Available online at www.sciencedirect.com
ScienceDirect
Journal of Economic Theory 145 (2010) 1757–1775
JOURNAL OF Economic Theory
www.elsevier.com/locate/jet

Simplicity and likelihood: An axiomatic approach ☆

Itzhak Gilboa [a,b,*], David Schmeidler [b,c]

[a] *HEC, Paris, France*
[b] *Tel-Aviv University, Israel*
[c] *The Ohio State University, United States*

Received 2 October 2008; final version received 3 September 2009; accepted 19 February 2010

Available online 27 March 2010

Abstract

We suggest a model in which theories are ranked given various databases. Certain axioms on such rankings imply a numerical representation that is the sum of the log-likelihood of the theory and a fixed number for each theory, which may be interpreted as a measure of its complexity. This additive combination of log-likelihood and a measure of complexity generalizes both the Akaike Information Criterion and the Minimum Description Length criterion, which are well known in statistics and in machine learning, respectively. The axiomatic approach is suggested as a way to analyze such theory-selection criteria and judge their reasonability based on finite databases.

JEL classification: C1; D8

Keywords: Maximum likelihood; Simplicity; Model selection; Akaike Information Criterion; Minimum Description Length; Axioms

1. Introduction

The selection of a theory based on observations is a fundamental problem that cuts across several disciplines. Finding the "right" way to select theories given evidence is at the heart of

☆ We thank Yoav Binyamini, Offer Lieberman, two anonymous referees and the associate editor for comments and suggestions. This project was supported by the Pinhas Sapir Center for Development and Israel Science Foundation Grant Nos. 975/03 and 355/06.

* Corresponding author at: Tel-Aviv University, Israel.
E-mail addresses: tzachigilboa@gmail.com (I. Gilboa), schmeid@tau.ac.il (D. Schmeidler).

0022-0531/$ – see front matter
doi:10.1016/j.jet.2010.03.010

philosophy of science, statistics, and machine learning. It is also highly relevant to rational models of learning, trying to capture the way that rational agents can make sense of the data available to them.

Two fundamental criteria for the selection of theories are simplicity and goodness of fit. The preference for simple theories is well known, and is often attributed to William of Occam (see Russell [11]). While the notion of simplicity is partly subjective and depends on language,[1] it is often surprising how much agreement one finds between the simplicity judgment of different people. For example, most people tend to agree that, other things being equal, a theory with fewer parameters is simpler than a theory with more parameters, or that a theory with a shorter description is simpler than a theory with a longer one. Whereas such claims depend on an agreement about a language, or a set of languages within which simplicity is measured, they do not seem to be vacuous statements. The suggestion that people tend to prefer simple theories to more complex ones can therefore be a meaningful empirical claim.

However, simplicity can only serve as an a priori argument for or against certain theories. How well a theory performs in explaining observed data should certainly also factor into our considerations in selecting theories. Sometimes, one may categorize theories dichotomously into theories that fit the data as opposed to theories that are refuted by the data, and then choose the simplest theory among the former. But in most problems in science, as well as in everyday life, theories are never categorically refuted. There is typically room for a measurement error, or, more generally, for probabilistic prediction. Therefore, a theory typically cannot be refuted by observations. Instead, theories can be ranked according to their goodness of fit, namely, the extent to which they match observations. In particular, the likelihood principle suggests to rank theories according to their likelihood function, that is, the a priori probability that the theory used to assign to the observed data before these data were indeed observed.

Viewed from a statistical point of view, the likelihood principle is a fundamental idea that neatly captures the notion of "goodness of fit" while relying on objective data alone. Choosing the theory that maximizes the conditional probability of the actually observed sample does not rely on any subjective a priori preferences, hunches, or intuitions of the reasoner. But for that very reason, the maximum likelihood principle cannot express preferences for simplicity. Due to this limitation, the applicability of this criterion is restricted to set-ups in which the set of possible theories is restricted a priori to a given class, within which complexity considerations might be ignored. When no such a priori restriction is available, the maximum likelihood principle is insufficient. More explicitly, if one considers all conceivable theories, one will always be able to find a theory that matches the observations perfectly. Such a theory will obtain the maximum conceivable likelihood value of 1, but it is likely to be "overfitting" the data. We tend not to trust a theory that matches the data perfectly if it appears very complex. Thus, maximum likelihood does not suffice to describe the totality of considerations that enter the theory selection process.[2]

We are therefore led to the conclusion that a reasonable criterion for the selection of theories based on observations has to take into account both the likelihood of a theory, or some other measure of goodness of fit, and its simplicity, or some other a priori preference for some theories versus others.[3] Indeed, combinations of likelihood and some measure of complexity are well

[1] See Goodman [4] and Sober [14].

[2] See Gilboa and Samuelson [6] who suggest an evolutionary argument for the preference for simplicity.

[3] Another relevant criterion is the theory's generality. In this paper we ignore the more involved three-way trade-off between goodness of fit, simplicity, and generality. We will only mention in passing that if we "normalize" the theories under comparison so that they have the same scope of applicability, preference for generality can be derived

known in the literature on statistics and on machine learning. Specifically, linear combinations of the logarithm of the likelihood function and a complexity measure appeared in both literatures. Akaike Information Criterion (AIC, Akaike [1]) suggests that, when comparing different statistical models, one adopts a model a that obtains the highest value for

$$\log\big(L(a)\big) - 2k$$

where $L(a)$ is the likelihood function of a, and k is the number of parameters used in model a.[4]

The machine learning literature often adopts Kolmogorov's complexity measure [3,9,10], which suggests that the complexity of a theory be measured by the minimal length of a program (say, a description of a Turing machine) that can be used to generate the theory's predictions. Solomonoff [15] suggested to use such a complexity measure as a basis for a theory of philosophy of science. Related concepts are the Minimal Message Length (MML, Wallace and Boulton [17]) and the Minimum Description Length (MDL) of a theory. Recent applications often trade-off a theory's likelihood with its simplicity by considering criteria of the form

$$\log\big(L(a)\big) - MDL$$

where *MDL* is the Minimum Description Length. (See Wallace and Dowe [18], and Wallace [16] for a more recent survey.)

Clearly, there could be many ways to trade-off a theory's likelihood with its complexity. Indeed, Schwartz Information Criterion (SIC, also known as BIC), suggests that the number of parameters be divided by the logarithm of the number of observations. How should we judge among such criteria? How should we trade off likelihood and complexity?

The present paper addresses this question in an axiomatic way. Our axiomatic approach does not presuppose particular measures of goodness of fit or of likelihood, let alone a particular combination thereof. Rather, we consider an abstract problem in which observations and theories are formal entities that are a priori unrelated, and are also devoid of any explicit content or mathematical structure. In particular, no statistical model is a priori assumed, and no likelihood functions are given. We only assume that a reasoner can rank theories given various databases of observations. Such rankings are modeled as weak orders (binary relations that are complete and transitive), and interpreted as "at least as plausible as" relations. We formulate certain conditions, or "axioms" on these weak orders, which can be viewed as notions of internal consistency: the axioms relate the rankings of theories given different databases of observations. The axioms do not restrict the inferences the reasoner may draw from any particular database, but they do exclude certain patterns of plausibility rankings given different databases. The main result of this paper is that the axioms imply the existence of a statistical model and a constant for each theory, such that for every database, theories are ranked according to the sum of the constant and their log-likelihood function.

Formally, theories are elements of a set $\mathbb{A}$ and observations – of a set $\mathbb{X}$. Neither set is endowed with any mathematical structure, and the two are a priori unrelated. Yet, the intended interpretation of the sets is that elements in $\mathbb{A}$ are distributions or densities on $\mathbb{X}$ with full support. As

from preference for simplicity. Specifically, if theory a is less general than theory b, one may augment a, for instance by using b when a was not defined. The resulting theory would be more cumbersome, and may be less preferred than b based on simplicity considerations.

[4] Observe that, as the sample size, n, grows to ∞, the expression above would typically tend to $-\infty$ for all models. One often divides this expression by n to obtain limits that can be meaningfully compared. Division by n obviously does not alter the ordinal ranking of models.

mentioned above, our formal result may well extend beyond this application. But it is useful to bear it in mind when judging the axioms.

