Statistics between Inductive Logic and Empirical Science

Jan Sprenger

Institut für Philosophie, Universität Bonn, 53113 Bonn, Germany

Abstract

Inductive logic generalizes the idea of logical entailment and provides standards for the evaluation of non-conclusive arguments. A main application of inductive logic is the generalization of observational data to theoretical models. In the empirical sciences, the mathematical theory of statistics addresses the same problem. This paper argues that there is no separable purely logical aspect of statistical inference in a variety of complex problems. Instead, statistical practice is often motivated by decision-theoretic considerations and resembles empirical science.

1 Inductive Logic

Deductive logic explicates the notion of a valid argument and develops a formalism how to discern *valid inferences* – inferences that preserve truth in passing from the premises to the conclusions. Then the premises logically entail the conclusion. Hence, deductive logic studies principles and criteria of truth-preserving inference. It is a formal science in the sense that the meaning of the symbols does not affect soundness or validity of the conclusions.
Inductive logic tries to generalize the idea of logical entailment to inferences where the truth of the premises does not guarantee the truth of the conclusions. Still, the truth of the premises might *indicate* the truth of the conclusion, and it is the point of inductive logic to make the vague and informal notion of truth-indication more precise (Hawthorne 2005). The central concepts become *confirmation* and *evidential support*: it is not asked whether the premises logically entail the conclusion but whether they give good reasons to assert the conclusion and to which degree they support it. In particular, inductive logic is supposed to quantify the effects of observation and measurement on the epistemic status of general hypotheses and theories. Most empirical sciences

Email address: `jan.sprenger@gmx.net` (Jan Sprenger).

infer from data to general hypotheses, and as deductive relations between theory and evidence seldom hold, the degree of support is of particular interest. There inductive logic comes into play, figuring out which hypotheses are best confirmed by the data.

Usually, inductive reasoning in science proceeds along the lines of the mathematical theory of probability. A probabilistic entailment has the general form

$$\phi_1^{x_1}, \ldots, \phi_n^{x_n} \models \psi^y \tag{1}$$

where ψ and ϕ_i denote sentences of a given language and y and x_i denote the corresponding probabilities. In particular, y denotes the posterior probability which the premises – sentences with a given probability – impose on the conclusion.[1] Bayesian inference is naturally embedded into such a framework: Bayes' theorem computes the (posterior) probability of a hypothesis H conditional on evidence E from the prior probability of H and the likelihood of E under the competing hypotheses:

$$P_{\text{new}}(H) := P(H|E) = \frac{P(H)P(E|H)}{P(H)\,P(E|H) + (1 - P(H))\,P(E|\neg H)} \tag{2}$$

The probabilities are interpreted epistemically, i.e. as rational degrees of belief. Hence, Bayesian updating is a special case of probabilistic inference, describing the transformation of subjective degrees of belief in the light of incoming evidence. When applying Bayesianism to modeling confirmation in science, external factors as the practical consequences of a fallacious inference are left out. That might be relevant to an inductive decision theory, but not to inductive *logic* and inductive *inference*. Inductive logic aims at isolating the logical part of inductive reasoning from pragmatic or decision-theoretic factors, in the very same way that is characteristic of deductive logic. Proposals along these lines were recently made by Howson (2003) and Fitelson (2005).

The impact of observation on scientific hypotheses is the core topic of statistical science. In the quantitative sciences, inductive inference is often tantamount to statistical inference. The latter comprises pattern detection as well as model building, hypothesis testing and many other forms of inference. To quote a famous statistician, "the theory of statistics deals in principle with the *general concepts underlying all aspects of such work* [analysis and interpretation of data]".[2] Statistical tools often differ from the tools of inductive logic by their higher mathematical sophistication. But the main difference between Bayesian inductive logic and Bayesian statistics seems to be the different tradition of the two fields: Confirmation theory and inductive logic

[1] See also the work of Hailperin (1996) and members of the PROGIC research group (e.g. Haenni et al. 2008). In their description, the x_i and y can also denote interval-valued probabilities.

[2] Cox (2006, xiii).

were mainly developed in logic and philosophy departments whereas statistics emerged from mathematics and the empirical sciences (e.g. the pioneer statistician R. A. Fisher was a leading geneticist, too). Therefore it is tempting to forge tighter links between inductive logic and statistical inference. Most prominently, the steady rise of subjective Bayesian inference within statistics speaks for such a close connection. Refining the pioneer work of Carnap (1962) and Hintikka (1966), this project was further pursued by many Bayesians in the field, e.g. Festa (1986, 1993) and more recently, Romeijn (2004, 2005). They try to integrate Bayesian statistical inference – and statistics in general – into the framework of a probabilistic inductive logic.

This paper is concerned with the scope of such unifying efforts in statistical science. Do statistical arguments have a logical aspect which is clearly separable from decision theory? Do they proceed along the lines of a probabilistic inductive logic? A position which affirms both questions shall henceforth be called the *logical view of statistical inference*. Apart from the aforementioned authors, this position is most prominently articulated by Howson and Urbach (2005) who argue for a Bayesian inductive logic as the best vehicle of inductive reasoning in science. Indeed, a lot of *frequentist* procedures, i.e. procedures that are justified by means of their favorable long-run properties, can be redescribed as Bayesian procedures with specific prior assumptions. [3] More and more statisticians realize that the alleged objectivity of many frequentist procedures falls prey to implicit prior assumptions that can be expressed in Bayesian terms. With regard to model analysis, Howson and Urbach even write:

> "Unlike the hotchpotch of ad hoc and unjustifiable rules of inference that constitute the classical approach, Bayes's theorem supplies a single, universally applicable, well-founded inductive rule which answers what Brandt calls the 'most important [...] need for integration of [...] model fitting into a coherent whole.' " [4]

Indeed, such a view seems to be attractive in virtue of current statistical practice. Especially in the social sciences, active researchers use statistics packages like SPSS as a "black box" that gives generalizable results out of raw data without further ado. Such mechanical procedures suggest that computer programs perform the "logical" part of statistical inference, transforming and compressing data to a form where we can base actual decisions on them. In such an image, any scientific expertise flows into the specification of prior probabilities. Those recent developments seem to advocate the unification of statistical techniques in the comprehensive framework of probabilistic logic, or more specifically, of Bayesian reasoning. Statistical research could then be

[3] Cf. Berger, Boukai and Wang (1997); Berger (2003).

