parallel I$$

along with an evidence value $Z_{I(\text{Skilling})}^D$ that I should report in case somebody else wants to assess my judgment. It's all a very clean and consistent methodology.

Although, correctly, we are allowed to make any hypothesis we want, actual data will favour priors that are decently matched. For example, a closed-minded fundamentalist (“God has *told* me that pieces of furniture are 1.287m long”) might not be well able to predict the length of my dining table. On the other hand, an excessively open-minded free thinker (“Hey man, anything goes, OK, say between 1 nanometre and 1000 kilometres”) will distribute his prior beliefs so widely across alternative states that any particular data will seem quite improbable. Between these extremes lie well-judged priors which predict the data tolerably well, without trying to be too precise about it. In fact, this is how good professionals always work. They already know something about the object in question, and use experience and intelligence to distribute their beliefs with good judgment. They can learn more effectively because they start from an appropriate base.

6. ESTIMATION

Usually, we are interested in some overall numerical property $\langle Q \rangle$ of the object in question, rather than the details of the individual probabilities p^j of the alternative states it may occupy. Abstractly, $\langle Q \rangle$ could be anything, but Cox-type consistency arguments applied to a 3-state system show that it can always be transformed onto a linear scale on which

$$\langle Q \rangle = \sum_j Q_j p^j$$

Here, we have Q_j as the property associated with state j alone and, as the choice of notation anticipated, $\langle Q \rangle$ is the mean (a.k.a. expectation) of the overall property. We are free to take non-linear functions of the real number $\langle Q \rangle$, such as $\exp\langle Q \rangle$ and $\langle Q \rangle^2$, but the underlying property is necessarily linear. We can also have $\langle Q^2 \rangle = \sum Q_j^2 p^j$ as a property, and thence acquire the variance

$$\text{var}(Q) = \langle Q^2 \rangle - \langle Q \rangle^2$$

which expresses our uncertainty about the mean value. Higher moments are available too, if we want them.

All this is perfectly standard. There's nothing new, and nothing difficult. The point is that this straightforward linear formulation is not just convenient, but *required*. There is no rational alternative.

One special property is

$$Q_j = \begin{cases} 1 & \text{if } j \text{ is one of the states in set } \Theta, \\ 0 & \text{otherwise.} \end{cases}$$

This has

$$\langle Q \rangle = \sum_{j \in \Theta} p_j$$

which estimates whether the object is in any of the states Θ .

Presentation of a general property Q is straightforward — we just evaluate it for each of n (a dozen or so) Monte Carlo samples of p . This gives us n random samples from the distribution $\Pr(Q)$, from which it's usually adequate to extract the mean and standard deviation. The sampling error decreases as $1/\sqrt{n}$, so is almost always buried beneath the statistical deviation δQ inherent in $\Pr(Q)$. To guard against the small risk that it pokes out, increase n to 100.

If we choose to measure Q , it then becomes just another likelihood factor $(L')_j = Q_j$ with data D' to be factored into the inference analysis. If we choose to measure Q many times, then the accumulation of such factors almost certainly forces the average value of those measurements to be very close to $\langle Q \rangle$. Given such data, hypothesis comparison favours that prior assignment which predicts the long-term averages correctly. In this way, the preferred value of p_j becomes the frequency ratio

$$p^j \sim \frac{\# \text{ times system was in state } j}{\# \text{ trials}}.$$

The old-fashioned “frequentist” paradigm [3] was to *define* probability through agreement with this long-term limit, in defiance of the obvious facts that many objects of interest are not repeatable at all, and no observation can be repeated for ever. I view matters the other way round, as a sanity check. It would be disturbing if long-term frequency ratios did not accord with the predictions from a solidly-based prior.

7. DECISIONS

The “object” we are investigating can include multiple futures.

If I do *this*, the future will bring me to states in set Θ , but if I do *that*, the future will bring me to states in set Φ . Do I do *this*, or do I do *that*?

Either choice will lead to a collapse of the space of possibilities, in this case to Θ or Φ . We can do this rationally by assigning valuations V to the various states. Pessimists reverse the sign and call them “loss functions”. Doing *this* will give me expectation value $\langle V \rangle_{\Theta}$, whereas *that* would give me $\langle V \rangle_{\Phi}$ instead. I will choose whichever option is of most value. Value to *me*, that is, though it is open to me to altruistically incorporate value to others.

Abstractly, V could be anything, but Cox-type consistency arguments applied to a 3-state system show that it can always be transformed onto a linear scale on which

$$V = \sum_j V_j p^j$$

Here, we have V_j as the value of state j alone. We could work in terms of $\exp(V)$ or V^3 or whatever, but the underlying valuation is necessarily linear. In our civilization, value is usually expressed as money, which is an approximation to the linear valuations required by decision-making.

A popular view [4] *defines* probability through monetary ratios in a world of ideal betting, in defiance of the obvious fact that money is not a mathematically-defined quantity so cannot form the logical basis of anything. Betting odds follow probabilistic beliefs, not the other way round, as introspection ought to confirm. I view the betting analogy as a sanity check. It would be disturbing if betting did not approximate to plausible valuations in a world allowing monetary exchange.