A database I is a function $I : \mathbb{X} \to \mathbb{Z}_+$, (where $\mathbb{Z}_+$ stands for the non-negative integers) with $\sum_{x\in\mathbb{X}} I(x) < \infty$, and $I(x)$ is interpreted as the number of times an observation x has appeared in the database described by I. We assume that, for each such database I, the reasoner has a ranking over theories, $\succsim_I \subset \mathbb{A} \times \mathbb{A}$, where $a \succsim_I b$ is interpreted as "given the observations in database I, theory a is at least as plausible as theory b". When applied to the empty database, $I = 0$, $\succsim_I$ would reflect the reasoner's a priori ranking over the theories, in the absence of any data.

We impose several axioms on the collection of rankings $\{\succsim_I\}_I$, that imply the following representation: for every theory a there exists $w(a) \in \mathbb{R}$ and for every observation x, also a number $v(a, x) \in \mathbb{R}$, such that, for any database I, and any two theories $a, b \in \mathbb{A}$, $a \succsim_I b$ iff

$$w(a) + \sum_{x\in\mathbb{X}} I(x)v(a, x) \geqslant w(b) + \sum_{x\in\mathbb{X}} I(x)v(b, x). \tag{1}$$

In this representation, one may interpret $v(a, x)$ as the log of $\Pr(x|a)$, and then $\sum_{x\in\mathbb{X}} I(x)v(a, x)$ is simply the log-likelihood of the theory a given the database I. If the theories are indeed given as distributions or densities over $\mathbb{X}$, it is natural to assume that the number $v(a, x)$, derived from the reasoner's rankings, will indeed coincide with the logarithm of the probability (or density) of observation x given theory a. However, in a formal model where theories have this additional structure, one would also need an additional assumption that would guarantee this equality.

The constant $w(a)$ reflects an a priori bias for the theory a, and it can be interpreted as a measure of the theory's simplicity, or some other subjective criterion for theory selection. Specifically, if there are finitely many theories, and the reasoner has an a priori subjective probability $p(a)$ that theory a actually governs the data generating process, then the a posteriori ranking of the theories given database I will follow (1) with $w(a) = \log(p(a))$ and $v(a, x) = \Pr(x|a)$ as above. While the Bayesian interpretation is not our preferred one (see the Discussion below), it is compatible with our axioms in many situations.

The axiomatic treatment may serve as a reason to select additive likelihood-complexity tradeoffs such as AIC and MDL, and perhaps to prefer them over other criteria that do not satisfy the axioms. It also serves to clarify the commonalities among simplicity-based criteria and the Bayesian approach to model selection.

This paper may be viewed as a contribution to the axiomatic analysis of statistical techniques. In Gilboa and Schmeidler [8] we provided an axiomatization of kernel estimation of density functions, kernel classification, as well as of maximum likelihood rankings.[5] Billot, Gilboa, Samet, and Schmeidler [2] and Gilboa, Lieberman, and Schmeidler [5] axiomatize kernel estimation of probabilities. One rationale for these papers is the attempt to ground statistical and machine learning methods in axiomatic derivations. The axiomatic approach offers consistency criteria that may help one select theories based on their abstract properties. Such criteria might be of interest especially when finite samples are concerned, and asymptotic behavior may not suffice as the sole guide for the selection of theories.

The rest of this paper is organized as follows. The next section describes the model and the result. The following one is devoted to a general discussion. Proofs and related analysis are to be found in Appendix A.

[5] As explained below, the present paper heavily relies on the results in Gilboa and Schmeidler [8].

2. Model and result

Let $\mathbb{X}$ be the set of (types of) *observations*. The set of *databases* is defined as

$$\mathbb{D} \equiv \left\{ I \;\middle|\; I : \mathbb{X} \to \mathbb{Z}_+,\ \sum_{x \in \mathbb{X}} I(x) < \infty \right\}.$$

A database $I \in \mathbb{D}$ is interpreted as a counter vector, where $I(x)$ counts how many observations of type x appear in the database represented by I.

Algebraic operations on $\mathbb{D}$ are performed pointwise. Thus, for $I, J \in \mathbb{D}$ and $k \geqslant 0$, $I + J \in \mathbb{D}$, and $kI \in \mathbb{D}$ are well-defined. Similarly, the inequality $I \geqslant J$ is read pointwise.

Let $\mathbb{A}$ be the set of *theories*. For $I \in \mathbb{D}$, $\succsim_I \subset \mathbb{A} \times \mathbb{A}$ is a binary relation on theories, where $a \succsim_I b$ is interpreted as "given the database I, theory a is at least as plausible as theory b". The asymmetric and symmetric parts of $\succsim_I$, $\succ_I$ and $\sim_I$, respectively, are defined as usual.

We now turn to describe our conditions. The first three, A1–A3, are "axioms" on the plausibility rankings. They are supposed to suggest appealing properties of the theory-selection criterion.[6] The last two are richness conditions. These conditions have no claim to suggest desirable properties of such criteria. Rather, they are conditions on the set-up of the model needed for our result to hold. For simplicity of notation, we refer to the richness conditions as "A4" and "A5", despite the fact that they are not proposed as "axioms". Correspondingly, axioms A1–A3 are also necessary for the representation (1), while A4–A5 are not. Weakenings or alternatives to A4 and A5 that will give rise to the representation (1) will certainly be of interest. (See the discussion in Appendix A.)

Formally, the only relation between the sets $\mathbb{A}$ and $\mathbb{X}$ is provided indirectly by the set of rankings $\{\succsim_I\}_{I \in \mathbb{D}}$. However, to fix ideas we ask the reader to bear in mind the classical statistical set-up, in which theories are simply distributions (or densities) over observations. We briefly comment on more general set-ups.

A1 Order. For every $I \in \mathbb{D}$, $\succsim_I$ is complete and transitive on $\mathbb{A}$.

A1 is a standard axiom in decision theory. Transitivity is typically considered to be a basic axiom of rationality: if theory a is at least as plausible as theory b, and the latter – at least as plausible as theory c, one would find it hard to argue that c is more plausible than a.

Completeness requires that any two theories can be compared for their plausibility, given any database. Typically, completeness is justified by necessity: once a database is given, the reasoner is asked to make some choice regarding which theory she will use for prediction. The completeness axiom requires that this choice be brought forth and explicitly modeled.

When completeness is applied to a database I consisting of one observation only (that is, $I(x) = 1$ and $I(y) = 0$ for $y \neq x$), it requires that the theories be "about" the observations. In the benchmark case, where theories are distributions over observations, a single observation x naturally induces a ranking of theories based on an a priori bias and the likelihood function. More generally, the completeness axiom still requires that, given a single observation, the reasoner will have a meaningful ranking of the theories. In particular, if the theories are about patterns of observations, rather than about single ones, completeness may not hold.

[6] Thus, our main interpretation is normative. Alternatively, the axioms can also be interpreted descriptively.

A2 Recombination. Suppose that $I, J, K, L \in \mathbb{D}$ are such that $I + J = K + L$. Then there are no $a, b \in \mathbb{A}$, for which $a \succsim_I b$ and $a \succsim_J b$, but $b \succsim_K a$ and $b \succ_L a$.

The essence of this axiom is that evidence gathered from observations is simply accumulated, and that there is no additional learning from the co-occurrence of different observations. Considering a violation of the axiom might serve to explain the type of learning that it rules out.[7] Suppose that a is a common disease, and that b is a rather rare disease. Disease a might manifest itself in symptom x or y, but it is very rare to have both symptoms present. In fact, when both x and y are observed, it is more likely to be disease b rather than a. Next assume that database I consists of two consecutive observations (for the same patient) of symptom x, i.e., $I(x) = 2$, $I(y) = 0$, and database J – of two observations of symptom y, $J(x) = 0$, $J(y) = 2$. Let $K = L$ with $K(x) = K(y) = 1$. Then $I + J = K + L$. Yet, according to our assumptions, each of I, J renders a more likely than b, whereas each of K, L suggests the opposite ranking, in violation of A2.