[4] Howson and Urbach (2005, 236).

described as a mutual adaption of formal theories of inference and statistical practice.

I would like to point out the restrictions to which the logical view of statistics is subjected. Taken as a description of the entire discipline of statistics, it distorts a large part of the actual practice. Section 2 highlights the constraints to which the widely applied frequentist estimation is subjected. For elementary problems, Bayesian estimation offers a way out of the problem, but it relies on the "Perfect Model Scenario": the assumption that the true model is included in the set of candidate models. This assumption becomes less plausible the more complex the situation gets, e.g. in a model selection problem. Under these circumstances, Bayesian probabilities lose their usual meaning as degrees of belief that a particular hypothesis is true, or so I argue (cf. section 3). In particular, we need a new interpretation and justification of Bayesian inference. Those problems of Bayesianism motivate the use of frequentist techniques in complex model selection problems although or all the more because such approaches are hardly separable from a decision-theoretic analysis. Many frequentist techniques are highly sensitive to underlying assumptions so that human expertise and scientific understanding are required for a sensible implementation (cf. section 4). Consequently, I conclude that statistics mainly addresses practical worries about using data in making decisions, predicting events or describing the mechanisms of a system. More precisely, statistics contains a patchwork of different approaches. Choosing one of them is highly sensitive to modeling assumptions, specification of goals, error tolerance etc., and there are no conclusive arguments for a particular method. Hence, comparison of different methods is only possible relative to far-reaching assumptions, blurring the prospects for conceptual unification of statistics. The way how genuinely scientific insights enter the statistical model analysis suggests that statistics resembles an empirical science more than a sophisticated inductive logic. This claim can be substantiated by the numerical turn in statistics: computer-based design of statistical methods and their simulation-based evaluation become more and more important.

2 Parameter estimation

My first example deals with an elementary practice of statistics: parameter estimation. [5] We assume that we have a family of statistical models $\mathcal{M}$ that are parametrized by a parameter ϑ whose true value we do not know. An *estimator* $\hat{\vartheta}$ is a function from the sample space – i.e. the set of possible observations – to the domain of the parameter. This function $\hat{\vartheta}$ is interpreted as a guess about

[5] I do not give any references since the example is contained in many statistics textbooks.

the true value of the parameter ϑ on the basis of the observations. When we estimate the unknown parameter, we would like to get an estimate that is in some sense close to the true value ϑ. In applied sciences, researchers use a standard software like SPSS: for a given set of data, the program yields a "best estimate" of the unknown parameter of interest. Put another way, after feeding in the data together with a family of models we mechanically obtain a best estimate of the parameter, together with some useful risk assessments. Such procedures seem to be tailor-made for integration into the framework of an inductive logic. I would like to stress, however, that the notion of a "best estimate" is far more complicated than we are inclined to think and that a lot of tacit assumptions are made when applying such a procedure. As always, an example will do the best illustration job. Let's begin with a frequentist analysis of the problem because it is the standard method in statistics packages.

Suppose you take part in a TV show where ten random integer numbers $x_1, \ldots, x_n$ are generated from the set $\{0, \ldots, \vartheta\}$, $\vartheta \in \mathbb{N}$.[6] Unfortunately, you do not know the value of ϑ. The better your guess approximates the true value, the more money are you are going to win. There are various reasonable ways to proceed. First, the double of the mean of the ten numbers ($\hat{\vartheta}_1(x) := 2/n \sum x_i$) gives an *unbiased estimate* of the true value ϑ. In other words, $\hat{\vartheta}_1$ is centered around the true value ($\mathbb{E}[\hat{\vartheta}_1] = \vartheta$). But this property does not tell us how strongly $\hat{\vartheta}_1$ will deviate on average from ϑ, i.e. we do not know the variance of $\hat{\vartheta}_1$. Hence, we cannot guarantee that $\hat{\vartheta}_1$ is a sensible estimator. For instance, $\hat{\vartheta}_1$ could be either much greater or much smaller than ϑ in a way that the errors cancel out on average. Second, we could choose the maximum of the ten numbers x_1-x_{10} ($\hat{\vartheta}_2(x) := \max x_i$). This is the smallest value of ϑ that is consistent with the data, and it makes the actually obtained result most likely. Therefore, we call $\hat{\vartheta}_2$ a *maximum likelihood estimator* (MLE). There is, however, an intuitively understandable drawback: for any large value of ϑ, the maximum of the ten random numbers will almost always be smaller than ϑ. In other words: It is quite unlikely that the maximum will actually be attained in a sample of ten random numbers. Hence, the estimator $\hat{\vartheta}_2$ *systematically underestimates* the true value of ϑ.

