Constraints as Data: A New Perspective on Inferring Probabilities

Manfred Jaeger
MPI Informatik
Stuhlsatzenhausweg 85, 66121 Saarbrücken, Germany
jaeger@mpi-sb.mpg.de

Abstract

We present a new approach to inferring a probability distribution which is incompletely specified by a number of linear constraints. We argue that the currently most popular approach of entropy maximization depends on a "constraints as knowledge" interpretation of the constraints, and that a different "constraints as data" perspective leads to a completely different type of inference procedures by statistical methods. With statistical methods some of the counterintuitive results of entropy maximization can be avoided, and inconsistent sets of constraints can be handled just like consistent ones. A particular statistical inference method is developed and shown to have a nice robustness property.

1 Introduction

Probabilistic representations of uncertainty usually consist of a single probability distribution over a large (but finite) domain of possible states $D = \{d_1, \ldots, d_n\}$. It is thus required to assign a probability value p_i to each state d_i. Usually, a direct, full assessment of all these values is very difficult or impossible. All one usually is able to obtain are partial descriptions of $\boldsymbol{p} = (p_1, \ldots, p_n)$ by constraints of e.g. the form $\boldsymbol{p}(A \mid B) \leq z$, $\boldsymbol{p}(A) + \boldsymbol{p}(B) \leq \boldsymbol{p}(C)$, or "$A$ and B are independent", where A, B, C are subsets of D. Such constraints can be derived by knowledge elicitation from an expert, by direct observations of the domain, or by any other information gathering process.

A set $c_1, \ldots, c_N$ of constraints defines the set $\Delta(c_1, \ldots, c_N)$ of probability measures on D that are consistent with the constraints. Very rarely will $\Delta(c_1, \ldots, c_N)$ consist of a single probability distribution. Instead, it will either contain more than one element, or be empty (i.e., the constraints are inconsistent). A fundamental problem in probabilistic reasoning then is to select from the admissible set $\Delta(c_1, \ldots, c_N)$ a distribution $\boldsymbol{p} =: \mathit{sel}(c_1, \ldots, c_N)$ as the best guess for the true distribution the constraints describe.

This *measure selection problem* is well studied in the literature, particularly for the case where the constraints are linear and consistent. It is almost unanimously suggested that in this case one should select the distribution with maximal entropy from $\Delta(c_1, \ldots, c_N)$ [Shore and Johnson, 1980; Lemmer and Barth, 1982; Jaynes, 1982; Cheeseman, 1983; Paris and Vencovská, 1990; Rödder and Meyer, 1996]. A more general class of constraints is considered by Drudzel and van der Gaag [1995] who then employ the center of mass selection rule (according to this rule one selects the center of mass of the admissible region).

In this paper we propose a new selection rule which is radically different from either maximum entropy or center of mass. It is motivated by the observation that in spite of the very compelling justifications it has been given [Shore and Johnson, 1980; Jaynes, 1982; Paris and Vencovská, 1990], maximum entropy selection has some rather counterintuitive properties. These are illustrated by the following examples.

Example 1.1 Overhearing two strangers talking at an airport, we hear the first one saying " ... Jones got at least 45% of the votes ... ", and the second replying " ... Smith didn't get any less than 5% either ... ". Before the two disappear in the crowd, we also hear them both agreeing on the fact that if anyone else had bothered to run for mayor, then neither Smith nor Jones would have had a chance of winning the election. Suppose, now, that we need to assess the probability $P(\mathit{Smith})$ of an arbitrary voter in the unnamed home town of the two strangers having voted for Smith. The information we have establishes a lower bound of 0.05 and an upper bound of 0.55 on $P(\mathit{Smith})$. Moreover, we have learned that the relevant underlying state space only consists of *Smith* and *Jones*. If we base our probability assessment on entropy maximization, then we will obtain $P(\mathit{Smith}) = 0.5$. Intuitively, this assessment appears to be overly optimistic from Smith's point of view.

Example 1.2 For the construction of a medical diagnosis system ten different experts are asked for bounds on the two crucial conditional probabilities $P_1 = P(\mathit{stylosis} \mid \mathit{polycarpia})$, and $P_2 = P(\mathit{xylopserosis} \mid \mathit{anameae})$. Assume that 0.41 and 0.51 are the greatest lower bound and smallest upper bound, respectively, mentioned by any expert for P_1. Having complete confidence in the experts, we will then take it as given that the true value for P_1 lies in the interval [0.41,0.51]. Let [0.49,0.61] be the correspondingly defined interval for P_2. Applying maximum entropy to find the best values for P_1 and P_2 for our expert system, we will determine $P_1 = P_2 = 0.5$. This appears somewhat counterintuitive because we have chosen the same value for both probabilities,

even though the information provided would seem to indicate a smaller value for P_1 than for P_2.