Thus, the recombination axiom requires that the learning from observations be done case-by-case, where no general picture is allowed to emerge from the totality of the observations. This will be satisfied if the observations are statistically independent. Our model does not assume any probabilistic model, and does not allow us to define independence in the standard statistics sense. But A2 may be viewed as a type of independence, stated in the language of rankings given databases.

The recombination axiom is a version of the "combination" axiom in Gilboa and Schmeidler [8]. The latter implied that $a \succsim_I b$ and $a \succsim_J b$ would necessitate $a \succsim_{I+J} b$. That is, a conclusion (theory a is at least as plausible as theory b) that is warranted given two disjoint databases separately should also be warranted given their union (modeled as $I + J$). This axiom is satisfied by maximum likelihood rankings. But it may be too restrictive when complexity considerations are introduced. Specifically, a simple theory a may be considered more likely than a more complex theory b given each of the databases I and J, separately, even if b fits the data in each database better. But when the two databases are considered in conjunction, the better fit provided by b may overwhelm the complexity considerations, rendering b more plausible than a given $I + J$. The recombination axiom we impose here considers a fixed set of observations, given by $I + J = K + L$. The axiom states that the same set of observations cannot be partitioned twice into two disjoint databases, such that in one partition both databases render a at least as plausible as b, and in the other – one renders b at least as plausible as a, and the other – strictly more plausible.

A3 Archimedean Axiom. Assume that $I, J \in \mathbb{D}$ and $a, b \in \mathbb{A}$ satisfy $b \succ_J a$ and $a \succsim_{J+I} b$. Then for every $K \in \mathbb{D}$ there exists $l \in N$ such that $a \succ_{K+lI} b$.

The antecedent of the Archimedean axiom assumes that, complexity considerations aside, database I renders a more likely than b: starting from $b \succ_J a$, the addition of the observations in I reverses the plausibility ranking. Since complexity considerations and other a priori biases for one theory over another do not change when we compare the database J to the database $J + I$, the switch from b to a can only be attributed to the fact that theory a provides a better fit to the observations in I than does theory b. In this case, the axiom demands that, for every database K,

[7] The following example is based on a suggestion of one of the referees.

the addition of sufficiently many replicas of database I should make a more plausible than b. That is, if a fits the data I better than does b, and we observe more and more databases identical to I, eventually we should prefer theory a to theory b, even if initial data (embodied in K) and complexity considerations originally gave preference to b.

If each theory is a distribution (or a density) over the observations, and the observations are i.i.d., the Archimedean axiom will be satisfied as long as the distributions have full support, that is, as long as no observation can completely refute a theory, that is, drive its likelihood function to zero. More generally, if one already assumes A2, the Archimedean axiom requires that the evidence gathered from different databases will always be comparable.

The last two axioms, or conditions, are not justified on a priori grounds. As mentioned above, they are used only because of the mathematical necessity and may well be weakened or replaced by other axioms. Having said that, we do not find them conceptually objectionable.

The first states that, for every list of four theories and any database, there is a possible continuation of the database that would rank the theories according to the order in the list.

A4 Diversity. For every list (a, b, c, d) of distinct elements of $\mathbb{A}$ and every $J \in \mathbb{D}$, there exists $I \in \mathbb{D}$, $I \geqslant J$ such that $a \succ_I b \succ_I c \succ_I d$. If $|\mathbb{A}| < 4$, the same applies to any permutation of the elements of $\mathbb{A}$.

A4 excludes, for instance, a situation in which one theory is always more plausible than another, regardless of the database. In particular, it excludes from the analysis theories that are tautologically true or tautologically false. More importantly, it does not allow us to include in $\mathbb{A}$ two theories a, b such that a is a generalization of b. Indeed, if each theory is a distribution (or a density) over the observations $\mathbb{X}$, one theory cannot be a generalization of another, and it cannot be always more plausible than another.

The diversity condition also imposes a certain richness condition on the observations. For instance, assume that the observations are the tosses of a coin, $\mathbb{X} = \{0, 1\}$. Suppose that the reasoner believes that the tosses are i.i.d., but does not know the parameter of the coin, so that the set of theories is $\mathbb{A} = [0, 1]$. In this case A4 will not hold, since the likelihood function over $[0, 1]$ has to be single-peaked. In this example there is a continuum of theories, but there aren't sufficiently many observations to allow us to differentiate among them in the sense of A4. More generally, this condition requires that the set of observations be sufficiently rich. If the theories are given by distribution (or density) functions, the diversity condition will be shown to require that the log-distribution (or log-density) of a theory not be weakly dominated by an affine combination of (up to) three other log-distributions (or log-density).

The reason that this condition is required to hold for every four theories but not for more is technical and will be clear in the course of the proof. It will also be clear, as explained in Gilboa and Schmeidler [8], that this condition can be somewhat weakened at the expense of simplicity. In that paper we also show why this axiom is needed: without it, one can construct counter-examples to the representation we seek. The same counter-examples can be used in the present context.

The second richness condition, which is our last condition, requires that for every database and every three theories there is a continuation of the database that renders the three theories equally plausible.

A5 Solvability. For every $\{a, b, c\} \subset \mathbb{A}$, and every $J \in \mathbb{D}$, there exists $I \in \mathbb{D}$, $I \geqslant J$ such that $a \sim_I b \sim_I c$.

The basic import of A5 is that for any three theories there are observations relative to which the "plausibility rankings" are in rational proportions. A counter-example in Appendix A shows why this axiom is necessary, using two observations and log-likelihood functions whose ratios are irrational. We view A5 as a richness condition because, like A4, it takes a given a subset of theories, and requires that there be at least one database that induces a particular ranking over these theories.

We can finally state our main result.

Theorem 1. *Let there be given* $\mathbb{X}$, $\mathbb{A}$, *and* $\{\succsim_I\}_{I\in\mathbb{D}}$ *as above. Assume that* $\{\succsim_I\}_{I\in\mathbb{D}}$ *satisfy* A1–A5. *Then there is a matrix* $v : \mathbb{A}\times\mathbb{X}\to\mathbb{R}$ *and a vector* $w : \mathbb{A}\to\mathbb{R}$ *such that*:

$$(*)\quad \begin{cases} \text{for every } I\in\mathbb{D} \text{ and every } a,b\in\mathbb{A}, \\ a\succsim_I b \quad \text{iff} \quad w(a)+\sum_{x\in\mathbb{X}} I(x)v(a,x) \geqslant w(b)+\sum_{x\in\mathbb{X}} I(x)v(b,x). \end{cases}$$

Furthermore*, in this case the matrix* v *and the vector* w *are unique in the following sense*: (v, w) *and* (u, y) *both satisfy* $(*)$ *iff there are a scalar* $\lambda > 0$, *a matrix* $\beta : \mathbb{A}\times\mathbb{X}\to\mathbb{R}$ *with identical rows* (*i.e., with constant columns*) *and a number* δ *such that* $u = \lambda v + \beta$ *and* $y = \lambda w + \delta$.

Observe that, in the tradition of axiomatizations in decision theory, the representation theorem above only suggests a possible representation. A reasoner whose rankings $\{\succsim_I\}_{I\in\mathbb{D}}$ satisfy our axioms can be thought of *as if* she had a likelihood function (whose logarithm is given by v) and a simplicity measure (given by w) such that she prefers theories with higher values of the sum of the log-likelihood and the simplicity measure. If it so happens that the theories involved are a priori given by a statistical model, so that a likelihood function $l(a|x)$ exists for $a\in\mathbb{A}$ and $x\in\mathbb{X}$, it does not follow that $v(a, x) = \log(l(a|x))$. Indeed, since the axioms make no reference to the likelihood function $l(a|x)$, such a conclusion would be impossible. To derive it, one has to impose additional axioms, relating the relations $\{\succsim_I\}_{I\in\mathbb{D}}$ to the supposedly given $\{l(a|x)\}_x$.

The situation is akin to Savage's derivation of a Bayesian prior: Savage's [12] axioms imply that there exists a probability measure such that the decision maker behaves in accordance with it (via the expected utility formula). If we observe a decision maker who faces a roulette wheel with given, objective probabilities, it stands to reason that her subjective prior would coincide with the measure governing the wheel's behavior. Yet such a conclusion requires an additional assumption and does not follow from the representation theorem itself.