These considerations highlight the drawbacks of both proposals, $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$. Still, we lack clear-cut criteria which of the two we ought to prefer. Not surprisingly, the answer depends on what "sufficiently close to the true value" means. Assume for example that the amount of money you can win roughly decreases as a quadratical function of the distance between the estimated and the true value. In order words, we assume a *quadratic loss function*. Then the

[6] By a "random integer number", I mean a random variable that is uniformly distributed on $\{0, \ldots, \vartheta\}$.

estimator ought to minimize the *mean quadratic error*, i.e.

$$Q(\hat{\vartheta}) = \mathbb{E}_\vartheta[(\hat{\vartheta} - \vartheta)^2] \tag{3}$$

(Note that this function typically depends on ϑ). It can be shown that this can be decomposed into a sum of two well-known quantities: the *variance* and the square of the *bias* of the estimator (e.g. the difference between the mean of the estimator and ϑ):

$$\begin{aligned} Q(\hat{\vartheta}) &= \mathbb{E}_\vartheta[(\hat{\vartheta} - \vartheta)^2] \\ &= \mathbb{E}_\vartheta[\hat{\vartheta}^2] - \mathbb{E}_\vartheta[\hat{\vartheta}]^2 + \mathbb{E}_\vartheta[\hat{\vartheta}]^2 - 2\vartheta\mathbb{E}_\vartheta[\hat{\vartheta}] + \vartheta^2 \\ &= V_\vartheta[\hat{\vartheta}] - (\mathbb{E}_\vartheta[\hat{\vartheta}] - \vartheta)^2 \end{aligned} \tag{4}$$

If we calculate the mean quadratic errors of our two tentative estimators, we obtain

$$\begin{aligned} Q(\hat{\vartheta}_1) &= \frac{1}{3n}\vartheta^2 \\ Q(\hat{\vartheta}_2) &= \frac{2}{(n+1)(n+2)}\vartheta^2 \end{aligned}$$

which suggests that $\hat{\vartheta}_2$ is superior to $\hat{\vartheta}_1$ for large values of n, in particular for $n = 10$. But this does not imply that the maximum likelihood estimator $\hat{\vartheta}_2$ is *optimal* – we have argued that it is biased. Indeed, we can show that both estimators, $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$, are inferior to the estimator $\hat{\vartheta}_3(x) := (n/(n-1)) \max x_i$ which corrects the bias inherent in $\hat{\vartheta}_2$ dependent on the sample size. $\hat{\vartheta}_3$ is unbiased and has an expected quadratic loss of

$$Q(\hat{\vartheta}_3) = \frac{1}{n(n+2)}\vartheta^2 \tag{5}$$

which is uniformly lower than the expected quadratic loss for both $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$. It can even be shown that $\hat{\vartheta}_3$ is the Q_ϑ-minimizing estimator among all unbiased estimators, i.e. $\hat{\vartheta}_3$ is an *efficient* estimator. This example illuminates that maximum likelihood estimators can be easily outperformed – their pleasant properties for large samples (e.g. asymptotical efficiency) are not brought to bear in the present case. But taking quadratic loss was not at all compulsory in the present case – the Euclidean distance between estimated and true value of ϑ would also have been a reasonable loss measure. That would, however, have yielded different results. More generally, there are lots of potential loss functions which impose different orderings on the available estimation strategies. This implies that “universal principles” like maximum likelihood estimation or efficient estimation cannot guide parameter estimation under all circumstances. For every inference problem, we have to examine which loss function best describes the specific situation – there is no simple epistemological recipe. For instance, most treasuries in the world use conservative estimates of tax revenues whereas climate policy might be based on pessimistic estimates, just because the consequences of overly optimistic action could be devastating. The choice of a loss function is a delicate and non-trivial endeavor which crucially affects the mathematical methods which we apply to determine the best estimator. Of course, it has become conventional to focus on quadratic losses because this is a plausible assumption for many estimation problems and because there exists a full-fledged mathematical theory scrutinizing this particular loss function. But this does not entail that Q_ϑ or any other mathematically beautiful loss function should be regarded as a distinguished measure of *verisimilitude* of point estimators.[7] Some measures may be more tractable and intuitive than others, but the remaining class is still large enough and there is no cogent point in favor of one of these measures apart from purely practical considerations. Estimation problems cannot be described as purely epistemic problems because any metric for measuring "verisimilitude" – the distance between estimated and true value – is conventional. The only way to make sense of a loss function consists in the translation of a mathematical distance into practical, possibly pecuniary losses. In other words, frequentist estimation theory is not separable from a decision-theoretic analysis. The loss function affects whether principles as efficient estimation can be trusted or should be dismissed. Hence, when statistics books advocate a particular estimation method, they make context-sensitive points. Generally, estimators are not comparable in terms of truthlikeness, but only in terms of adequacy for a certain question.

According to the logical view of statistical inference, there is, however, a purely logical aspect of point estimation. Bayesian estimation theory purports to solve the problem – we assign a prior distribution over the potential values of ϑ which we subsequently transform into a posterior distribution. This distribution summarizes our total evidence – prior beliefs as well as learning from data – and serves as the sole input to the decision-theoretic analysis. In the TV show example, we would assign a prior distribution over the natural numbers (recall that $\vartheta \in \mathbb{N}$) and update it on any incoming random number x_i. Hence, the proper logical part of statistical inference seems to be independent of the loss function. At first sight, this step seems to solve the problem, but assigning reasonable prior probabilities is far from trivial. We may be more or less ignorant about the probable range of ϑ, and eliciting a reasonable subjective prior distribution may be impossible.[8] Lacking any further evidence about ϑ, the

[7] A measure of verisimilitude that corresponds to Euclidean distance is adopted by Festa (1986, 277-279), but he does not give any reason apart from tractability and intuitive plausibility (Festa 1993, 38-40).

[8] Of course, we could recourse to objective prior distributions, but then it is questionable we still perform a genuine subjective Bayesian analysis. It is sometimes ar-

assignment of prior beliefs about the value of ϑ is probably neither sensible nor sufficiently motivated. Again, the TV show example may serve as an illustration. Thus, the Bayesian approach involves a substantial idealization with regard to the assignment of prior probabilities. Still, there is a distinct logical aspect in Bayesian estimation. I do not want to admit, however, that this is generalizable to more complex problems. This will come out more clearly in the next section.