There is a warning, though. Decisions need not commute, meaning that they can interfere with each other. The difficulty can already be seen in 3-state systems, but it's more eye-catching in bigger ones. Suppose 9 states are equiprobable, and have the values shown.

7	1	3
1	8	1
5	1	6

An initial decision between rows would select the bottom row, and subsequent decision on columns would select “6”. On the other hand, an initial decision between columns would select the left, and subsequent decision on rows would select “7”. Deciding both together would select the middle “8”, which would presumably have been the actual aim. Ideally, a single final decision should be taken all at once.

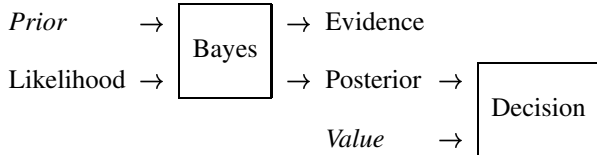
8. PHILOSOPHY

We have reached a very simple and coherent position.

Underlying my (personal) analysis of any object is a probability distribution p^j representing my belief about

which state j it plausibly occupies. I have to start by assigning a (personal) prior distribution π^j . I can observe it through one or more (objective) likelihood functions L_j . I can estimate its (objective) properties through components Q_j . I can assign its (personal) value through valuations V_j .

Probability calculus (forced) tells me how to incorporate data into my beliefs, through simple additions and multiplications only. My (personal) decision to collapse the possible states is made on the basis of maximum value, which is the only place where maximisation enters the framework.



Nothing in this scheme gives any hint of how priors and values should be assigned. We are free to adopt any prior position, because our background information I could legitimately be anything. The language of probability is defined, but what we say in it is arbitrary.

We are also free to assign arbitrary valuations, because our needs could legitimately be anything. The method of decision is defined, but our context for it is arbitrary. Nevertheless, principles can be discerned in the pattern of equations.

To make wise decisions, one should discover what the situation actually is. To discover the situation, one should start with good judgment, and use all relevant data properly. First comes the cool, disinterested search for the truth, undistorted by one's eventual aims. This is the arena of probabilistic analysis. Afterwards and not before comes decision-making, based on good judgment about values, and about choices. Collapsing the space of possibilities should be done as late as possible to avoid early damage. In short, rational thinking yields practical power, and irrational thinking diminishes it. That is a fact too often ignored by wishful thinkers who see only what they *hope* to see. They are, thereby, less effective. Perhaps this should be explained to our political leaders.

9. GEOMETRY

We have already noted that there is really only one sort of probability, and whether it is called prior π^j , likelihood L_j^k , evidence Z^k , posterior P^j or whatever is a mere matter of context. I chose to write probability distributions as superscripted vectors. The discerning reader may already have noticed that there is really only one way of accessing probabilities, and that is through subscripted vectors. Whether these are called likelihoods L_j^\bullet , properties Q_j , valuations V_j or whatever is a mere matter of context. They can all be called "*observables*" — a term deliberately adopted from quantum mechanics where it has practically the same meaning.

Observable access to a probability is always through linear summation and suddenly this looks like the contraction of indices

$$\mathbf{X} \cdot \mathbf{p} = \sum_i X_i p^i$$

in the scalar product of Riemannian geometry. The observable is being written in covariant form X_j , while the probability p^j is contravariant. Moreover, if we were to change coordinates by re-grouping and/or sub-dividing the states, we would find that the scalar product is, correctly, invariant. This is very suggestive. Is there a *metric*, to complete the geometry?

Indeed there is a natural metric. It is easiest to derive between observables, where we have already noted that observable $\langle Q \rangle = \sum Q_i p^i$ carries with it $\langle Q^2 \rangle = \sum Q_i^2 p^i$. Applying the same idea to the observable product AB of two different observables, we have [5]

$$\langle AB \rangle = \sum_i A_i B_i p^i$$

We identify this as the scalar product $\mathbf{A} \cdot \mathbf{B}$, because if we want to assign a geometry at all this is the only definition that can make sense for observables. In Riemannian geometry, the scalar product between covariant vectors defines the contravariant metric g^{ij} ,

$$\mathbf{A} \cdot \mathbf{B} = \sum_{ij} A_i B_j g^{ij}$$

Equating these expressions, we get the metric.

$$g^{ij} = \begin{cases} p^i & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Meanwhile, the corresponding covariant metric g_{ij} is the inverse matrix to the contravariant form, namely

$$g_{ij} = \begin{cases} 1/p^i & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

This varies with position, but we can still use it to define local lengths and angles between contravariant distributions. In particular the length $\delta\ell$ of a small displacement δp is [6]

$$(\delta\ell)^2 = \sum_{ij} (\delta p^i) (\delta p^j) g_{ij} = \sum_i \frac{(\delta p^i)^2}{p^i}$$

This is valid within the simplex of normalised distributions ($\sum p^i = 1$), and within that domain the geodesic (shortest) distance between different probability distributions \mathbf{p} and \mathbf{q} turns out to be

$$\ell(\mathbf{p}, \mathbf{q}) = 2 \arccos \sum_i \sqrt{p_i q_i}$$

Most computational exploration of probability distributions requires control of distance, and this is the natural distance, so it may be useful.