Much of the appeal of Savage's theorem is in that it does not assume an objectively given probability measure, but derives one from preferences. Thus he defines subjective probabilities where no objective probabilities are given. By the same token, our theorem can be said to derive a statistical model (or a likelihood function) even when such a model is not a priori given.

3. Discussion

3.1. The recombination axiom

The statement of the recombination axiom (A2) might bring to mind Simpson's paradox [13], which appears to constitute a violation of the axiom. Consider, for example, the famous Berkeley Sex Bias Case, in which the percentage of men admitted to graduate school is higher than the percentage of women admitted, while the converse is true for each department separately.

(Historically, the converse was true in *almost* all departments.[8]) For simplicity, assume that there are only two departments. In this case, splitting the database by departments would yield two databases (say, I and J) in each of which women appear to be favored to men. By contrast, splitting the same database randomly would yield two other databases (say, K and L), each supporting the opposite conclusion.

However, this application of our model is inappropriate, because the single observations are not directly related to the theories discussed. In fact, in this example even the completeness axiom is problematic: given a single case, be it of a man or a woman, admitted or not, it is not at all obvious how one should rank two theories such as "women are favored at admission" vs. "men are favored at admission". These theories are about comparisons of *sets* of observations (to be precise, comparisons of percentages of admitted applicants within two sub-populations), and they do not directly say anything about a particular observation.

To deal with the Berkeley Sex Bias Case, one would have to consider "observations" that are directly relevant to the theories. For example, an observation might be a pair of candidates that are similar in all respects apart from their gender, one of whom was admitted by a certain department and the other – denied admission by the same department. Such an observation would indeed constitute a direct evidence of unequal treatment of the genders. But it is easy to see that Simpson's paradox cannot be replicated using such observations of disjoint pairs, as the paradox relies on unequal proportions of women and men applying to different departments.

Similar difficulties with the recombination axiom might arise when one considers various theories that are not about specific observations, but rather about patterns of observations. For instance, if one is to judge whether a sequence of observations is random, one may easily construct counter-examples to the recombination axiom. Again, in such examples the theories discussed do not say anything about specific observations, only about patterns thereof. This is highlighted by similar difficulties with the completeness axiom applied to databases of single observations. Having but a single observation, one cannot rationally judge whether it comes from a random sequence or not.

To conclude, our model should only be applied to theories and observations that are directly related, in the sense that every theory is relevant to every observation. Differently put, every single observation should have meaningful implications about the plausibility of the theories. When attention is restricted to such applications, the completeness axiom is not too demanding, and the recombination axiom appears reasonable.

3.2. Methods of classical statistics

It appears that maximum likelihood is a reasonable criterion only when the set of theories is a priori restricted in one way or another. For instance, one may face a regression problem and consider only linear or quadratic theories. But in this case the set of theories under discussion is subjectively chosen. That is, the model does not purport to explain why the particular set of theories – say, linear – was chosen to begin with. Assuming the model as given, likelihood maximization offers an objective ranking of theories. But the choice of the model itself remains subjective, and sometimes arbitrary.

Statistical theory offers a variety of tools to cope with the problem of overfitting data as a result of likelihood maximization. The trade-off between a good fit and the theory's complexity is

[8] See http://en.wikipedia.org/wiki/Simpson's_paradox#_note-3.

familiar from model selection criteria in parametric set-ups (such as adjusted R^2, LASSO, Ridge Regression, and others) as well as in non-parametric set-ups (Akaike Information Criterion, BIC, etc.). The present paper addresses this question axiomatically, describing an inductive learning process that does not impose arbitrary restrictions on the set of theories.

3.3. Bayesian analysis

As mentioned in the Introduction, the ranking by (1) can be interpreted as a Bayesian ranking where $w(a)$ is taken to be the logarithm of theory a's prior probability. However, several distinctions should be borne in mind. First, the numbers $w(a)$ in our set-up are not unique. It is readily seen that they can all be multiplied by a positive constant (alongside the numbers $v(a,x)$) without changing the rankings in (1). Hence, a reasoner who satisfies our axioms can be viewed as Bayesian, but her Bayesian beliefs are not uniquely determined by her rankings $\{\succsim_I\}_{I\in\mathbb{D}}$.

Among the pieces of information that are missing in $\{\succsim_I\}_{I\in\mathbb{D}}$ in order to determine the reasoner's prior probability are the rankings of subsets of theories. A Bayesian reasoner, who has a prior over the space of theories, has a prior probability for every measurable subset of theories. By contrast, our reasoner is only assumed to rank specific theories.

3.4. The measurement of complexity

The measurement of complexity is not a trivial issue. It is very appealing to use some notion of Kolmogorov's complexity, namely the length of the minimal program that implements a theory. But the minimal description length of a theory gives equal weight to bits that describe the algorithm of the program and to bits that describe arbitrary parameters. For instance, the MDL of the theory $y = 1.30972x$ is much higher than the MDL of the theory $y = 2x$. For applications to everyday human reasoning, as well as to scientific reasoning in the social sciences, a "simple" parameter such as 2 need not have any privileged status as compared to a "complicated" parameter such as 1.30972. Differently put, if the bits needed to describe 1.30972 were used to encode logical computation steps, one may have a theory that is much more complicated than the linear relationship $y = 1.30972x$. This suggests that the length of the description of a program in bits, including all numerical parameters, isn't an intuitive measure of the theory's complexity. The appropriate choice of a measure of complexity is beyond the scope of the axiomatic investigation taken in this paper.

Appendix A. Proofs and related analysis

A.1. A basic result

We will rely on the following result, which appears in Gilboa and Schmeidler [7,8]. To state it, we first define a matrix $v : \mathbb{A}\times\mathbb{X}\to\mathbb{R}$ to be *diversified* if there are no elements $a,b,c,d\in A$ with $b,c,d\neq a$ and $\lambda,\mu,\theta\in\mathbb{R}$ with $\lambda+\mu+\theta=1$ such that $v(a,\cdot)\leqslant\lambda v(b,\cdot)+\mu v(c,\cdot)+\theta v(d,\cdot)$. That is, v is diversified if no row in v is dominated by an affine combination of three (or fewer) other rows. The axioms used for the theorem are:

A1* Order. For every $I\in\mathbb{D}$, $\succsim_I$ is complete and transitive on $\mathbb{A}$.

A2* Combination. For every $I,J\in\mathbb{D}$ and every $a,b\in\mathbb{A}$, if $a\succsim_I b$ ($a\succ_I b$) and $a\succsim_J b$, then $a\succsim_{I+J} b$ ($a\succ_{I+J} b$).

A3* Archimedean Axiom. For every $I, J \in \mathbb{D}$ and every $a, b \in \mathbb{A}$, if $a \succ_I b$, then there exists $l \in N$ such that $a \succ_{lI+J} b$.

A4* Diversity. For every list (a, b, c, d) of distinct elements of $\mathbb{A}$ there exists $I \in \mathbb{D}$ such that $a \succ_I b \succ_I c \succ_I d$. If $|\mathbb{A}| < 4$, then for any strict ordering of the elements of $\mathbb{A}$ there exists $I \in \mathbb{D}$ such that $\succ_I$ is that ordering.