3 Bayesian model selection

A lot of the modern debate in statistics and applied sciences focuses on the issues of model selection – how to filter a set of candidate models as to obtain a predictively successful and explanatorily helpful model. To select a model which can be used in further study of the phenomena is such an important decision that we ought to treat it as an integral part of statistical inference. Model selection thus involves the fitting of models to empirical data as well as decisions on the complexity of the model and finding the causally relevant factors. A suitable selection strategy has to evaluate the *model selection uncertainty*, i.e. to account for the problem that the same data which are used for selecting a model family are also used for fitting the model and estimating the parameters which leads to undue optimism towards the selected model. This problem, sometimes also called *selection bias*, is a serious problem for statistical inference:

> “Statisticians admit this privately, but they(we) continue to ignore the difficulties because it is not clear what else could or should be done.” [9]

To base statistical inference on several sensible candidate models is a natural attempt to mitigate the problem. This is the rationale of *model averaging*: Instead of using a single fitted model as the basis of statistical inference, the inference is based on an *average* of all candidate models. Assume, for instance, that ϑ is a quantity of interest and that $\hat{\vartheta}_k$ is the estimator of ϑ in a model M_k where $M_1, \ldots, M_n$ denote the candidate models. Instead of *selecting* a specific value of k, model averaging assigns weights ρ_k to any estimator $\hat{\vartheta}_k$ so that the model-averaged estimator of ϑ takes the form

$$\hat{\vartheta} := \sum_{k=1}^{n} \rho_k \hat{\vartheta}_k \tag{6}$$

gued that as a matter of fact, objective priors tend to minimize risk (cf. Williamson 2007), but this is the very kind of decision-theoretic argument that proponents of the logical view of statistics should avoid.

[9] Chatfield (1995, 421).

By avoiding the selection of a single model, the problem of selection bias seems to be circumvented. However, it is not clear how to assign the weights ρ_k. We might set up a loss function that measures the goodness of $\hat{\vartheta}$ so that the estimation problem becomes an optimization problem over the range of the ρ_k. Apart from the fact that such an analysis poses serious computational problems, it again entangles statistical inference and decision-theoretic analysis. This is exactly what proponents of the logical view of statistics would oppose to. For them, it would be more natural to calculate a posterior distribution over the quantity of interest ϑ. Therefore they normally adhere to *Bayesian model averaging* and assign prior degrees of belief to the candidate models M_k. Then, the distribution of ϑ given the data x can be decomposed into

$$P(\vartheta|x) = \sum_{k=1}^{n} P(\vartheta|M_k, x)P(M_k|x) \tag{7}$$

where the terms on the right hand side are calculated according to Bayes' theorem. We interpret $P(\vartheta|x)$ as the posterior distribution of the quantity of interest, thus separating the logical and the decision-theoretic part of the analysis.

The main virtues of the Bayesian approach consist in conceptual simplicity and unification since Bayes' theorem offers a simple roadmap to the most probable model. Moreover, averaging future predictions by means of the posterior distribution reduces the model selection bias, i.e. the risk that the most probable model is overrepresented in future applications (Kass and Raftery 1995; Wasserman 2000). Other Bayesian approaches to model selection and model averaging include a Laplace approximation of the term $\log P(E|H)$ in Bayes' theorem, the so-called Bayesian Information Criterion (BIC, Schwarz 1978). Yet another strategy examines which models are favored by the data by means of intrinsic or fractional Bayes factors. [10] Most of these approaches have to rely on a prior distribution which may be hard to elicit. How to proceed in the face of ongoing uncertainty about prior weights is the subject of serious statistical research – see Hoeting et al. (1999, 390) for an overview. But that is not the main problem. In Bayesian model analysis we make the general assumption that a true model exists and that it is included in the set of candidate models. Then, it is the goal to pick out the models which are most probable a posteriori. But the claim that models are literally true in science can often be doubted, with consequences for the entire Bayesian approach. For instance, all models have a finite number of components whereas the true model is of infinite order and thus never included in the set of candidate models. Furthermore, scientific modeling necessarily involves some idealization or neglect of external factors so that we cannot speak of literally true models. Bayesian model averaging mitigates this criticism because several different models are taken into account, thus decreasing the risk of being utterly mistaken. But the

[10] See Berger and Perrichi (1996) for a description of this method.

principal problem remains. For example, when the goal is prediction in very complex systems, the Bayesian assumption that the true model is included in the set of candidate models is no longer tenable. Several mistaken predictions need not average out – they might just all be completely wrong. This is especially salient in sciences as geophysics where it is common knowledge that all suggested models are hopelessly wrong due to the enormous complexity of the systems under investigation (e.g. the Earth climate). This is commonly called the failure of the Perfect Model Scenario (cf. Smith 2003). Such a failure is especially plausible if the systems under study have nonlinear or even chaotic character as it frequently occurs in geophysics. If we were to offer bets according to the probabilities of our tentative models of Earth climate or tectonic activity, we would lose money all the time. To make this clear, note that we are actually restricting ourselves to a very small subset of the possible Earth climate models, and it is very unlikely that the true model, granted that it exists at all, is included in this tiny subspace. Therefore the results of Bayesian updating should not necessarily lead to confidence in the most probable model: The relative superiority of a particular model does not make up an argument for truth. We might answer, of course, that a Bayesian does not aim at literal truth, but at "truth up to a reasonable degree of approximation". This move does not help either because we lack a definite metric for deciding when a false model is sufficiently close to the true model in order to count as "approximately true". We encounter the same problem as in section 2: there is no distinct measure of verisimilitude and "closeness to the true value". Moreover, even "approximate truth" may be impossible to achieve. The attractiveness of the Bayesian approach in philosophy of science largely turns on results as the Gaifman-Snir theorem (Gaifman and Snir 1982) – if the true model is among the candidate models and certain regularity conditions hold, the posterior probability of the true model will eventually converge to 1. But in most real-world situations, those idealized conditions do not apply. (Compare that to the TV show example where we *know* that some natural number corresponds to the true value of ϑ!) So how shall we justify a Bayesian approach to the model selection problem? I believe that we should adopt an *instrumental justification*: First, under specific circumstances, Bayesian model selection may select those models that minimize a certain discrepancy measure to the true model – more on that in section 4. Second, the model which is favored by a Bayesian model selection procedure may be the best game in town in so far as it is likely to yield the most accurate predictions among all candidate models. Hence, Bayesian model selection tells us which models we should work with and not which models we shall *believe.* The prior probabilities to the candidate models merely mirror our expectations which of them will perform best for the specific purpose. Working with the "most probable model" is often more likely to minimize actual, real-world losses, thus forging a link to an socioeconomic analysis. Smith (2003) writes:

"It is easily observed that many talented meteorologists dismiss the idea of

> a two-way exchange with socioeconomics out of hand. They state, rather bluntly, that meteorologists should stick to the 'real science' of modeling the atmosphere [...]. Interestingly, if only coincidentally, these same scientists are often those most deeply embedded within PMS." [11]

In such a socioeconomic approach, we focus on the probabilities of observable, economically relevant events that the models try to predict. This is indeed the approach advertised in Roulston et al. (2003) – the agent has to choose between the different forecasts which are obtained from different information consolidation strategies. Then, the agent prices the value of the strategies by means of Monte Carlo resampling from the empirical distribution. Due to the natural connection to decision and expected utility theory, such an analysis could be equally performed in Bayesian terms. Of course, the fact that some models yield better predictions than others does not entail that they should be trusted. But preferring the predictions of a certain model to the predictions of another model may be justified even when we know that all models are wrong because we are interested in their relative real-world value. Here, I do not argue against Bayesian model analysis as a statistical technique, but against an overoptimistic interpretation as the "logic" of statistical inference.

A Bayesian model analysis may have other benefits, too. When we have very large amounts of data and a corresponding number of degrees of freedom, most candidate models will be predictively accurate. The differences between relatively complex and relatively simple models will not be found in the predictive performance but in terms of *interpretability*: A more parsimonious model is explanatorily more valuable because the salient dependencies among the variables are picked out. When all models perform decently in terms of prediction, we would like to select a model that correctly identifies the relevant causal variables and contributes to our scientific understanding. [12] Both theory and simulation results indicate that many Bayesian model selection methods as the BIC tend to *underfit* the true model. In other words, there is a systematical bias in favor of simple and easily interpretable models. So simple models may be favored over complex models although they generally fit the data worse. Besides, complex models are more likely to include spurious covariates. To summarize, Bayesian model selection and Bayesian model averaging are useful tools in statistical model analysis. However, Bayesian probabilities cannot be understood naively, as degrees of belief that a certain model is actually true. This assumption is often untenable in real science. Instead, we better interpret them in an instrumental way – as indicating the relative weight of models when it comes to an evaluation of their predictions. Since PMS often fails, there is no unified logic of statistical model analysis which Bayesian reasoning could reveal, thus falsifying the logical view of statistics. The next

[11] Smith (2003, 236).

[12] Cf. Burnham and Anderson (1998, 171-173).

section introduces alternatives.

4 Model selection: information criteria

The previous section has pointed out virtues and vices of the Bayesian approach to model selection. That gives us a reason to have a closer look at classical frequentist techniques, in particular at the various *information criteria* that generalize the principle of maximum likelihood estimation. Maximum likelihood estimators have several good properties when the sample size increases, e.g. they are asymptotically unbiased, efficient and normal (cf. section 2). In other words, they satisfy several desirable conditions when they are generated by a "very large" amount of data. However, it is not clear how to best apply MLE principles to the problem of model selection. Typically, there are several families of parameterized models with a different degree of complexity. The overall selection error of a model selection criterion consists of two components: the *approximation error* – the error in the selection of a specific model family – and the *estimation error* – the error in fitting the model parameters to the data. Simple MLE estimation advocates the selection of a complex model because the existence of a lot of parameters helps to improve the fit to the data. However, it can be shown that "variance [of an estimator] is directly related to [model] complexity, being far greater for complex models than for simple ones" [13] : Complex models tend to absorb random noise during the parameter fitting process. In other words: if the models are too complex, the parameter values carry little information because the estimation error becomes too big (cf. Zucchini 2000). Hence, a naive application of maximum likelihood principles does not deserve consideration. Therefore, many statisticians pursue a different approach: they choose a function that measures the discrepancy between the true and the candidate model. They impose the single constraint on the true model that it exists somewhere "out there" and that it generates the observed data. Then, they search for the candidate model that minimizes the expected discrepancy to the true model in the same way that a point estimate is supposed to minimize the distance to the true value. These model selection criteria constitute the family of the *minimum total discrepancy* (MTD) criteria. The choice of the metric introduces a conventional element, but most frequently, the discrepancy is measured by means of the *Kullback-Leibler (K-L) divergence* [14]

$$H(f,g) = \int f(x) \log \left(\frac{f(x)}{g_\vartheta(x)} \right) dx \qquad (8)$$

[13] Myung (2000, 195).

[14] This quantity is also called K-L discrepancy, K-L distance, K-L information or relative entropy.

where the integral is taken over the entire sample space and ϑ is the (usually multidimensional) parameter of the approximating model or candidate model g. The standard motivation for using K-L divergence stems from coding theory: H describes the expected loss in transmitting data when an approximating distribution g is used instead of the true distribution f.[15] $\log[f(x)/g_\vartheta(x)]$ measures the information loss between f and g_ϑ for any data point x, and averaging the loss according to $f(x)$ (the relative frequency of a datum x) yields the expected information loss.[16] In fact, the Kullback-Leibler divergence is mathematically tractable and has an enormous significance in the theory of statistics.

