From ordinal ordering to a prior probability[*]

Shiri Alon[†] and Gabi Gayer[‡]

– Preliminary draft –

November 8, 2020

Abstract

Billot, Gilboa, Samet and Schmeidler [1] (BGSS) offer a model for the formation of a prior probability over states of nature based on data. According to their model, the decision maker possesses a similarity function over observations, and, given a database, adopts the prior that is the similarity-weighted frequency of outcomes within that database. BGSS thus simplify the task of forming a prior probability over states, reducing it to the question of forming a similarity function over observations. Still, the task of forming a similarity function remains, and may not always be straightforward. We characterize two relatively simple procedures for the formation of a similarity function, one which requires placing observations on an integer scale, and the other which is essentially ordinal.

Keywords:

JEL classification: D71, D81

[*] The authors thank Itzhak Gilboa and David Schmeidler for their helpful comments, and the Israeli Science Foundation, grant number 1188/14, for financial support. Part of this research was done while the first author was visiting the Department of Economics at the University of Washington. The author is grateful to the department for its hospitality.

[†] Bar-Ilan University, Ramat-Gan 5290002, Israel. e-mail: Shiri.Alon-Eron@biu.ac.il

[‡] Bar-Ilan University, Ramat-Gan 5290002, Israel. e-mail: Gabriella.Gayer@biu.ac.il

1 Introduction

In classic models of decision making under uncertainty a decision maker faces alternatives which may yield different outcomes depending on a state of nature that will be realized. The Bayesian approach models such a decision maker as evaluating alternatives using a prior probability over the states of nature. Facing a decision problem the decision maker therefore needs to form such a prior probability.

Billot, Gilboa, Samet and Schmeidler [1] (BGSS) examine a decision maker facing a current problem in which various outcomes (states of nature) can be realized. The decision maker has access to a database of past cases that may be relevant to the current problem considered, and forms a prior probability over the outcomes that may be realized based on these data. BGSS describe a decision maker whose prior is formed on the basis of a similarity function, a function returning, for each observation, the degree to which this observation is relevant to the current case under consideration. According to the BGSS model, given a database, the prior formed by the evaluator will be the similarity-weighted frequency of occurrences of outcomes over that database. The problem of forming a data-based prior is therefore reduced to the problem of ascribing similarity values to observations.

Assigning values to express the degree of relevance of observations to a current case may sometimes be easy, for example if all available observations are deemed equally relevant to the problem at hand or if only observations whose attributes are identical to those of the current problem are considered relevant. Alternatively, the similarity function can be estimated empirically. Such an empirical approach is taken in Gilboa, Lieberman, and Schmeidler [2], where a statistical theory for estimating the similarity function is developed, and further implemented in Gayer, Gilboa and Lieberman [3] in the context of real estate pricing. However, the empirical method requires the specification of a functional form of the similarity function, which is not always straightforward. Therefore in some situations the question of forming a similarity function remains difficult.

We characterize axiomatically two simple procedures for assigning similarity values to attribute vectors. The first procedure consists in placing similarity values on an integer scale, namely assigning a number $k \in \{0, 1, 2, \ldots\}$ to each observation, yielding a similarity function that is measured in finite resolution. The second is ordinal, requiring the decision maker to order observations according to their degree of relevance to the current case under consideration.

2 Setup

We first present the model of BGSS that is adjusted to our framework. In BGSS an individual is interested in predicting the outcome (or state of nature) $\omega \in \Omega$ that will obtain in a current problem, by forming a probabilistic belief over $(\Omega, 2^{\Omega})$. It is assumed, in BGSS [1] as well as in this paper, that Ω is finite and $|\Omega| \geq 3$.

When forming a belief, the individual draws upon experience in similar past cases, which we refer to as the individual's private database. An observation in this database is characterized by a list of attributes that were observed in a past case, and that are believed to affect the probability of outcomes. An observation is therefore a pair (a, ω), where a is that observation's vector of attributes, and $\omega \in \Omega$ is its realized outcome. It is assumed here that attributes $a \in \mathbb{A}$ may take on only a finite number of values, restricting them to be discrete variables (this may also include continuous variables that are given in a finite resolution). For example, within observations that describe patients, attributes may include age, gender, and blood pressure level - low/normal/high. A database is a finite sequence of observations, $D \in (\mathbb{A} \times \Omega)^r$ for $r \in \mathbb{N}, r \geq 1$, where $\mathbb{D} = \bigcup_{r \geq 1} (\mathbb{A} \times \Omega)^r$ denotes the set of all possible databases.