Theorem 2. *Let there be given $\mathbb{X}$, $\mathbb{A}$, and $\{\succsim_I\}_{I\in\mathbb{D}}$ as above. Then the following two statements are equivalent*:

(i) $\{\succsim_I\}_{I\in\mathbb{D}}$ *satisfy* A1*–A4*;
(ii) *There is a diversified matrix $v : \mathbb{A} \times \mathbb{X} \to \mathbb{R}$ such that*:

$$(**) \quad \begin{cases} \text{for every } I \in \mathbb{D} \text{ and every } a, b \in \mathbb{A}, \\ a \succsim_I b \quad \text{iff} \quad \sum_{x\in\mathbb{X}} I(x)v(a,x) \geqslant \sum_{x\in\mathbb{X}} I(x)v(b,x). \end{cases}$$

***Furthermore**, in this case the matrix v is unique in the following sense*: *v and u both satisfy* (**) *iff there are a scalar $\lambda > 0$, a matrix $\beta : \mathbb{A} \times \mathbb{X} \to \mathbb{R}$ with identical rows* (*i.e., with constant columns*) *such that $u = \lambda v + \beta$.*

A.2. Proof of Theorem 1

The strategy of the proof is as follows. We define a set of auxiliary relations, $\{\succsim'_I\}_I$ on $\mathbb{A}$, interpreted as follows: $a \succsim'_I b$ suggests that the observations contained in I are at least as probably under a than under b. Thus, if we were to ignore complexity considerations or other a priori biases for one theory over the other, we would expect a to be more plausible than b given I. The relation $\succsim'_I$ will correspond to the summation of the v entries in our representation. That is, $a \succsim'_I b$ will turn out to be equivalent to

$$\sum_{x\in\mathbb{X}} I(x)v(a,x) \geqslant \sum_{x\in\mathbb{X}} I(x)v(b,x)$$

which is the numerical representation we seek if the w's are all set to zero.

The first step in the proof consists of showing that the relations $\{\succsim'_I\}_I$ satisfy the conditions of Theorem 2. This identifies the matrix v up to the transformations allowed by Theorem 2, namely, up to addition of constants to columns and multiplication of the entire matrix by a positive number. We fix one such representing matrix v. This step does not make use of axiom A5.

The next step in the proof is to show that for every two theories a, b there exists a number α^{ab}, with $\alpha^{ba} = -\alpha^{ab}$, such that, for every I, $a \succsim_I b$ iff

$$\alpha^{ab} + \sum_{x\in\mathbb{X}} I(x)v(a,x) \geqslant \sum_{x\in\mathbb{X}} I(x)v(b,x),$$

which is the desired representation for the case of two theories. Finally, the we wish to prove that for each theory a there exists a number $w(a)$ such that, for every a, b, $\alpha^{ab} = w(a) - w(b)$.

A.2.1. Step 1: The matrix v

For $a, b \in \mathbb{A}$ and $I \in \mathbb{D}$, define $a \succ'_I b$ if there exists $J \in \mathbb{D}$ such that $b \succsim_J a$ and $a \succ_{J+I} b$. That is, $a \succ'_I b$ if the evidence contained in I is sufficient to reverse the ordering between a and b.

Lemma 1. *For $a, b \in \mathbb{A}$ and $I \in \mathbb{D}$, it is impossible that both $a \succ'_I b$ and $b \succ'_I a$.*

Proof. Assume, to the contrary, that there are $J, K \in \mathbb{D}$ such that $b \succsim_J a$, $a \succ_{J+I} b$, $a \succsim_K b$, and $b \succ_{K+I} a$. Since $J + (K + I) = (J + I) + K$, this contradicts A2. □

Lemma 2. *For $a, b \in \mathbb{A}$ and $I \in \mathbb{D}$, if there exists $J \in \mathbb{D}$ such that $b \succ_J a$ and $a \succsim_{J+I} b$, then $a \succ_{J+2I} b$.*

Proof. If not, $b \succsim_{J+2I} a$, and then by defining $K = L = J + I$ and $I' = J + 2I$, we obtain $a \succsim_K b$, $a \succsim_L b$, $b \succsim_{I'} a$, $b \succ_J a$ while $K + L = I' + J = 2J + 2I$, a contradiction to A2. □

Lemma 3. *For $a, b \in \mathbb{A}$ and $I \in \mathbb{D}$, $a \succ'_I b$ iff there exists $J \in \mathbb{D}$ such that $b \succ_J a$ and $a \succsim_{J+I} b$.*

Proof. Assume first that there exists $J \in \mathbb{D}$ such that $b \succ_J a$ and $a \succsim_{J+I} b$. If $a \succ_{J+I} b$, then $a \succ'_I b$ follows from the definition of $\succ'_I$. Otherwise, $a \sim_{J+I} b$. Define $J' = J + I$, and note that $b \succsim_{J'} a$. But Lemma 2 implies that $a \succ_{J'+I} b$, which yields $a \succ'_I b$.

Conversely, assume that $a \succ'_I b$. By A4 there exists L such that $b \succ_L a$. By A3, there exists k such that $a \succ_{L+kI} b$. Let k' be the minimal $k \geqslant 1$ such that $a \succsim_{L+kI} b$ and define $J = L + (k - 1)I$. □

Define, for $a, b \in \mathbb{A}$ and $I \in \mathbb{D}$, $a \sim'_I b$ if neither $a \succ'_I b$ nor $b \succ'_I a$. Clearly, $\sim'_I$ is reflexive and symmetric. We observe the following.

Lemma 4. *For $a, b \in \mathbb{A}$ and $I \in \mathbb{D}$, the following are equivalent*:

(i) $a \sim'_I b$,

(ii) *for every $J \in \mathbb{D}$*

$$a \succsim_J b \quad \Leftrightarrow \quad a \succsim_{J+I} b,$$

(iii) *for every $J \in \mathbb{D}$*

$$a \succsim_J b \quad \Leftrightarrow \quad a \succsim_{J+I} b$$

and

$$b \succsim_J a \quad \Leftrightarrow \quad b \succsim_{J+I} a.$$

Proof. We prove that (i) ⇒ (iii) ⇒ (ii) ⇒ (i). Since (iii) ⇒ (ii) is obvious, only two steps are needed.

To prove that (i) ⇒ (iii), assume that $a \sim'_I b$. Consider $J \in \mathbb{D}$. If $a \succsim_J b$ but $a \succsim_{J+I} b$ fails to hold, then $b \succ_{J+I} a$ and $b \succ'_I a$ by definition of $\succ'_I$, contradicting $a \sim'_I b$. If $a \succsim_{J+I} b$ but $a \succsim_J b$ doesn't hold, we have $b \succ_J a$ and then Lemma 3 implies that $a \succ'_I b$, again a contradiction. Similarly, $b \succsim_J a \Leftrightarrow b \succsim_{J+I} a$.

To prove that (ii) $\Rightarrow$ (i), assume that for every $J \in \mathbb{D}$ we have $a \succsim_J b \Leftrightarrow a \succsim_{J+I} b$. If $a \sim'_I b$ does not hold, then either $a \succ'_I b$ or $b \succ'_I a$. If $b \succ'_I a$, by definition of $\succ'_I$ there exists J with $a \succsim_J b$ but $b \succ_{J+I} a$, contradicting $a \succsim_J b \Rightarrow a \succsim_{J+I} b$. If, however, $a \succ'_I b$, by Lemma 3 there exists J such that $b \succ_J a$ and $a \succsim_{J+I} b$, a contradiction to $b \succsim_{J+I} a \Rightarrow b \succsim_J a$. □

Lemma 5. *For $a, b \in \mathbb{A}$ and $I \in \mathbb{D}$, the following are equivalent*:

(i) $a \succ'_I b$,
(ii) *there exist $J \in \mathbb{D}$ and $k \geqslant 1$ such that $b \succsim_J a$ and $a \succ_{J+kI} b$*,
(iii) *there exist $J \in \mathbb{D}$ and $k \geqslant 1$ such that $b \succ_J a$ and $a \succsim_{J+kI} b$*,
(iv) *for every $J \in \mathbb{D}$ there exists $k \geqslant 0$ such that for every $l \geqslant 0$*

$$a \succ_{J+lI} b \quad \Leftrightarrow \quad l \geqslant k.$$

Proof. We show that (i) is equivalent to each of (ii), (iii), and (iv).

We begin with (i) $\Leftrightarrow$ (ii). If (i) holds, then (ii) holds for $k = 1$. Conversely, if (ii) holds, let $l = \min\{l \mid a \succ_{J+lI} b\}$, where $l > 0$ because $b \succsim_J a$. Denoting $J' = J + (l-1)I$ we have $b \succsim_{J'} a$ but $a \succ_{J'+I} b$, that is, $a \succ'_I b$.

The proof that (i) $\Leftrightarrow$ (iii) is almost identical, defining $l = \min\{l \mid a \succsim_{J+lI} b\}$ and invoking Lemma 3.