The attentive reader will have noticed that in (8), g is not a single model. Rather, it denotes a family of models which are indexed by the parameter ϑ. This mirrors that the candidate models have to be fitted to the data during the model analysis. The point of the MTD criteria is to minimize K-L divergence between the true model f and a family of candidate models g_ϑ. Thus, it is important how to choose the parameter ϑ, and this is the point where the maximum likelihood principle comes in: we estimate ϑ by the maximum likelihood estimate $\hat{\vartheta}$ and then, we try to minimize the *estimated K-L divergence* $\hat{H}(f,g)$, where $\hat{\vartheta}$ is estimated from the actual data y.

$$\begin{aligned}\hat{H}(f,g) &= \int f(x) \log \left(\frac{f(x)}{g_{\hat{\vartheta}(y)}(x)} \right) dx \\ &= \int f(x) \log f(x) dx - \int f(x) \log g_{\hat{\vartheta}(y)}(x) dx \qquad (9)\end{aligned}$$

In a comparison of different candidate models, the first term drops out since it merely depends on the true model f. In other words: When we want to compare the estimated Kullback-Leibler divergence of different candidate models, it is sufficient to focus on the second term in (9). The mutual first terms will cancel out and can be neglected for the purpose of *model comparison*. In other words: We are less interested in the absolute distance between candidate and true model than in the relative truth-distance of the miscellaneous candidate models. This corresponds to our practical worry to select the relatively best candidate model, regardless of whether it is indeed close to the true model (cf. section 3).

The MLE $\hat{\vartheta}(y)$ is based on the actual data. However, it is sensible to average over the sampling distribution in order to eliminate the *sampling error*: If $\hat{\vartheta}(y)$ is based on unrepresentative data y, the estimated K-L divergence will

[15] Cf. Kullback and Leibler (1951); Shannon and Weaver (1949).

[16] Note that the Kullback-Leibler divergence is not symmetric – it would not make sense to average the loss according to the approximating density instead of the true density.

not be close to the true divergence. So, eliminating the sampling error helps to improve the predictive performance of the selected model. Hence,

$$-E_y E_x \left[\log g_{\hat{\vartheta}(y)}(x)\right] = -\int f(y) \left(\int f(x) \log g_{\hat{\vartheta}(y)}(x) dx \right) dy \qquad (10)$$

plus a constant describes the *expected estimated Kullback-Leibler divergence* between f and g. Now, it might look tempting to estimate (10) by means of the log-likelihood of the actual data under g. Here, Akaike's fundamental contribution comes in: He managed to show that *if the set of approximating (candidate) models contains the true model*, the bias of the log-likelihood as an estimator of the expected estimated K-L discrepancy asymptotically corresponds to the number of estimable parameters K of the model family g:

$$\mathrm{AIC}(g, y) := -\log g_{\hat{\vartheta}(y)}(y) + K(g) \qquad (11)$$

is an asymptotically unbiased estimator of (10).[17]

Akaike's result forges a link between the method of maximum likelihood estimation and model selection by means of minimizing K-L divergence. This approach is typical of the MTD criteria. Under which circumstances are we allowed to rely on Akaike's results? First, the method of information criteria is relative to the set of families of candidate models over which the K-L discrepancy is minimized. Without sound scientific judgment, we cannot specify such a *global model*: Both taking too many and too few models into account raises the probability of an erroneous selection. – Second, Kullback-Leibler divergence can in principle be replaced by the Hellinger distance, the Gauss distance, etc. The situation is analogous to the parameter estimation problem: goodness of a point estimator is relative to a loss function, goodness of a model selection criterion is relative to a discrepancy function. Model selection criteria are often tailor-made to a specific discrepancy measure. For instance, the positive properties of Akaike's selection criterion are all derived for Kullback-Leibler divergence. Of course, Kullback-Leibler divergence is a widely accepted and sensible measure of discrepancy for a variety of purposes, but external circumstances might determine which discrepancy measure best describes the actual inference problem. – Third, some MTD criteria tend to focus on the approximation error, e.g. AIC asymptotically eliminates the bias of the MLE, but it scarcely addresses the estimation error. Take the phenomenon of overfitting: If the selected model contained significantly more parameters than the true model, the estimation error would become uncontrollable due to variance explosion. Therefore it is important to check whether some information criteria are particularly prone to such errors and to identify the circumstances that favor such errors. By means of theoretical results, simulations and practice we can determine which criteria tend to overfit or underfit

[17] Cf. Burnham and Anderson (1998, 43-48).

the true model. By now, simulation, practice and theoretical considerations all suggest that the AIC often chooses too complex models, in particular for small samples.[18] This reiterates a previously made point: Unbiasedness of an estimator does not guarantee reliable behavior. It does neither control the variance nor does it imply consistency, i.e. that the estimator converges to the true value with increasing sample size. In our example, AIC is prone to massive overfitting errors. Still, the AIC is of enormous practical value in a variety of real problems and can be refined as to address the above problems. The precise circumstances where AIC should be replaced by a "corrected" version is the subject of a lot of statistical investigation. But in any case, it is way too simplified to state that Akaike's method of model selection gives a general tradeoff between simplicity and data-model fit, as it is often done in philosophy of science.[19]

Indeed, further refinements of the MTD idea address those problems. For instance, Takeuchi generalized Akaike's idea to the situation where the true model is not part of the candidate models and the model families are non-nested, obtaining the so-called Takeuchi Information Criterion (Takeuchi 1976). This step was very important since the restriction that the true model be among the candidate models is a very heavy one – compare the preceding section. Bozdogan's ICOMP criterion (Bozdogan 1988, 2000) is based on a more general notion of model complexity, taking into account not only the number of parameters, but also parameter stability, random error structure, etc. Theoretical results, however, are far too weak to guarantee the overall superiority of a certain model selection criterion. Instead, computer simulations can indicate how the criteria perform in practice. But even those simulations are relative to a very specific choice of circumstances: the sample size, the number of candidate model families, nested vs. non-nested models, etc. So the results of simulations have to be analyzed with great care and caution.[20] The performance of any criterion is relative to the specific circumstances and there is no criterion which could claim general optimality. I believe that we can trace back the lack of such a criterion to the fundamental problem of estimation theory: the lack of a general, loss-independent best estimate. Although some criteria (e.g. Bozdogan's ICOMP) appear to be more generally applicable and more stable than their competitors, it is not justified to conclude that they are uniformly best among all MTD criteria. Therefore, a crucial task for the practicing statistician consists in the deliberation which information criterion might be most adequate for the specific circumstances and the specific goal of inquiry. For instance, the tenability of the vital assumption that the candidate

[18] Cf. Taper (2004, 497-500). Speaking about overfitting is meaningful when applied to the AIC because in Akaike's asymptotical analysis, the true model is a part of the candidate model set.