The reasons why the maximum entropy solution appears counterintuitive in the two examples are very similar. In the first example, an equal percentage of 50% of votes for both Smith and Jones seems implausible, because the constraints are highly asymmetric. Experience tells us that the disparity of the given lower bounds probably reflects a similar disparity of the actual values, which will rather be assumed to be approximately 90% for Jones and 10% for Smith. Such an assessment could be based on a natural explanation for how the constraints were generated in the first place: one might suspect, for instance, that the constraints report the partial count of 50% of the votes, among which 45% were found to be for Jones, and 5% for Smith. In the second example it appears unlikely that the experts would systematically state larger upper and lower bounds for P_2 than for P_1 if these two probabilities were really the same.

In both examples we have thus argued that the maximum entropy distribution is a counterintuitive solution of the selection problem, because the given constraints are unlikely to be observed when this is the true distribution. Underlying this argument is a view of constraints that is fundamentally different from the view which (implicitly) underlies the use of the maximum entropy principle: entropy maximization is predicated on the view that the given constraints are just a description of a state of knowledge: the knowledge that the true distribution is a member of the admissible region defined by the constraints. We call this the *constraints as knowledge* perspective. In our examples – and, we would claim, in most cases where we encounter the measure selection problem – the given constraints are not only a description of our knowledge, they also are the source of our knowledge. They thereby carry not only the face-value information consisting of a restriction of the admissible region; they also carry the meta-information consisting of the fact that we observed exactly these constraints. This meta-information is relevant for the solution of the measure selection problem as it allows us to reason about the likelihood of observing the given constraints for different true distributions. We call the view of constraints that tries to take into account this meta-information the *constraints as data* perspective: constraints are seen as randomly sampled pieces of information. The distribution of this constraint data (the *constraint distribution*) is in part determined by the true distribution on the domain (the *domain distribution*), which we want to determine. In other words, the domain distribution is a parameter of the constraint distribution. Our problem thus becomes a statistical one: to infer a parameter of a distribution from random samples drawn from that distribution.

All statistical methods rely in part on considerations of likelihood. The most direct way to use likelihood is by maximum likelihood inference: select the parameter that gives highest probability to the observed sample. The measure selection rule we develop in this paper is likelihood maximization for the observed constraints. The main problem we face in a formal development of this intuitive principle is that statistical methods usually require a specific model on how the distribution of observed data depends on the parameter of interest, i.e. the stipulation of some underlying parametric family. Our goal, however, is to define a general rule for measure selection that does not require any knowledge about the random mechanism that produces the constraints. Our approach towards solving this dilemma is that of robust statistics: we do propose a specific model for the random generation of constraints, but this model is chosen such that in the long run it will lead to correct inferences for a fairly wide class of constraint distributions.

The constraints as data perspective coupled with statistical approaches to measure selection permits us to handle inconsistent sets of constraints just like consistent ones. Our statistical model for the constraint observation only must allow for the observation of wrong constraints, i.e. constraints not satisfied by the true distribution (as an erroneous assessment given by an expert, the premature and incorrect report of an election result, etc.). Such a model then assigns nonzero likelihoods to inconsistent sets of constraints, and a maximum likelihood solution can be found just as for consistent constraint sets.

The idea of measure selection by likelihood maximization for the observed constraints was already expressed in [Jaeger, 1998], but no concrete formalization of the idea was developed. The view of constraints as data has also been taken in somewhat different form by Dickey [1980], who proposed a model in which partial specifications of a probability distribution P were treated as random variables with a distribution depending on P. A major difference between Dickey's and our work is that Dickey does not consider partial specifications by arbitrary linear constraints, but only by values for a fixed set of "aspects" of P. It is interesting to note that Dickey takes it for granted that in most cases the specified aspects will overdetermine the model, i.e. be inconsistent, whereas authors in artificial intelligence assume underdetermined models.

In this paper we can only give an overview of our maximum likelihood approach to measure selection. Goal of this paper is to convey the main ideas, and to provide some insight into the feasibility of their mathematical development. More technical details, including proofs of the theorems here stated, will be given in a full technical paper.