The problem under consideration is characterized by the same attributes that characterize observations in the database. The individual forms a probability over possible outcomes $\omega \in \Omega$ given the attributes of the current problem, where this probability is a function of the information available to him or her. That is, if $\Delta(\Omega)$ denotes the set of all possible probabilities over $(\Omega, 2^{\Omega})$, then the individual's priors is a function $p : \mathbb{D} \longrightarrow \Delta(\Omega)$, so that when the database available to the individual is $D \in \mathbb{D}$, she or he forms a probability $p_D \in \Delta(\Omega)$. Note that the attributes characterizing the current problem are a feature of the evaluation problem, and are independent of the database available to the individual.

In the probability assessment problem that was axiomatized in BGSS [1] the individual judges the degree of relevance of each past observation (a, ω) to the problem at hand, quantifying this relevance by a similarity value $s(a) > 0$ assigned to each possible vector a of attributes. Allowing only positive similarity values reflects a presumption that the individual's probability is not based on cases that are expected to yield opposite outcomes to those in the current case, and moreover, every observation has some influence on the resulting probability (values are strictly positive). Similarity values attached to attribute vectors $a \in \mathbb{A}$ will typically depend on the attributes of the current problem, with higher similarity values associated with attributes a that are deemed closer to those of the current problem, thus more relevant for the evaluation question at hand. This is in accordance with the underlying reasoning that from causes which appear similar we expect similar effects (a principal that is attributed to Hume).

The axioms in BGSS characterize a priors function that is a similarity-weighted frequency over available data. That is, under the BGSS model there exists a similarity function $s : \mathbb{A} \times \Omega \longrightarrow \mathbb{R}_{++}$, such that for every database $D \in \mathbb{D}$ and every $\omega' \in \Omega$,

$$p_D(\omega') = \frac{\sum_{(a,\omega)\in D} s(a)\mathbb{1}(\omega = \omega')}{\sum_{(a,\omega)\in D} s(a)} \ ,$$

where $\mathbb{1}(\omega = \omega')$ is an indicator function that assumes the value 1 if ω is equal to ω', and zero otherwise.[1] Note that the similarity is independent of the outcome ω, expressing the fact that the degree to which past cases are relevant for the current problem is the result of the similarity in characteristics, and cannot change if the realized outcome was different. [2]

For our purposes we supplement the above with the possibility that some observations are completely irrelevant to the question at hand. These irrelevant observations will have no effect on the resulting prior probability and will be assigned a similarity of

[1] The basic BGSS prior, characterized by two axioms, is somewhat more general, but in their paper two additional, simple axioms are listed that yield the specific form that we use.

[2] This independence follows from the assumption suggested in BGSS [1] that permuting the states of nature results in a corresponding permutation of the probability vector.

zero, however, they allow for more flexibility in assigning similarity values. In the second type of similarity that we characterize they will serve as a starting point for the ordering of similarity values in increasing relevance order. The next definition identifies irrelevant databases.

Definition 1. *A database E is irrelevant according to the evaluator if for every database $D \in \mathbb{D}$, $p_D = p_{D \circ E}$. Otherwise E is relevant.*

Given a database D, the individual will be able to form a meaningful prior probability so long as D contains information that is relevant to the question at hand. If the individual considers D to be irrelevant, he or she will return a fixed, predetermined prior (for instance, a uniform distribution as a response to the lack of information in D).

With irrelevant databases as a possibility, the Concatenation axiom that appears in BGSS [1] needs to be adjusted:

Concatenation of Relevant Databases (CRD). Let $D, E \in \mathbb{D}$ be two relevant databases, then $D \circ E$ is also relevant, and $p_{D \circ E} = \lambda p_D + (1 - \lambda) p_E$, for $\lambda \in (0, 1)$.