[19] Cf. Forster and Sober (1994).

[20] Cf. Forster (2000, 202) and Myung (2000, 208).

model is contained in the set of true models sensitively depends on the particular problem. So even if we agree on measuring loss between candidate and true model by means of Kullback-Leibler divergence, there is still room for disagreement on the model selection method. To resolve such a disagreement requires insights from the underlying empirical science as well as mathematical sophistication. In a manifold way, the needs and demands of the field of application and the particularities of the observed data play a crucial role for the model selection procedure. It goes without saying that these findings sharply contradict the logical view of statistics. As the nature of those problems is often quite messy, model selection seems to be closer to empirical science than to inductive logic – we engage in a kind of empirically based engineering work, including the difficult fine-tuning of model selection methods and specific problems. There is no unified scheme of translating scientific insights and observed data to a final decision – a lot hinges on intermediate plausibility arguments. Therefore we cannot neatly decompose MTD model selection into sets of premises and conclusions which are connected by a unified logic of inductive inference. It rather seems that the premises determine which method of inference to apply. Since there is no universal recipe for solving those kind of problems, the reader might by now understand why statistical modeling and model analysis are thought to be an art as well as a science.

5 Model selection: Conclusions

The preceding investigations teach us that simulation- and practice-based evaluation of information criteria go hand in hand with theoretical considerations. In other words: The theoretical properties which we can deduce about a method, be it Bayesian or frequentist, do not decide alone over its adequacy. There are a lot of different error types, and none of the available model selection criteria takes care of all of them. Instead, simulations that resemble typical applications are used in order to study the properties of the proposed criterion. They help us to see whether the criterion is sufficiently robust and applicable in a variety of circumstances. In particular, it is important to check whether the constraints given by the intended application are adequately transformed to the parameters of the simulation (e.g. number of candidate models, linearity, etc.). This evaluation of the model selection criteria has a quasi-empirical character, and due to the increasing computer power, such approaches become more and more popular. In particular, the results of the simulation analysis can lead to the introduction of *ad hoc* information criteria that are adapted to model selection under specific circumstances. [21]

It should be clear by now that a solution of the model selection problem is more

[21] See the SIC(n) family in Taper (2004, 497-500).

than the solution of an intricate mathematical problem. Human expertise is required to decide which form of modeling is most appropriate. It is clear that these priorities must be set by scientists, not by mathematicians. Only they understand the objects of mathematical modeling sufficiently well to assess the adequacy of a particular discrepancy function or the importance of model parsimony in the relevant context. Bayesian model selection is exempted from these reservations only in very idealized circumstances which seldom hold in real science. Formal theories of inductive inference can thus be useful tools, but ultimately, they remain just tools.

The results thus suggest a close collaboration between mathematically minded statisticians and working scientists in order to find the most adequate model selection method in a particular problem. Indeed, this is the route statistics has taken in the last decade, with a lot of statistical literature stemming from researchers that are not located in a mathematics or statistics department. The increased interest in statistical methods among researchers whose primary interests are outside mathematics and statistics shows that a crucial point has been realized: In order to design efficient and helpful statistical methods, scientific understanding and mathematical sophistication have to go hand in hand.

6 Summary

In this paper, I do *not* argue that probabilistic logics in general and Bayesian inductive logic in particular do not have value in statistics. Quite to the contrary, Bayesianism can be extremely useful. Instead, I am concerned with the logical view of statistics that claims the existence of a clearly separable and unified logic of inductive inference whose results serve as a basis for decision-making. Nothing could be further from actual practice. The first case study – parameter estimation – has highlighted some basic problems of frequentist estimation: The quality of an frequentist estimate is relative to the choice of a loss function. This might be resolved in a Bayesian framework, but the more complex case of model selection illuminates problems of Bayesian inference due to the collapse of the Perfect Model Scenario. Bayesian reasoning is then best embedded into a decision-theoretic framework, thus dismissing the logical view of statistical inference. The subsequent discussion of frequentist model selection methods show that satisfactory inference methods are highly sensitive to prior assumptions, goals of inference and substantial scientific insights into the underlying process. Statistical methods are optimal only relative to a variety of external, pragmatic factors: Which types of error do we want to address? What are the practical consequences of a fallacious inference? What is the structure of the random error? Do we have nested or non-nested, linear or non-linear models? And so forth.

It turns out to be impossible to make a neat separation between the logical and the decision-theoretic part in statistical inference. Statistics must not be described as a branch of mathematics that miraculously transforms messy data and vague assumptions into a trustworthy posterior distribution. This would neglect the many uncertainties in the process. Instead, statistics seems to be much closer to empirical work and scientific modeling: The most interesting and fruitful questions about models in science deal with the interplay of scientific inquiry and mathematical modeling. Being able to address such questions with the help of statistical tools has yielded an incredible progress, making statistics an indispensable part of empirical science. These questions are beyond the realms of formal theories of inference as inductive logic.

Acknowledgements

The paper has greatly benefited from extensive discussions with Stephan Hartmann, Jan-Willem Romeijn and Lenonard Smith. Apart from that, I would like to thank the audience at the Third PROGIC Workshop in Canterbury, UK, for helpful discussion and two anonymous referees of this journal for valuable advice.