2 The Constraint Sample Space

To treat constraints as random samples we have to view them as elements of some sample space on which probability distributions can be defined. Throughout we assume that the constraints refer to a distribution on a domain of n elements. The set of all these distributions can be identified with

$$\Delta^n := \{(p_1, \ldots, p_n) \in \mathbb{R}^n \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1\}.$$

A linear constraint then has the general form

$$c: \quad x_1 p_1 + \ldots + x_n p_n \leq z \qquad (x_1, \ldots, x_n, z \in \mathbb{R}). \quad (1)$$

We could identify this constraint with its parameters $x_1, \ldots, x_n, z$, and thus take $\mathbb{R}^{n+1}$ as our sample space. However, this would mean to view two equivalent constraints like

$p_1 - 2p_2 \leq 0.2$ and $2p_1 - 4p_2 \leq 0.4$ as different sample points. As it does not seem sensible that our method should depend on such representational variants of constraints, we prefer to distinguish constraints only according to the subsets of distributions they define. This can be done by writing constraints in a normal form

$$s_1 p_1 + \ldots + s_n p_n \leq 0, \tag{2}$$

where $\boldsymbol{s} := (s_1, \ldots, s_n)$ is an element of the $n-1$-dimensional unit sphere

$$S^{n-1} = \{(s_1, \ldots, s_n) \mid \sum_i s_i^2 = 1\}.$$

As every linear constraint (1) can be transformed into a unique normal form (2), we henceforward identify linear constraints c with points $\boldsymbol{s} \in S^{n-1}$, and model randomly observed constraints by probability distributions on S^{n-1}.

In the binomial case ($n = 2$), a constraint (2) is a (nontrivial) lower bound on p_1 iff $s_1 < 0$ and $s_2 > 0$; it is a (nontrivial) upper bound iff $s_1 > 0$ and $s_2 < 0$. The following definition generalizes this classification of constraints.

Definition 2.1 A *sign vector* is any vector with components in $\{-1, 0, 1\}$. For $r \in \mathbb{R}$ we define $\mathit{sign}(r)$ as $-1, 0$ or 1, depending on whether $r < 0$, $r = 0$, or $r > 0$. The sign vector $\mathit{sign}(\boldsymbol{s})$ for $\boldsymbol{s} \in S^{n-1}$ is the vector $(\mathit{sign}(s_i))_{i=1,\ldots,n}$. Each sign-vector ζ of length n defines a *sector* S^ζ in S^{n-1}:

$$S^\zeta := \{\boldsymbol{s} \in S^{n-1} \mid \mathit{sign}(\boldsymbol{s}) = \zeta\}. \tag{3}$$

The intuition behind this definition is that sectors contain constraints of the same qualitative type. The classification of constraints according to sectors gives rise to the following coarser, three-way distinction: a constraint $\boldsymbol{s}$ is *vacuous* iff $\mathit{sign}(s_i) \neq\in \{-1, 0\}$ for all i (a vacuous constraint is satisfied by all $\boldsymbol{p} \in \Delta^n$); $\boldsymbol{s}$ is a *support constraint* iff $\mathit{sign}(s_i) \in \{0, 1\}$ for all i (a support constraint is satisfied by all $\boldsymbol{p} \in \Delta^n$ whose set of support is a subset of $\{i \mid \mathit{sign}(s_i) = 0\}$); $\boldsymbol{s}$ is *proper* iff $\mathit{sign}(s_i) = 1$ and $\mathit{sign}(s_j) = -1$ for some i, j (a proper constraint $\boldsymbol{s}$ divides the interior of Δ^n, i.e. there exist $\boldsymbol{p} \in \mathit{int}\,\Delta^n$ that satisfy $\boldsymbol{s}$, and $\boldsymbol{p}' \in \mathit{int}\,\Delta^n$ that do not satisfy $\boldsymbol{s}$).

Figure 1 illustrates constraints from different sectors. Shown in the figure is the polytope Δ^3 with its 3 vertices corresponding to domain distributions that assign unit mass to one of the states in D. Six different constraints from three different sectors are represented by the halfplanes of points satisfying the constraint. Halfplanes are shown by their boundary line, and a shading that indicates to which side of the boundary the halfplane extends.

For the rest of the paper we make two simplifying assumptions. *Assumption 1:* All constraints in the observed sample are proper. *Assumption 2:* The model $\boldsymbol{p} \in \Delta^n$ we want to determine lies in the interior $\mathit{int}\,\Delta^n$ of Δ^n. The two assumptions are somewhat connected. Non-proper constraints are essentially constraints on the set of support of $\boldsymbol{p}$. Thus, both assumptions will be satisfied if in an initial inference step we use all observed non-proper constraints to determine a set of support for our model, and then use the method we shall develop on the remaining proper constraints to determine $\boldsymbol{p}$ with

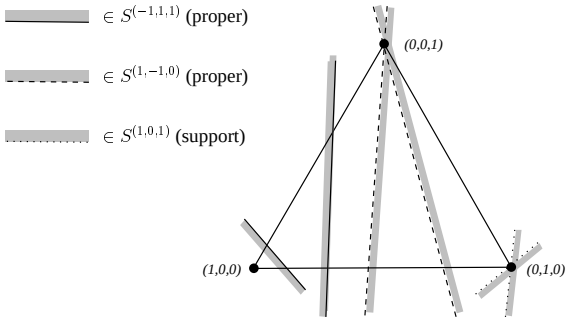

Figure 1: Constraints from different sectors

that set of support. With these assumptions, the measure selection problem consists of finding selection rules in the following sense.