According to the definition of an irrelevant database, if two databases D and E are irrelevant then $D \circ E$ is irrelevant as well. In the other direction, the definition together with the above version of Concatenation implies that a database D is irrelevant if and only if any sub-database of it is irrelevant as well. In particular, D is irrelevant if and only if each of its observations is irrelevant. Observations that are perceived as irrelevant are assigned a zero similarity value within the model. The similarity in our framework thus obtains nonnegative, but not necessarily positive, values.

Leshno (2014) generalizes BGSS to apply to databases that belong to different classes of relevance which are ordered lexicographically. Leshno (2014) and the present paper share the feature that any amount of irrelevant observations is overwhelmed by just a single relevant observation. However, while Leshno (2014) allows the prior that is based on irrelevant information to still vary depending on that data, interpreting irrelevant information as belonging to a different relevance class, we choose to set the prior based on irrelevant information to a fixed one that is interpreted as the evaluator's response to lack of information.

The characterization of probabilistic beliefs in BGSS [1], when supplemented with irrelevant observations and under the adjusted Concatenation axiom, implies an individual whose prior probability given a database is the similarity-weighted frequency over that database, as specified in the assumption below. To keep the discussion interesting, the assumption also states that not all databases are irrelevant.

Assumption 1. *There exists a similarity function* $s : \mathbb{A} \times \Omega \longrightarrow \mathbb{R}_+$*, unique up to multiplication by a positive number, satisfying:*

(i) For every (a, ω)*,* $s(a) = 0$ *if and only if* (a, ω) *is irrelevant*

(ii) There exists $a^* \in \mathbb{A}$ *with* $s(a^*) > 0$

(iii) If $D \in \mathbb{D}$ *is relevant, then,*

$$\forall \omega' \in \Omega \quad p_D(\omega') = \frac{\sum_{(a,\omega)\in D} s(a)\mathbb{1}(\omega = \omega')}{\sum_{(a,\omega)\in D} s(a)} . \tag{1}$$

Otherwise, $p_D = p_0$ *for a fixed* $p_0 \in \Delta(\Omega)$*.*

To simplify notation we denote the sum over the similarity values in a database D by $s(D) = \sum_{(a,\omega)\in D} s(a)$. Under the above assumption, it obtains that D is relevant if and only if $s(D) > 0$.

The representation in Assumption 1 is obtained by first setting aside all irrelevant observations and applying the result in BGSS [1]. Next, these observations are added back by setting their similarities to zero, which corresponds to them not modifying the probability given any database. Finally the original order of observations in any database (mixing relevant and irrelevant observations in any order) can be restored due to the Invariance axiom (see BGSS [1] for that axiom). In accordance with the prior (1), the probability of an event F given a relevant database D is,

$$p_D(F) = \frac{\sum_{(a,\omega)\in D} s(a)\mathbb{1}(\omega \in F)}{\sum_{(a,\omega)\in D} s(a)} .$$

In the simple case where the similarity function is constant, the probability of a state is evaluated by its frequency of occurrence in the entire data. If, on the other hand, a positive weight is assigned only to observations whose attributes are identical to those of the current problem, the probability of a state is evaluated by the frequency of occurrence of that state within identical observations alone. In general, the evaluated probability of a state is a weighted average of the frequency of occurrence in the data, with weights that are determined by the similarity of the observations to the current case.

3 Finite Resolution Similarity

The first procedure that we characterize simplifies the assignment of similarity values by requiring the evaluator to set those values on an integer scale. Namely, for every observation, depending on its attributes, the similarity value ascribed should be chosen out of $\{0, 1, 2, \ldots\}$. If there is no specific functional form that the evaluator adopts for her or his similarity function it is easier to choose a value to express a degree of relevance from among a limited, simple set of values, than to choose when every nonnegative value is possible.

As a similarity function is unique up to multiplication by a positive number, adopting an integer-scaled similarity function is the same as adopting a similarity function that only obtains values of the form $k\delta$, $k \in \mathbb{N}$, for $\delta > 0$. An integer-scaled similarity function is thus in fact a similarity function that is measured in finite resolution. We therefore term this similarity a *finite resolution similarity* .