We now show (i) $\Leftrightarrow$ (iv). Assume (i) holds. Given J, consider $N = \{l \geqslant 0 \mid a \succ_{J+lI} b\}$. By A3, $N \neq \emptyset$. Let k be the minimal element in N. If, for $l > k$, $b \succsim_{J+lI} a$, then, by the implication (iii) $\Rightarrow$ (i), we obtain $b \succ'_I a$, a contradiction to Lemma 1. Hence $a \succ_{J+lI} b$ iff $l \geqslant k$.

Conversely, assume that (iv) holds. By A4 there exists J such that $b \succ_J a$. Let k be defined by (iv), and use the implication (ii) $\Rightarrow$ (i). □

Define $a \succsim'_I b$ if $a \succ'_I b$ or $a \sim'_I b$.

Lemma 6. *For $a, b \in \mathbb{A}$ and $I, J \in \mathbb{D}$*

(i) $a \succ_J b$ *and* $a \succsim'_I b$ *imply* $a \succ_{J+kI} b$ *for all* $k \geqslant 1$,
(ii) $a \succsim_J b$ *and* $a \succ'_I b$ *imply* $a \succ_{J+kI} b$ *for all* $k \geqslant 1$,
(iii) $a \sim_J b$ *and* $a \sim'_I b$ *imply* $a \sim_{J+kI} b$ *for all* $k \geqslant 1$,
(iv) $a \succsim_J b$ *and* $a \succsim'_I b$ *imply* $a \succsim_{J+kI} b$ *for all* $k \geqslant 1$,
(v) $a \sim_J b$ *and* $a \sim_{J+kI} b$ *for some* $k \geqslant 1$ *imply* $a \sim'_I b$.

Proof. (i) Assume $a \succ_J b$ and $a \succsim'_I b$. If for some $k \geqslant 1$, $b \succsim_{J+kI} a$, then Lemma 5 ((iii) $\Rightarrow$ (i)) implies that $b \succ'_I a$, a contradiction.

(ii) If $a \succ_J b$, the conclusion follows from (i). Assume, then, that $a \sim_J b$ and $a \succ'_I b$. By Lemma 5 ((ii) $\Rightarrow$ (i)) we know that $a \succsim_{J+kI} b$ for all $k \geqslant 1$. Also, Lemma 5 ((i) $\Rightarrow$ (iv)) implies that there exists $k \geqslant 1$ such that for every $l \geqslant 0$, $a \succ_{J+lI} b \Leftrightarrow l \geqslant k$ and therefore $a \sim_{J+lI} b \Leftrightarrow l < k$. If $k > 1$, consider J, $I' = J + kI$, $K = J + I$, and $L = J + (k-1)I$. Observe that $J + I' = K + L = 2J + kI$. Moreover, $a \sim_J b$, $a \sim_K b$, $a \sim_L b$, but $a \succ_{I'} b$, in contradiction to A2.

(iii) follows from Lemma 4.

(iv) follows from (i)–(iii).

(v) follows from (ii). □

We now show that $\{\succsim'_I\}_I$ satisfy axioms A1*–A4* of Theorem 2.

Lemma 7. *For every $I \in \mathbb{D}$, $\succsim'_I$ is a weak order.*

Proof. Completeness of $\succsim'_I$ follows from its definition. We need to prove transitivity. Assume that $a, b, c \in \mathbb{A}$ satisfy $a \succsim'_I b$ and $b \succsim'_I c$, and show $a \succsim'_I c$. We distinguish between four cases:

Case 1: $a \succ'_I b$ and $b \succ'_I c$.

By A4, there exists J such that $c \succ_J b \succ_J a$. Since $a \succ'_I b$, by Lemma 5 there exists k_1 such that $a \succ_{J+lI} b$ for $l \geqslant k_1$. Similarity, $b \succ'_I c$ implies that there exists k_2 such that $b \succ_{J+lI} c$ for $l \geqslant k_2$. Hence, there exists l (for instance, $l = \max(k_1, k_2)$) such that $a \succ_{J+lI} b \succ_{J+lI} c$, hence $a \succ_{J+lI} c$. By Lemma 5, $a \succ'_I c$.

Case 2: $a \succ'_I b$ and $b \sim'_I c$.

By A4, there exists J such that $b \succ_J c \succ_J a$. Let k be such that $a \succ_{J+kI} b$. By Lemma 4, $b \sim'_I c$ and $b \succ_J c$ imply that $b \succ_{J+kI} c$. By transitivity, $a \succ_{J+kI} c$, and $a \succ'_I c$ follows from Lemma 5.

Case 3: $a \sim'_I b$ and $b \succ'_I c$.

By A4, there exists J such that $c \succ_J a \succ_J b$. Let k be such that $b \succ_{J+kI} c$. By Lemma 4, $a \sim'_I b$ and $a \succ_J b$ imply that $a \succ_{J+kI} b$. Hence $a \succ_{J+kI} c$, and $a \succ'_I c$ follows as above.

Case 4: $a \sim'_I b$ and $b \sim'_I c$.

If $a \succ'_I c$, then applying Case 2 (with the roles of b and c reversed) implies $a \succ'_I b$, a contradiction. Similarly, $c \succ'_I a$ would imply $c \succ'_I b$. □

Lemma 8. $\{\succsim'_I\}_I$ *satisfy the Combination Axiom* A2*.

Proof. We need to show that, for every $I, J \in \mathbb{D}$ and every $a, b \in \mathbb{A}$, if $a \succsim'_I b$ ($a \succ'_I b$) and $a \succsim'_J b$, then $a \succsim'_{I+J} b$ ($a \succ'_{I+J} b$).

Assume first that $a \sim'_I b$ and $a \sim'_J b$. In this case, Lemma 4 implies that, for every K, $a \succsim_K b \Leftrightarrow a \succsim_{K+I} b$ and $a \succsim_K b \Leftrightarrow a \succsim_{K+I} b$. We wish to show that, for every $K \in \mathbb{D}$, $a \succsim_K b \Leftrightarrow a \succsim_{K+I+J} b$, thus establishing (by Lemma 4 again) that $a \sim'_{I+J} b$.

Let there be given such K. If $a \succsim_K b$, we have $a \succsim_{K+I} b$, and, by considering $K' = K + I$, also $a \succsim_{K+I+J} b$. Conversely, if $a \succsim_{K+I+J} b$ but $a \succsim_K b$ fails to hold, we have $b \succ_K a$. In this case $b \succ_{K+I} a$ (or else $a \succ'_I b$) and then also $b \succ_{K+I+J} a$ (otherwise $a \succ'_J b$), a contradiction. It follows that the combination axiom holds in this case.

We now turn to the case in which one of the relations $a \succsim'_I b$ and $a \succsim'_J b$ is strict. Without loss of generality, assume that $a \succ'_I b$. Hence there exists $K \in \mathbb{D}$ such that $b \succsim_K a$ but $a \succ_{K+I} b$. If $b \succsim_{K+I+J} a$, then $b \succ'_J a$ by Lemma 3. Hence, $a \succ_{K+I+J} b$. Combined with $b \succsim_K a$, this implies $a \succ'_{I+J} b$. □

Lemma 9. $\{\succsim'_I\}_I$ *satisfy the Archimedean Axiom* A3*.

Proof. We need to show that, for every $I, J \in \mathbb{D}$ and every $a, b \in \mathbb{A}$, if $a \succ'_I b$, then there exists $l \in N$ such that $a \succ'_{lI+J} b$. Consider K with $b \succ_K a$. If $a \succsim_{K+J} b$, then by Lemma 6(ii) (for $k = 1$) we have $a \succ_{K+J+I} b$, and it follows that $a \succ'_{I+J} b$, i.e., the conclusion is obtained for $l = 1$. Otherwise, we have $b \succ_{K+J} a$. In this case, apply Lemma 5 ((i) ⇒ (iv) and $J' = K + J$) to conclude that there exists $l \geqslant 1$ such that $a \succ_{K+J+lI} b$, which, combined with $b \succ_K a$, implies that $a \succ'_{lI+J} b$. □

Lemma 10. $\{\succsim'_I\}_I$ *satisfy the Diversity Axiom* A4*.