Definition 2.2 A *selection rule* is any function that for every $n \in \mathbb{N}$ maps any tuple $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N \in (S^{n-1})^N$ ($N \in \mathbb{N}$) of proper constraints to a set $\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) \subseteq \mathit{int}\,\Delta^n$.

This definition of a selection rule is not directly tied to constraints as data, and is very similar to Paris and Vencovská's [1990] notion of an *inference process*. It is very general in two respects: first, it is not required that $\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ consists of a unique point. While it is obviously desirable that a selection rule yields unique solutions as often as possible, one needs to take the possibility into account that no principled statistical method can guarantee unique solutions in all cases. Second, it is not demanded that $\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ be a subset of $\Delta(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ (the distributions on D that actually satisfy the constraints). Such a demand, which is natural from the constraints as knowledge perspective, is not required from the constraints as data perspective. To see why, recall that in order to deal with inconsistent constraint sets (and also for greater realism) we should work with probabilistic models according to which it is possible to observe false constraints. This means that even for consistent constraint sets we must take the possibility into account that it contains false constraints, and that therefore the true domain distribution does not belong to $\Delta(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$.

3 Invariance and Equivariance

A selection rule in the sense of Definition 2.2 is a maximum likelihood selection rule, if it takes the following form: for every $N \in \mathbb{N}$, and for every $\boldsymbol{p} \in \mathit{int}\,\Delta^n$ a probability distribution $F^N_{\boldsymbol{p}}$ on $(S^{n-1})^N$ is defined, and for a sample $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ of N constraints we select the distributions in $\mathit{int}\,\Delta^n$ that maximize the likelihood of the sample:

$$\mathit{sel}_F(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) := \arg\max_{\boldsymbol{p}} f^N_{\boldsymbol{p}}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N), \tag{4}$$

where $f^N_{\boldsymbol{p}}$ is the density function of $F^N_{\boldsymbol{p}}$. Usually, constraints will be assumed to be independent, so that with $f_{\boldsymbol{p}} := f^1_{\boldsymbol{p}}$

$$f^N_{\boldsymbol{p}}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) = \prod_{i=1}^N f_{\boldsymbol{p}}(\boldsymbol{s}_i). \tag{5}$$

Whenever some information about the random process that generates the constraints is available, then one will choose in (4) a family $(F_{\boldsymbol{p}}^N)_{\boldsymbol{p}}$ that is a plausible model for this random process. The question we shall be concerned with, however, is what to do when no particular information about the generation of the constraints is available. Thus, we address the same inference problem as addressed by maximum entropy inference: input of the selection problem is a set of constraints, and nothing else.

The justification of the maximum entropy principle, in broad outline, takes the following form: because there is no information except the constraints, one should select the domain distribution $\boldsymbol{p}$ that encodes the least additional information beyond the face-value information provided by the constraints. Minimal information content, in turn, is realized by distributions $\boldsymbol{p}$ that, roughly speaking, maximize independencies and uniformity. Our approach to dealing with the lack of information is somewhat similar, only applied to the constraint distribution: because we have no information on the family $(F_{\boldsymbol{p}}^N)_{\boldsymbol{p}}$, we should assume the least specific structure of this family. In particular, we will assume that the constraints are independent, and that the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ is homogeneous in $\boldsymbol{p}$, in a sense that will be formalized by the notion of *G-invariance*, which is developed in this section.

We derive the concept of G-invariance from the semantic concept of a random constraint generating mechanism that works uniformly for all $\boldsymbol{p}$. Additional support for the G-invariance assumption on $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ will be given by the observation that maximum likelihood selection with respect to G-invariant families is *G-equivariant*, and that this can be understood as the formalization of the intuitive principle that a uniform shift applied to all constraints should induce a similar shift of the selected distributions (cf. Example 1.2).