A finite resolution similarity is characterized by one axiom, stated below (on top of the BGSS axioms). Additional notation is needed for the statement of the axiom: for a database D and a positive integer k, $kD = \underbrace{D \circ \ldots \circ D}_{k \text{ times}}$ is the database composed of k copies of D.

Equatability. For outcomes $\omega, \omega' \in \Omega$ and databases $D, E \in \mathbb{D}$, if $p_D(\omega) > p_D(\omega')$ and $p_E(\omega') > p_E(\omega)$ then there are $m, k \in \mathbb{N}$ such that $p_{(mD)\circ(kE)}(\omega) = p_{(mD)\circ(kE)}(\omega')$.

Proposition 1. *Suppose $p : \mathbb{D} \longrightarrow \Delta(\Omega)$ that satisfies Assumption 1. Then* ***Equatability*** *is satisfied, if and only if, s can be chosen so that $s(a) \in \{0, 1, 2, ...\}$ for every $a \in \mathbb{A}$.*

4 Simple Ordering Similarity

In this section we characterize a similarity which is a special case of the one in the previous section. This similarity is the result of an ordinal procedure, consisting in the individual ordering attribute vectors in an increasing order of degree of relevance. Under two axioms, this ordering of relevance of attribute vectors induces a similarity function which assigns a value of zero to irrelevant attribute vectors, 1 to the attribute vectors that are next in order, 2 to the ones next above, and so on. We use the term *simple ordering similarity* to describe such a similarity function.

The characterization of an individual employing a simple ordering similarity consists of two axioms (in addition to the axioms that characterize a similarity-weighted frequency individual as in Assumption 1). First, we suppose that an irrelevant observation exists. This is employed as a starting point for ordering attribute vectors that are relevant to varying degrees.

Irrelevant Observation (IO). There exists an observation (a_0, ω) such that for every database D, $p_D = p_{D \circ \{(a_0, \omega)\}}$.

Together with the other BGSS axioms, Irrelevant Observation implies that for some vector of attributes a_0, $s(a_0, \omega') = 0$ for every $\omega' \in \Omega$.

Next, we postulate that in specific kinds of ranking reversals there are attribute vectors that can induce equality of probabilities. The reversals under considerations are those in which one outcome is more probable than another given a database, but this ranking is reversed given a database when just one observation with the more probable outcome occurred under different attributes, then there is a third set of attributes with which the two outcomes are equally probable.

Attribute Equatability (AE). For outcomes $\omega, \omega' \in \Omega$, attribute vectors $a, c \in \mathbb{A}$, and a database $D \in \mathbb{D}$, if

$$p_{D\circ(a,\omega)}(\omega) > p_{D\circ(a,\omega)}(\omega') \quad \text{but} \quad p_{D\circ(c,\omega)}(\omega') > p_{D\circ(c,\omega)}(\omega) \ ,$$

then there exists an attributes vector $b \in \mathbb{A}$ such that $p_{D\circ(b,\omega)}(\omega) = p_{D\circ(b,\omega)}(\omega')$.

An individual forming a similarity-weighted frequency prior satisfies the above two axioms, if and only if, this individual's similarity is a simple ordering similarity. This is stated in the following theorem.

Theorem 1. *Suppose $p : \mathbb{D} \longrightarrow \Delta(\Omega)$ that satisfies Assumption 1. Then (**IO**) and (**AS**) are satisfied, if and only if, s can be chosen so that $s(\mathbb{A}) = \{0, 1, ..., k\}$ for some $k \in \mathbb{N}, \ k > 0$. Moreover, this choice is unique up to multiplication by a positive number.*

According to this theorem, an individual who wishes to form a prior over outcomes according to the data-based similarity-weighted frequency rule, and who finds the two axioms above appealing, needs only to pinpoint an irrelevant attribute vector and then continue to order all other attribute vectors in an increasing order of degree of relevance. The similarity that would then generate this individual's data-based priors is the one assigning the value k to the attributes vector that is k-th in order. Under the two axioms above the procedure of forming a data-based prior over outcomes is therefore greatly simplified, as it boils down to an ordinal task, whereby attribute vectors are ordered in an increasing degree of relevance.