Proof. Assume first that $|\mathbb{A}| \geqslant 4$. (The proof for the case $|\mathbb{A}| < 4$ is identical.) We need to show that, for every list (a, b, c, d) of distinct elements of $\mathbb{A}$ there exists $I \in \mathbb{D}$ such that $a \succ'_I b \succ'_I c \succ'_I d$. By A4 there exists J such that $d \succ_J c \succ_J b \succ_J a$. Using A4 again, this time for J, we conclude that there exists $K \in \mathbb{D}$, $K \geqslant J$ such that $a \succ_K b \succ_K c \succ_K d$. Since $K \geqslant J$, we can define $I = K - J \in \mathbb{D}$. Observe that $a \succ'_I b \succ'_I c \succ'_I d$. □

We therefore conclude that $\{\succsim'_I\}_I$ satisfy axioms A1*–A4*. By Theorem 2, there exists a diversified matrix $v : \mathbb{A} \times \mathbb{X} \to \mathbb{R}$ such that:

$$(**) \quad \begin{cases} \text{for every } I \in \mathbb{D} \text{ and every } a, b \in \mathbb{A}, \\ a \succsim'_I b \quad \text{iff} \quad \sum_{x \in \mathbb{X}} I(x) v(a, x) \geqslant \sum_{x \in \mathbb{X}} I(x) v(b, x). \end{cases}$$

Furthermore, the matrix v is unique in the following sense: v and u both satisfy $(*)$ iff there are a scalar $\lambda > 0$, a matrix $\beta : \mathbb{A} \times \mathbb{X} \to \mathbb{R}$ with identical rows (i.e., with constant columns) such that $u = \lambda v + \beta$. We fix a particular matrix v for the rest of the existence proof.

A.2.2. Step 2: Representation for pairs of theories

In order to uniquely identify the constants α^{ab} such that, for every I,

$$a \succsim_I b \quad \text{iff} \quad \alpha^{ab} + \sum_{x \in \mathbb{X}} I(x) v(a, x) \geqslant \sum_{x \in \mathbb{X}} I(x) v(b, x), \tag{2}$$

and to further find a vector w such that $\alpha^{ab} = w(a) - w(b)$, we need to use A5. (See the following subsection for examples illustrating the difficulties one encounters in the absence of A5.)

Fix $a, b \in \mathbb{A}$. Given matrix v, define

$$v_{ab}(I) = \sum_{x \in \mathbb{X}} I(x) v(a, x) - \sum_{x \in \mathbb{X}} I(x) v(b, x) \in \mathbb{R}. \tag{3}$$

Evidently, $v_{ab}(I) = -v_{ba}(I)$. Observe that, by $(**)$, $v_{ab}(I) \geqslant (>) 0$ if and only if $a \succsim'_I (\succ'_I) b$. Using this notation, the representation we seek is

$$a \succsim_I b \quad \text{iff} \quad \alpha^{ab} + v_{ab}(I) \geqslant 0. \tag{4}$$

Choose $I \in \mathbb{D}$ with $a \sim_I b$. Define

$$\alpha^{ab} = -v_{ab}(I).$$

Define also $\alpha^{ab} = -\alpha^{ba}$. We wish to show that this α^{ab} satisfies (4).

Lemma 11. *For every* $J \in \mathbb{D}$,

(i) $v_{ab}(J) + \alpha^{ab} > 0$ *implies that* $a \succ_J b$,
(ii) $v_{ab}(J) + \alpha^{ab} = 0$ *implies that* $a \sim_J b$,
(iii) $v_{ab}(J) + \alpha^{ab} < 0$ *implies that* $b \succ_J a$.

Proof. Let there be given $J \in \mathbb{D}$. Consider $K = I + J$. By A5, there exists $L \in \mathbb{D}_{\geqslant K}$ such that $a \sim_L b$. Since $K \geqslant I, J$ and $L \geqslant K$, $I' \equiv L - I$, $J' = L - J \in \mathbb{D}$.

Since $a \sim_I b$ and $a \sim_L b$, Lemma 6(v) implies that $a \sim'_{I'} b$. Hence $v_{ab}(I') = 0$. Also, $v_{ab}(L) = v_{ab}(I) + v_{ab}(I') = v_{ab}(I)$. We now separate the three cases.

(i) The assumption on J is that $v_{ab}(J) > v_{ab}(I)$. Since $v_{ab}(I) = v_{ab}(L) = v_{ab}(J) + v_{ab}(J')$, we obtain $v_{ab}(J') < 0$, that is, $b \succ'_{J'} a$. If $b \succsim_J a$, Lemma 6(ii) would imply $a \succ_L b$, a contradiction. Hence $a \succ_J b$ is established.

(ii) In this case, $v_{ab}(J) = v_{ab}(I)$ and it follows that $v_{ab}(J') = 0$ and $b \sim'_{J'} a$. If $a \succ_J b$ ($b \succ_J a$), $a \succ_L b$ ($b \succ_L a$) would follow by Lemma 6(i). Hence $a \sim_J b$.

(iii) If $v_{ab}(J) < -\alpha^{ab} = v_{ab}(I)$, $v_{ab}(J') > 0$ and $a \succ'_{J'} b$. If $a \succsim_J b$, Lemma 6(ii) would imply $a \succ_L b$, hence $b \succ_J a$. □

Observe that we also have $b \succsim_I a$ iff $\alpha^{ba} + v_{ba}(I) \geqslant 0$.

Finally, we note that, given the matrix v, α^{ab} and α^{ba} are unique. Moreover, if $u = \lambda v + \beta$ also satisfies $(**)$, the constants α_u^{ab} corresponding to u is $\alpha_u^{ab} = \lambda\alpha^{ab}$.

A.2.3. Step 3: Representation for all theories

Given v satisfying $(**)$, $(\alpha^{ab})_{a,b\in\mathbb{A}}$ are defined as above. Consider a triple $a, b, c \in \mathbb{A}$. Let I satisfy $a \sim_I b \sim_I c$. Then, by Lemma 11,

$$\alpha^{ab} + v_{ab}(I) = 0,$$
$$\alpha^{bc} + v_{bc}(I) = 0,$$
$$\alpha^{ca} + v_{ca}(I) = 0.$$

Summing up, and noticing that, for every a, b, c and every I,

$$v_{ab}(I) + v_{bc}(I) + v_{ca}(I) = 0$$

we obtain that

$$\alpha^{ab} + \alpha^{bc} + \alpha^{ca} = 0.$$

Fix $a \in \mathbb{A}$ and set $w(a) = 0$. For $b \neq a$ define $w(b) = w(a) - \alpha^{ab}$. Thus,

$$\alpha^{ab} = w(a) - w(b).$$

For $b, c \neq a$, observe that

$$\begin{aligned}\alpha^{bc} &= -\alpha^{ab} - \alpha^{ca}\\ &= \big(w(b) - w(a)\big) + \big(w(a) - w(c)\big)\\ &= w(b) - w(c).\end{aligned}$$

Hence, for all $a, b \in \mathbb{A}$,

$$a \succsim_I b \quad \text{iff} \quad w(a) + \sum_{x\in\mathbb{X}} I(x)v(a,x) \geqslant w(b) + \sum_{x\in\mathbb{X}} I(x)v(b,x).$$

Clearly, the vector w is unique up to a shift by an additive constant, leaving the differences $w(a) - w(b) = \alpha^{ab}$ unchanged. This completes the proof of the theorem.

A.3. Necessity and counter-examples

The theorem does not provide an exact characterization of the collections of relations $\{\succsim_I\}_{I\in\mathbb{D}}$ that satisfy A1–A5. While axioms A1–A3 are clearly necessary for the representation (∗), A4 and A5 are not.

As shown in Theorem 2, A4 holds only if the matrix v is diversified. Correspondingly, if $\{\succsim_I\}_{I\in\mathbb{D}}$ satisfy A1–A5, the resulting matrix v will also be diversified.