To motivate the concept of a random constraint generating mechanism that works uniformly for all $\boldsymbol{p}$, reconsider Example 1.1, and the subsequently given explanation of how the constraints might have been generated from a partial count of 50% of the votes. If the true distribution is indeed $\hat{\boldsymbol{p}} = (\hat{p_1}, \hat{p_2}) = (.1, .9)$ $(= (P(\textit{Smith}), P(\textit{Jones})))$, then the observation of the constraints $.05 \leq p_1 \leq .55$ follow as a result of a sequence of chance events: the partial count of 50% of the votes becomes known, this partial count happens to be an accurate projection of the full count, and two strangers happen to mention these partial counts in their conversation. This sequence of chance events does not depend on the distribution $\hat{\boldsymbol{p}}$. If the true distribution was $\boldsymbol{p}^* = (.4, .6)$, then the same sequence of events would occur with the same likelihood, but now generate the constraints $.2 \leq p_1 \leq .7$.

Generalizing from this example, we obtain the (yet informal) notion of a random constraint generating process that works uniformly for all $\boldsymbol{p}$: the constraint generating mechanism will produce a constraint $\hat{\boldsymbol{s}}$ when the true domain distribution is $\hat{\boldsymbol{p}}$ with the same likelihood as it will produce a constraint $\boldsymbol{s}^*$ corresponding to $\hat{\boldsymbol{s}}$ when the true distribution is $\boldsymbol{p}^*$. To make this idea precise, we have to find a transformation g on constraints that maps every $\boldsymbol{s} \in S^{n-1}$ to a corresponding $g(\boldsymbol{s}) \in S^{n-1}$, such that an observation of $\boldsymbol{s}$ under the true distribution $\hat{\boldsymbol{p}}$ corresponds to an observation of $g(\boldsymbol{s})$ under $\boldsymbol{p}^*$. The correspondence expressed by g should preserve elementary qualitative properties of constraints. Two natural preservation conditions are:

Sector preservation: g maps every sector S^ζ bijectively onto itself.

Implication preservation: For all $k \in \mathbb{N}$, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$:

$$\bigcap_{i=1}^{k-1} \Delta(\boldsymbol{s}_i) \subseteq \Delta(\boldsymbol{s}_k) \Leftrightarrow \bigcap_{i=1}^{k-1} \Delta(g(\boldsymbol{s}_i)) \subseteq \Delta(g(\boldsymbol{s}_k)) \quad (6)$$

Sector preservation means that two corresponding constraints should be of the same qualitative type, as expressed by their membership in a sector. Implication preservation says that logical relationships between constraints should be preserved. Implication preservation is equivalent to the conjunction of two simpler conditions: $g(-\boldsymbol{s}) = -g(\boldsymbol{s})$ for all $\boldsymbol{s}$, and $\bigcap_{i=1}^{k-1} \Delta(\boldsymbol{s}_i) = \emptyset \Leftrightarrow \bigcap_{i=1}^{k-1} \Delta(g(\boldsymbol{s}_i)) = \emptyset$ (consistency preservation).

The following definition introduces a class of transformations that satisfy both properties.

Definition 3.1 Let $\boldsymbol{r} = (r_1, \ldots, r_n) \in (\mathbb{R}^+)^n$. The transformation $g_{\boldsymbol{r}} : S^{n-1} \to S^{n-1}$ is defined by

$$g_{\boldsymbol{r}}\ ((s_1, \ldots, s_n)) := \frac{(r_1 s_1, \ldots, r_n s_n)}{\|(r_1 s_1, \ldots, r_n s_n)\|}.$$

We write G_n for the set $\{g_{\boldsymbol{r}} \mid \boldsymbol{r} \in (\mathbb{R}^+)^n\}$.

It is obvious that transformations $g_{\boldsymbol{r}}$ satisfy sector preservation. They also satisfy a slightly strengthened version of implication preservation. For this, denote by $H(\boldsymbol{s}) \subseteq \mathbb{R}^n$ the set of all real solutions of (2), without the restriction to solutions $\boldsymbol{p} \in \Delta^n$ (so that $\Delta(\boldsymbol{s}) = H(\boldsymbol{s}) \cap \Delta^n$). In analogy to (6) we can then define *global implication preservation* of g by the condition

$$\bigcap_{i=1}^{k-1} H(\boldsymbol{s}_i) \subseteq H(\boldsymbol{s}_k) \Leftrightarrow \bigcap_{i=1}^{k-1} H(g(\boldsymbol{s}_i)) \subseteq H(g(\boldsymbol{s}_k)). \quad (7)$$

With condition (7) we look at constraints as defining sets of real numbers, not sets of probability distributions. In our context condition (6) seems to be the more pertinent one. We nevertheless here introduce the global version (7), because with this version we can prove the following representation theorem.

Theorem 3.2 Let $n \geq 3$, $g : S^{n-1} \to S^{n-1}$. g preserves sectors and is globally implication preserving iff $g \in G_n$.