More recently [7], a school of thought has developed the analogous metric

$$g_{ij}(\theta) = \sum_k \frac{1}{p^k} \frac{\partial p^k}{\partial \theta^i} \frac{\partial p^k}{\partial \theta^j}$$

for distributions p parameterised by θ to lie in a sub-manifold of the probability simplex. I have grave doubts about the long-term utility of this, principally on the grounds that its magnitude is dominated by the small-scale differential shape of the manifold, whereas practical inference should be dominated by larger-scale summations. Two distributions can be close in the simplex but far apart on the sub-manifold, simply because the manifold happens to be locally rough. Practical computation will relate them directly, and will not want to follow the intricate small-scale details of a geodesic path confined to such a manifold.

10. OLD STATISTICS

For light relief, let's have some fun with old-fashioned "orthodox" statistics. Consider hypothesis testing. As a student in the 1960s, I was taught [3] to use data to reject a "null hypothesis" H_0 if some test criterion (χ^2 , F , t , ... the choice wasn't always clear) exceeded some plausible bound. The idea, basically, was to start worrying if the likelihood L of the data fell below the 1% quantile of what random variation would expect. Rather grandly, one would then "reject the null hypothesis at the 99% significance level". Yes, but *what then?* Am I supposed to invent new symmetry-breaking physics because I tossed a die and it produced the same face 5 times in a row? There's no point in rejecting a hypothesis if one doesn't have an alternative.

I was given some timings on falling objects, and told to test the null hypothesis that the gravitational acceleration g was 9.81 m s^{-2} . But, if there was one thing I knew with 99.999-recurring percent certainty it was that the real number g was *not* exactly 9.81000 -recurring m s^{-2} . Was I supposed to put something known to be false into my analysis? Would not that enable me to "prove" anything at all?

Anyway, how much data was I supposed to use? There might already be lots of data, or perhaps solid physics, underpinning a good sharp prior. Was I supposed to reject all that because of a 1-in-100 fluke in just the most recent dataset?

Consider parameter estimation. I was taught to do that by inventing (how?) an "unbiased estimator" that would be correct on average, and feeding my data into that. Take this problem. Cans of foodstuff are injected with a preservative which keeps the food fresh for an unknown time T . Thereafter, the food decays exponentially at unit rate per week, $\Pr(t) = e^{-t}$. The lifetimes of various cans are observed as data D_1, D_2, \dots . What is T ? Well, the av-

erage decay time $\langle t \rangle$ is unity, so the obvious unbiased estimator is $\hat{T} = \langle D \rangle - 1$.

Actually, three cans were observed to decay at times 8, 9, 13 weeks. The unbiased estimate is

$$\hat{T} = (8 + 9 + 13)/3 - 1 = 9 \text{ weeks.}$$

Yet we know for certain that $T \leq 8$ because one of the cans had already decayed by then. This estimator is unbiased on average, but it's hopeless for the particular data in hand.

Ordinary probabilistic analysis is properly straightforward. The likelihood factor for the k th observation is

$$\Pr(D^k | T) = H(D^k - T) \exp(T - D^k)$$

where H is the unit step function. For several data, the likelihood factors multiply to

$$\Pr(D | T) = H(D_{\min} - T) \exp \sum (T - D^k)$$

For the particular data quoted, the likelihood is

$$\Pr(D | T) = H(8 - T) \exp(3T - 30)$$

So T is definitely less than 8 weeks, and (unless there is strong prior reason otherwise) probably about $\frac{1}{3}$ week less. Easy!

11. CONCLUSIONS

Probability calculus gives a coherent framework for rational inference. It is a language in which we can submit any "prior" hypothesis to our data, and receive in return a revised "posterior" and the "evidence" value for comparing alternative hypotheses. We can estimate quantities and make decisions, all in the same system.

Applications of any significant size appear to need exponential computing power, but this can be evaded by random sampling. As it happens, programs designed to sample probability distributions randomly also have the exploratory power needed for maximum-value selection.

REFERENCES

- [1] Cox R.T., 1946, Probability, frequency and reasonable expectation, *Am. J. Phys.* **14**, 1–13.
- [2] Jaynes E.T., 2003, *Probability theory: the logic of science*, Cambridge University Press.
- [3] Edwards A.W.F., 1972, *Likelihood*, Cambridge University Press.
- [4] de Finetti B., 1937, La prévision: ses lois logiques, ses sources subjectives, *Ann. Inst. H. Poincaré* **7**, 1–68.
- [5] Wootters W.K., 1981, Statistical distance and Hilbert space, *Phys. Rev. D* **23**, 357–362.
- [6] Fisher R.A., 1956, *Statistical methods and scientific inference*, Oliver and Boyd, London.
- [7] Amari S., 1985, *Differential-geometry methods in statistics*, Lecture notes in statistics, Springer-Verlag.