However, not every diversified v will guarantee that the relations $\{\succsim_I\}_{I\in\mathbb{D}}$ defined by v and a vector w via (∗) will also satisfy A5. In fact, the matrix-vector pairs (v, w) that guarantee A5 as well are precisely those that satisfy the following condition:

(v, w)-solvability. For every $a, b, c \in \mathbb{A}$ there exists $I \in \mathbb{D}$ such that

$$w(a) + \sum_{x\in\mathbb{X}} I(x)v(a,x) = w(b) + \sum_{x\in\mathbb{X}} I(x)v(b,x) = w(c) + \sum_{x\in\mathbb{X}} I(x)v(c,x).$$

Adding diversity of v and (v, w)-solvability, one may obtain a version of Theorem 1 which is a precise characterization. Since the main point of the theorem from a conceptual viewpoint is the sufficiency result, and since it is also the less trivial direction, we chose to omit this condition from the statement of the theorem, leaving it with only one implication.

To see that (v, w)-solvability is not too restrictive, consider the following condition: for every $a_1, a_2, a_3 \in \mathbb{A}$ there are $x_1, x_2, x_3 \in \mathbb{X}$ such that all the numbers $\{w(a_i), v(a_i, x_j)\}_{i,j\leqslant 3}$ are rational (or, to be precise, generate only rational ratios).

However, dropping A5, our result may not hold. In the following, we retain the following notation from the proof: given $\{\succsim_I\}_I$, the relations $\{\succsim'_I\}_I$ derived from them as above. For given $\mathbb{A}$ and $\mathbb{X}$, v denotes a real-valued matrix, $v : \mathbb{A}\times\mathbb{X}\to\mathbb{R}$. In the following examples, v will represent the relations $\{\succsim'_I\}_I$ by (∗∗). We also retain the notation

$$v_{ab}(I) = \sum_{x\in\mathbb{X}} I(x)v(a,x) - \sum_{x\in\mathbb{X}} I(x)v(b,x) \in \mathbb{R}$$

for $I \in \mathbb{D}$, $a, b \in \mathbb{A}$.

We first show that in the absence of A5 uniqueness may fail.

Example 1. Let $\mathbb{A} = \{a, b\}$, $\mathbb{X} = \{x, y\}$ and

$$v = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

For every I, define $a \succ_I b$ if $v_{ab}(I) \geqslant 0$ (i.e., $I(x) \geqslant I(y)$) and $b \succ_I a$ otherwise. In this case, $\{\succsim_I\}_{I\in\mathbb{D}}$ can be represented by (v, w) via (∗) for v above and for every w with

$$w(a) - w(b) \in (0, 1).$$

That is, the representation is not unique. Using the representation, we know that $\{\succsim_I\}_{I\in\mathbb{D}}$ satisfy A1–A3, and A4 can readily be verified. Clearly, A5 is violated in this example.

Second, the following example shows that without A5 representation as in (1) may not be possible:

Example 2. Let $\mathbb{A} = \{a, b\}$, $\mathbb{X} = \{x, y\}$ and

$$v = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{2} \end{pmatrix}.$$

For every I, define $a \succ_I b$ if $v_{ab}(I) \geqslant 0$ and $b \succ_I a$ otherwise.

Observe that, for all $I \neq 0$, $v_{ab}(I) \neq 0$. Hence one may use the matrix v and the constants $w(a) = w(b) = 0$ to represent $\{\succsim_I\}_{I \neq 0}$ via $(*)$. However, (v, w) cannot represent all of $\{\succsim_I\}_{I \in \mathbb{D}}$ because $v_{ab}(0) = 0$, hence $a \succ_0 b$, but $v_{ab}(0) = w(b) - w(a)$.

We claim that no other pair, (v', w'), may represent $\{\succsim_I\}_{I \in \mathbb{D}}$ via $(*)$. To see this, assume that such a pair (v', w') is given. Normalize v' such that the minimal value in each column is 0 and the maximal value in column x is 1. Hence, $v' = v$. Observe that

$$\begin{aligned} \mathit{range}(v_{ab}) &= \left\{ v_{ab}(I) \;\middle|\; I \in \mathbb{D} \right\} \\ &= \{k - l\sqrt{2} \mid k, l \in \mathbb{Z}_+\} \end{aligned}$$

is dense in $\mathbb{R}$. If $w'(b) - w'(a) > 0$, there exists $I \neq 0$ such that $v_{ab}(I) \in (0, w'(b) - w'(a))$ and then (v, w') cannot represent $\succsim_I$ (because (v, w) does). Similarly, $w'(b) - w'(a) < 0$ implies the existence of $I \neq 0$ with $v_{ab}(I) \in (w'(b) - w'(a), 0)$ and the same conclusion follows.

To conclude the proof we need to verify that $\{\succsim_I\}_{I \in \mathbb{D}}$ satisfy A1–A4. In the presence of only two alternatives, A1 only means completeness, which is directly verified from the definition. To see that A2 holds, assume that I, J, K, L are given, with $I + J = K + L$. Assume further that $a \succsim_I b$ and $a \succsim_J b$, but $b \succsim_K a$ and $b \succ_L a$. Observe that $a \succsim_I b$, which is only possible if $a \succ_I b$, implies that $v_{ab}(I) \geqslant 0$, with a strict equality unless $I = 0$. Hence $a \succsim_I b$ and $a \succsim_J b$ imply $v_{ab}(I), v_{ab}(J) \geqslant 0$, and $b \succsim_K a$, $b \succ_L a$ imply $v_{ab}(K), v_{ab}(L) \leqslant 0$. Since $v_{ab}(I) + v_{ab}(J) = v_{ab}(K) + v_{ab}(L)$, this is possible only if $v_{ab}(I) = v_{ab}(J) = v_{ab}(K) = v_{ab}(L) = 0$, and therefore $I = J = K = L = 0$. But then $b \succsim_K a$ and $b \succ_L a$ can't hold. To see that A3 holds, assume that $I, J \in \mathbb{D}$ satisfy $b \succ_J a$ and $a \succsim_{J+I} b$. In this case, $I \neq 0$ and $v_{ab}(I) > 0$ follows. Hence, for every $K \in \mathbb{D}$ there exists $l \in N$ such that $a \succ_{K+lI} b$. Finally, A4 clearly holds because no row in v dominates another.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (6) (1974) 716–723.

[2] A. Billot, I. Gilboa, D. Samet, D. Schmeidler, Probabilities as similarity-weighted frequencies, Econometrica 73 (2005) 1125–1136.

[3] G.J. Chaitin, On the length of programs for computing binary sequences, J. Assoc. Comp. Machines 13 (1966) 547–569.

[4] N. Goodman, Fact, Fiction, and Forecast, Harvard University Press, Cambridge, MA, 1955.

[5] I. Gilboa, O. Lieberman, D. Schmeidler, Empirical similarity, Rev. Econ. Statist. 88 (2006) 433–444.

[6] I. Gilboa, L. Samuelson, Subjectivity in inductive inference, mimeo, 2009.

[7] I. Gilboa, D. Schmeidler, A Theory of Case-Based Decisions, Cambridge University Press, Cambridge, 2001.

[8] I. Gilboa, D. Schmeidler, Inductive inference: an axiomatic approach, Econometrica 71 (2003) 1–26.

[9] A.N. Kolmogorov, On tables of random numbers, Sankhya Ser. A (1963) 369–376.

[10] A.N. Kolmogorov, Three approaches to the quantitative definition of information, Probab. Inf. Transm. 1 (1965) 4–7.

[11] B. Russell, A History of Western Philosophy, Allen & Unwin, Great Britain, 1946.

[12] L.J. Savage, The Foundations of Statistics, John Wiley and Sons, New York, 1954; second edition, Dover, 1972.

[13] E.H. Simpson, The interpretation of interaction in contingency tables, J. Roy. Statist. Soc. Ser. B 13 (1951) 238–241.

[14] E. Sober, Simplicity, Clarendon Press, Oxford, 1975.
[15] R. Solomonoff, A formal theory of inductive inference I, II, Inf. Control 7 (1964) 1–22, 224–254.
[16] C.S. Wallace, Statistical and Inductive Inference by Minimum Message Length, Inf. Sci. Stat., Springer, 2005.
[17] C.S. Wallace, D.M. Boulton, An information measure for classification, Comput. J. 13 (1968) 185–194.
[18] C.S. Wallace, D.L. Dowe, Minimum message length and Kolmogorov complexity, Comput. J. 42 (1999) 270–283.