The theorem does not hold for $n = 2$. The proof is by reduction to a classical representation result in projective geometry which characterizes mappings that preserve collinearity. We may conjecture that the theorem also holds when the condition of global implication preservation is replaced by implication preservation in our preferred sense (6). A proof of this modified theorem appears to be considerably harder, however.

In light of Theorem 3.2 we see the transformations $g_{\boldsymbol{r}} \in G_n$ as the adequate realization of the concept of correspondence of constraints. To relate this correspondence to different domain distributions, we define dual transformations on Δ^n.

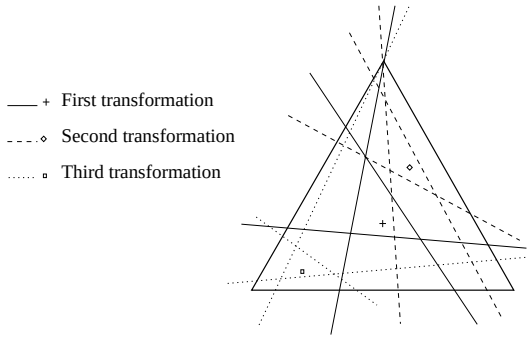

Figure 2: G-transformations and equivariant selection

Definition 3.3 Let $\boldsymbol{r} = (r_1, \ldots, r_n) \in (\mathbb{R}^+)^n$. The transformation $\bar{g}_{\boldsymbol{r}} : \Delta^n \to \Delta^n$ is defined by

$$\bar{g}_{\boldsymbol{r}}\ ((p_1, \ldots, p_n)) := \frac{(p_1/r_1, \ldots, p_n/r_n)}{\sum_{i=1}^n p_i/r_i}.$$

We write $\bar{G}_n$ for the set $\{\bar{g}_{\boldsymbol{r}} \mid \boldsymbol{r} \in (\mathbb{R}^+)^n\}$.

The mapping $\bar{g}_{\boldsymbol{r}}$ is dual to $g_{\boldsymbol{r}}$ in that it is the only transformation of Δ^n such that for all $\boldsymbol{p}$,$\boldsymbol{s}$:

$$\boldsymbol{p} \in \Delta(\boldsymbol{s}) \quad \Leftrightarrow \quad \bar{g}_{\boldsymbol{r}}(\boldsymbol{p}) \in \Delta(g_{\boldsymbol{r}}(\boldsymbol{s})). \tag{8}$$

Our initial intuition of corresponding observations of constraints now can be phrased as follows: the observation of constraint $\boldsymbol{s}$ under the true domain distribution $\boldsymbol{p}$ corresponds to the observation of constraint $g_{\boldsymbol{r}}(\boldsymbol{s})$ under the true domain distribution $\bar{g}_{\boldsymbol{r}}(\boldsymbol{p})$. Figure 2 shows three different transformations of a set of three constraints, and the dual transformations of one probability measure inside the admissible region of the constraints. Each of the three sets of constraints can be transformed into any of the other two sets by unique $g_{\boldsymbol{r}} \in G_n$. The dual transformations $\bar{g}_{\boldsymbol{r}}$ at the same time transform the indicated points in Δ^3 into each other.

With the transformations G_n and $\bar{G}_n$ we can now finally formalize the idea of a constraint generating mechanism that works uniformly for all $\boldsymbol{p}$:

Definition 3.4 Let $(F_{\boldsymbol{p}})_{\boldsymbol{p} \in \mathit{int}\, \Delta^n}$ be a family of distributions on S^{n-1}. The family is called *G-invariant* if for all $g_{\boldsymbol{r}} \in G$ and all $\boldsymbol{p} \in \mathit{int}\, \Delta^n$:

$$g_{\boldsymbol{r}}(F_{\boldsymbol{p}}) = F_{\bar{g}_{\boldsymbol{r}}(\boldsymbol{p})}. \tag{9}$$

When the distributions $F_{\boldsymbol{p}}$ are represented by densities $f_{\boldsymbol{p}}$ relative to a suitably chosen underlying measure on S^{n-1}, then (9) can be expressed by the condition

$$f_{\boldsymbol{p}}(\boldsymbol{s}) = f_{\bar{g}_{\boldsymbol{r}}(\boldsymbol{p})}(g_{\boldsymbol{r}}(\boldsymbol{s})). \tag{10}$$

By using such appropriate density functions, it is immediate that the maximum likelihood selection rule given by (4) and (5) becomes *G-equivariant* in the sense of the following definition.