5 Proofs

5.1 Proof of Proposition 1

Suppose that $p : \mathbb{D} \longrightarrow \Delta(\Omega)$ satisfies Assumption 1 and **Equatability**. Let $\omega, \omega' \in \Omega$ be two states, and $a \in \mathbb{A}$ a relevant attributes vector. Then $p_D(\omega) > p_D(\omega')$ for $D =$

$\{(a^*, \omega)\}$ and $p_E(\omega') > p_E(\omega)$ for $E = \{(a, \omega')\}$. Therefore, by Equatability, there are $m, k \in \mathbb{N}$ such that $p_{(mD)\circ(kE)}(\omega) = p_{(mD)\circ(kE)}(\omega')$, implying, $ms(a^*) = ks(a)$. That is, for every $a \in \mathbb{A}$, $s(a) = rs(a^*)$ for a nonnegative rational number r. By uniqueness of the similarity function up to multiplication by a positive number, we may multiple s by a large enough positive integer, and divide it by $s(a^*)$, to obtain that the resulting similarity values are all nonnegative integers, as stated in the proposition.

In the other direction, suppose that the similarity s obtains only nonnegative integer values. Suppose that $p_D(\omega) > p_D(\omega')$ and $p_E(\omega') > p_E(\omega)$ for databases D and E and states $\omega, \omega' \in \Omega$. Denote by D_ω the sub-database in D containing only observations in which ω is the state that was realized, and similarly for ω' and for E. Employing the representation in 1, the inequalities over probabilities translate to: $s(D_\omega) > s(D_{\omega'})$ and $s(E_{\omega'}) > s(E_\omega)$, where all these similarity values are integers. Denote $s(D_\omega) - s(D_{\omega'}) = k$ and $s(E_{\omega'}) - s(E_\omega) = m$, then k, m are positive integers, and it holds that $km = m\,(s(D_\omega) - s(D_{\omega'})) = k\,(s(E_{\omega'}) - s(E_\omega))$, hence $ms(D_\omega) + ks(E_\omega) = ms(D_{\omega'}) + ks(E_{\omega'})$, which according to the formula for generating a prior given a database, implies, $p_{(mD)\circ(kE)}(\omega) = p_{(mD)\circ(kE)}(\omega')$, as required.

5.2 Proof of Theorem 1

Suppose that Assumption 1 holds, and denote the corresponding non-negative similarity function by s. Assume that **IO** and **AS** hold. Since there is a finite number of attributes vectors, we may order similarity values in an increasing order. Namely, there are $a_0, a_1, \ldots$ such that $s(a_0) < s(a_1) < \ldots$. If there is more than one attributes vector with the same similarity value, then the corresponding a_k is one of those attribute vectors. According to **IO**, $s(a_0) = 0$, and following the non-degeneracy part (ii) of Assumption 1, there are at least two attribute vectors, a_0 and a_1, such that $s(a_0) < s(a_1)$.

To show that s is a simple ordering similarity, it should be proved that $s(a_k) = ks(a_1)$ for every k. If $s(a_1)$ is maximal then this claim is trivially true, so suppose that $s(a_1)$ is not maximal. The proof continues by induction. Consider the database $D =$

$\{(a_1,\omega),(a_2,\omega')\}$, and suppose by negation that $s(a_2) \neq 2s(a_1)$. If $s(a_2) > 2s(a_1)$, then since $s(a_1) > 0$ and $s(a_2) > 2s(a_1)$, it holds that,

$$\begin{aligned} p_{D\circ(a_2,\omega)}(\omega) &> p_{D\circ(a_2,\omega)}(\omega') \ , \\ p_{D\circ(a_1,\omega)}(\omega') &> p_{D\circ(a_1,\omega)}(\omega) \ . \end{aligned}$$