References

Hirotugu Akaike (1973): “Information Theory as an Extension of the Maximum Likelihood Principle”, in: B. N. Petrov, F. Csaki (ed.), *Second International Symposium on Information Theory*, 267-281. Akademiai Kiado, Budapest.

James O. Berger, Luis R. Perrichi (1996): “The Intrinsic Bayes Factor for Model Selection and Prediction”, *Journal of the American Statistical Association* 91, 109-122.

James O. Berger, Benzion Boukai, Y. Wang (1997): “Unified Bayesian and Frequentist Testing of a Precise Hypothesis (with Discussion)”, *Statistical Science* 12, 133-160.

James O. Berger (2003): “Could Fisher, Jeffreys and Neyman Have Agreed on Testing? (with Discussion)”, *Statistical Science* 18, 1-32.

Hamparsum Bozdogan (1988): “ICOMP: A New Model Selection Criterion”, in: Hans H. Bock (ed.), *Classification and Related Methods of Data Analysis*, 599-608. Elsevier Science, Amsterdam.

Hamparsum Bozdogan (2000): “Akaike’s Information Criterion and Recent Developments in Information Complexity”, *Journal of Mathematical Psychology* 44, 62-91.

Kenneth P. Burnham, David R. Anderson (1998): *Model Selection and Inference: a Practical Information-Theoretic Approach.* Springer, New York.

Rudolf Carnap (1962): *Logical Foundations of Probability.* Second Edition. The University of Chicago Press, Chicago.

C. Chatfield (1995): "Model Uncertainty, Data Mining and Statistical Inference (with discussion)", *Journal of the Royal Statistical Society, series A* 158, 419-466.

David Cox (2006): *Principles of Statistical Inference.* Cambridge University Press, Cambridge.

Roberto Festa (1986): "A measure for the distance between an interval hypothesis and the truth", *Synthese* 67, 273-320.

Roberto Festa (1993): *Optimum Inductive Methods.* Kluwer, Dordrecht.

Branden Fitelson (2005): "Inductive Logic", in: J. Pfeifer, S. Sarkar (ed.), *Philosophy of Science: An Encyclopedia.* Routledge, London.

Malcolm Forster (2000): "Key Concepts in Model Selection: Performance and Generalizability", *Journal of Mathematical Psychology* 44, 205-231.

Malcolm Forster, Elliott Sober (1994): "How to Tell when Simpler, More Unified or Less Ad Hoc Theories Provide More Accurate Predictions", *British Journal for the Philosophy of Science* 45, 1-35.

Haim Gaifman, Marc Snir (1982): "Probabilities over Rich Languages", *Journal of Symbolic Logic* 47, 495–548.

Rolf Haenni, Jan-Willem Romeijn, Gregory Wheeler and Jon Williamson (2008): "Possible Semantics for a Common Framework of Probabilistic Logics", in: V. N. Huynh (ed.), *Proceedings of the International Workshop on Interval/Probabilistic Uncertainty and Non-Classical Logics.* Advances in Soft Computing, Ishikawa.

Theodore Hailperin (1996): *Sentential Probability Logic: Origins, Development, Current Status and Technical Applications.* Lehigh University Press, Bethlehem/PA.

James Hawthorne (2005): "Inductive Logic", in: *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/logic-inductive.

Jaakko Hintikka (1966): "A Two-Dimensional continuum of Inductive Methods", in: Jaakko Hintikka, Patrick Suppes (ed.), *Aspects of Inductive Logic*, 113-132. North-Holland, Amsterdam.

Jennifer Hoeting, David Madigan, Adrian Raftery, Chris Volinsky (1999): "Bayesian Model Averaging: A Tutorial", *Statistical Science* 14, 382-417.

Colin Howson (2003): *Hume's problem.* Oxford University Press, Oxford.

Colin Howson, Peter Urbach (2005): *Scientific Reasoning: The Bayesian Approach.* Third Edition, La Salle: Open Court.

Robert Kass, Adrian Raftery (1995): "Bayes Factors", *Journal of the American Statistical Association* 90, 773-790.

S. Kullback, R. A. Leibler (1951): "On information and sufficiency", *Annals of Mathematical Statistics* 22, 79-86.

Jae Myung (2000): "The Importance of Complexity in Model Selection", *Journal of Mathematical Psychology* 44, 190-204.

Jan-Willem Romeijn (2004): "Hypotheses and Inductive Predictions", *Synthese* 141, 333-364.

Jan-Willem Romeijn (2005): "Theory Change and Bayesian Statistical Inference", *Philosophy of Science* 72, 1174-1186.

M.S. Roulston, D.T. Kaplan, J. Hardenberg, L.A. Smith (2003): "Using medium-range weather forecasts to improve the value of wind energy production", *Renewable Energy* 28, 585-602.

Gideon Schwarz (1978): "Estimating the Dimension of a Model", *Annals of Statistics* 6, 461-464.

Claude Shannon, Warren Weaver (1949): *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Leonard Smith (2003): "Predictability Past, Predictability Present", in: *Proceedings on the ECMWF Seminar on Predictability*, 219-242. ECMWF, Reading.

K. Takeuchi (1976): "Distribution of Information Statistics and a Criterion of Model Fitting", *Suri-Kagaku (Mathematical Sciences)* 153, 12-18.

Mark Taper (2004): "Model Identification from Many Candidates", in: Mark Taper, Subhash Lele (ed.), *The Nature of Scientific Evidence*, 488-524 (with discussion). The University of Chicago Press, Chicago & London.

Larry Wasserman (2000): "Bayesian Model Selection and Model Averaging", *Journal of Mathematical Psychology* 44, 92-107.

Jon Williamson (2007): "Motivating objective Bayesianism: from empirical constraints to objective probabilities", in: William Harper, Gregory Wheeler (ed.), *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr.*, 151-179. College Publications, London.

Walter Zucchini (2000): "An Introduction to Model Selection", *Journal of Mathematical Psychology* 44, 41-61.