Definition 3.5 A selection rule *sel* is called *G-equivariant* iff for samples $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ of constraints, and every $g_{\boldsymbol{r}} \in G_n$

$$\mathit{sel}(g_{\boldsymbol{r}}(\boldsymbol{s}_1), \ldots, g_{\boldsymbol{r}}(\boldsymbol{s}_N)) = \bar{g}_{\boldsymbol{r}}(\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)). \tag{11}$$

Note that the concept of G-equivariance does not, in turn, rely on maximum likelihood selection rules. Indeed, independently from the constraints as data interpretation, G-equivariance captures the idea that when the given constraints undergo a shift in one directions, then the selected distributions should undergo a similar shift. In Figure 2, a G-equivariant rule would have to select the distribution indicated by a cross given the solid constraints iff it selects the distribution indicated by a diamond given the dashed constraints iff it selects the distribution indicated by a box given the dotted constraints.

A second homogeneity assumption one will make about $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ in the absence of any information to the contrary is *permutation invariance*: if π is any permutation of $1, \ldots, n$, then for all $\boldsymbol{p}$

$$\pi(F_{\boldsymbol{p}}) = F_{\pi \boldsymbol{p}}. \tag{12}$$

Maximum likelihood selection with respect to a permutation invariant family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ leads to a *permutation equivariant* selection rule:

$$\mathit{sel}(\pi \boldsymbol{s}_1, \ldots, \pi \boldsymbol{s}_N) = \pi(\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)). \tag{13}$$

4 Robust Estimation

In the previous section we have argued that when no particular information about the constraint generating family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ is given, then reasonable assumptions on this family are G- and permutation invariance. These assumptions alone are not nearly sufficient to identify a unique such family: if F is any distribution on S^{n-1} that satisfies $F(\pi \boldsymbol{s}) = F(\boldsymbol{s})$ for all $\boldsymbol{s} \in S^{n-1}$ and all permutations π, then F gives rise to a G- and permutation invariant family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ by letting $F_{\boldsymbol{u}} := F$, where $\boldsymbol{u} = (1/n, \ldots, 1/n)$ is the uniform domain distribution, and $F_{\boldsymbol{p}} := g_{\boldsymbol{r}}(F_{\boldsymbol{u}})$, where $g_{\boldsymbol{r}} \in G$ is the transformation uniquely determined by $\boldsymbol{p} = \bar{g}_{\boldsymbol{r}}(\boldsymbol{u})$. Conversely, every G- and permutation invariant family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ is uniquely determined by its member $F_{\boldsymbol{u}}$, which has to satisfy $F_{\boldsymbol{u}}(\pi \boldsymbol{s}) = F_{\boldsymbol{u}}(\boldsymbol{s})$.

In the following, we define a particular family $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$ by way of defining $L_{\boldsymbol{u}}$. The motivation for this family comes out of the robustness of maximum likelihood selection with respect to this family (Theorem 4.2). $L_{\boldsymbol{u}}$ can be thought of as a mixture of multivariate Laplace distributions that are separately defined on each sector. The usual multivariate Laplace distribution on $\mathbb{R}^n$ has a density $f(\boldsymbol{x})$ that depends on the Euclidean distance between $\boldsymbol{x}$ and the mean $\boldsymbol{m}$ of the distribution. To define Laplace-like distributions on the sectors of S^{n-1}, we first introduce a suitable metric on sectors:

Definition 4.1 Let $\zeta \in \{-1, 0, 1\}^n$, $\boldsymbol{s}, \boldsymbol{s}' \in S^\zeta$. Define

$$d^\zeta(\boldsymbol{s}, \boldsymbol{s}') := \left(\sum_{i,j:\, \zeta_i \neq 0, \zeta_j \neq 0} \mathit{Log}^2\left(\frac{s'_i s_j}{s'_j s_i}\right) \right)^{1/2}. \tag{14}$$

A density $l_{\boldsymbol{u}}(\boldsymbol{s})$ now is defined as a function of the distance between $\boldsymbol{s}$ and a reference constraint $\boldsymbol{m}(\zeta) \in S^\zeta$, which can be thought of as the mean constraint in sector S^ζ.

In order for $L_{\boldsymbol{u}}$ to satisfy $L_{\boldsymbol{u}}(\pi \boldsymbol{s}) = L_{\boldsymbol{u}}(\boldsymbol{s})$ for all permutations π, the $\boldsymbol{m}(\zeta)$ have to be chosen such that $\boldsymbol{m}(\pi\zeta) =$

$\pi \boldsymbol{m}(\zeta)$ for all sign vectors ζ and permutations π. Apart from this condition, no restriction has to be imposed on the choice of the $\boldsymbol{m}(\zeta)$ in order to obtain our robustness result. We therefore only assume at this point that some $\boldsymbol{m}(\zeta) \in S^\zeta$ have been appropriately fixed, and define

$$l_{\boldsymbol{u}}(\boldsymbol{s}) := exp(-d^\zeta(\boldsymbol{s}, \boldsymbol{m}(\zeta))). \quad (\boldsymbol{s} \in S^\zeta) \qquad (15)$$

The function $l_{\boldsymbol{u}}(\boldsymbol{s})$ is the density of a probability distribution $L_{\boldsymbol{u}}$, which induces a G- and permutation invariant family $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$. The maximum likelihood selection rule sel_L based on this family is distinguished by the following robustness property.