AS implies that there exists an attributes vector b such that $p_{D\circ(b,\omega)}(\omega') = p_{D\circ(b,\omega)}(\omega)$, that is, such that $s(b) = s(a_2) - s(a_1)$, hence by the negation assumption, $s(a_1) < s(b) < s(a_2)$. Contradiction. If, on the other hand, $s(a_2) < 2s(a_1)$, then,

$$\begin{aligned} p_{D\circ(a_1,\omega)}(\omega) &> p_{D\circ(a_1,\omega)}(\omega') \ , \\ p_{D\circ(a_0,\omega)}(\omega') &> p_{D\circ(a_0,\omega)}(\omega) \ , \end{aligned}$$

which by **AS** implies that there exists an attributes vector b such that $p_{D\circ(b,\omega)}(\omega') = p_{D\circ(b,\omega)}(\omega)$, hence, $s(b) = s(a_2) - s(a_1)$, yielding a contradiction, as $s(a_0) < s(a_2) - s(a_1) < s(a_1)$.

Suppose that for every $j = 1, \ldots, k$, $s(a_k) = ks(a_1)$, and suppose by negation that $s(a_{k+1}) > (k+1)s(a_1)$. Let $D = \{(a_1,\omega),(a_{k+1},\omega')\}$. Following the negation assumption and the fact that $s(a_1) > 0$,

$$\begin{aligned} p_{D\circ(a_{k+1},\omega)}(\omega) &> p_{D\circ(a_{k+1},\omega)}(\omega') \ , \\ p_{D\circ(a_1,\omega)}(\omega') &> p_{D\circ(a_1,\omega)}(\omega) \ . \end{aligned}$$

AS implies that there exists an attributes vector b such that $p_{D\circ(b,\omega)}(\omega') = p_{D\circ(b,\omega)}(\omega)$, hence, $s(b) = s(a_{k+1}) - s(a_1)$. However, since by the induction assumption, and by the negation assumption, $s(a_k) = ks(a_1) < s(a_{k+1}) - s(a_1) < s(a_{k+1})$, a contradiction is inflicted.

To rule out the case $s(a_{k+1}) < (k+1)s(a_1)$, consider $E = \{(a_k,\omega),(a_{k+1},\omega')\}$. The inequalities

$$\begin{aligned} p_{E\circ(a_1,\omega)}(\omega) &> p_{E\circ(a_1,\omega)}(\omega') \ , \\ p_{E\circ(a_0,\omega)}(\omega') &> p_{E\circ(a_0,\omega)}(\omega) \ , \end{aligned}$$

together with **AS**, deliver that there exists an attributes vector b such that $s(b) = s(a_{k+1}) - s(a_k)$, which by the preceding assumptions should satisfy, $s(a_0) < s(b) <$

$s(a_1)$. Contradiction. It is thus concluded that $s(a_{k+1}) = (k+1)s(a_1)$, and the similarity values of s are equally spaced, starting from $s(a_0) = 0$.

In the other direction, **IO** is immediate. For **AS**, suppose that for a database D,

$$\begin{aligned} p_{D\circ(a,\omega)}(\omega) &> p_{D\circ(a,\omega)}(\omega')\ , \\ p_{D\circ(c,\omega)}(\omega') &> p_{D\circ(c,\omega)}(\omega)\ . \end{aligned}$$

Then it should hold that $s(a) > s(D_{\{\omega'\}}) - s(D_{\{\omega\}}) > s(c)$. Since all similarities are integer-valued, the difference $s(D_{\{\omega'\}}) - s(D_{\{\omega\}})$ must be an integer. Being between $s(c)$ and $s(a)$, and since the range of s is an interval within the integers, it must equal the similarity of some attributes vector.

References

[1] Billot, A., I. Gilboa, D. Samet, and D. Schmeidler, (2005) “Probabilities as Similarity-Weighted Frequencies,” *Econometrica*, 73 , 1125-1136

[2] Gilboa, I., O. Lieberman, and D. Schmeidler, (2006),“Empirical Similarity,” *Review of. Economics and. Statististics*, 88, 433–444.

[3] Gayer, G., I. Gilboa, and O. Lieberman (2007), “Rule-based and case-based reasoning in housing prices” *The BE Journal of Theoretical Economics* 7(1).