Theorem 4.2 Let $n \geq 3$. Let $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ be a G- and permutation-invariant family of probability distributions on proper constraints such that $F_{\boldsymbol{u}}(S^\zeta) > 0$ for all proper sectors S^ζ. Let $F_{\boldsymbol{p}}^\infty$ denote the distribution of an infinite sequence $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots$ of independent constraints drawn according to $F_{\boldsymbol{p}}$. Then

$$F_{\boldsymbol{p}}^\infty(lim_{N\to\infty} sel_L(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) = \boldsymbol{p}) = 1. \qquad (16)$$

The theorem says that in the long run we will select with probability 1 the correct distribution $\boldsymbol{p}$ by using sel_L, even when the constraints are actually generated according to distributions $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$. The conditions $n \geq 3$ and $F_{\boldsymbol{u}}(S^\zeta) > 0$ make sure that with probability 1 $sel_L(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ will be a unique point for all sufficiently large N. To obtain an analogous result for $n = 2$ a mild additional condition on $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ must be added. The proof of theorem 4.2 follows the proof of a general robustness result given as theorem 1 in [Huber, 1967].

Theorem 4.2 provides a good justification for using sel_L on "large" samples. It does not provide any guarantee that sel_L will show a sensible behavior on small samples. In particular, the behavior on small samples can be strongly affected by the special choice of the reference constraints $\boldsymbol{m}(\zeta)$. Thus, the definition of sel_L and Theorem 4.2 do not yet provide a full answer to the measure selection problem from the constraints as data perspective. To extend these first results towards a fully satisfactory solution, one will have to develop suitable criteria by which to judge the performance of a maximum likelihood selection rule on small samples, and to specialize or modify the definition of sel_L to obtain a selection rule that performs well according to these criteria (but retains the asymptotic behavior (16)).

5 Conclusion

We have seen that an interpretation of constraints as data, not as knowledge, leads to a completely new perspective on the measure selection problem. This perspective calls for statistical methods of parameter estimation as the tool for measure selection. The key problem we then face is that statistical methods call for a statistical model for the data generation, but that (according to the traditional problem statement that we deal with) no information about the appropriate statistical model is given. We have argued that in the absence of any such information G- and permutation invariance are natural homogeneity assumptions for the constraint distributions. $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$ is a relatively simple G- and permutation invariant family of constraint distributions that leads to a robust maximum likelihood selection rule.

Future work will have two major directions: first, the definition of sel_L will be refined in order to obtain a sensible small sample behavior of the selection rule. Second, it will be explored to which degree the assumptions made in Theorem 4.2 on the constraint generating family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ can be relaxed without loosing (16) for sel_L.

References

[Cheeseman, 1983] P. Cheeseman. A method of computing generalized Bayesian probability values for expert systems. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 198–202, 1983.

[Dickey, 1980] J. M. Dickey. Beliefs about beliefs, a theory of stochastic assessment of subjective probabilities. In J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, editors, *Bayesian Statistics*, pages 471–487. Valencia, Spain: University Press, 1980.

[Druzdzel and van der Gaag, 1995] M. J. Druzdzel and L. C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 141–148, Montreal, Quebec, Canada, 1995.

[Huber, 1967] P.J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol.1*, pages 221–233. University of California Press, 1967.

[Jaeger, 1998] M. Jaeger. Measure selection: Notions of rationality and representation independence. In *Proceedings of UAI-98*, 1998.

[Jaynes, 1982] E.T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[Lemmer and Barth, 1982] J. F. Lemmer and S. W. Barth. Efficient minimum information updating for Bayesian inferencing in expert systems. In *Proceedings of AAAI 82*, pages 424–427, 1982.

[Paris and Vencovská, 1990] J.B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:183–223, 1990.

[Rödder and Meyer, 1996] W. Rödder and C.-H. Meyer. Coherent knowledge processing at maximum entropy by SPIRIT. In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence (UAI–96)*, pages 470–476, Portland, Oregon, 1996.

[Shore and Johnson, 1980] J.E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26(1):26–37, 1980.