Constraints as Data: A New Perspective on Inferring Probabilities

Manfred Jaeger
Submitted to IJCAI – Please do not circulate

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 $D = \{d_1, \dots, d_n\}$. 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 $\boldsymbol{p} = (p_1, \dots, p_n)$ 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 $c_1, \dots, c_N$ of constraints defines the set $\Delta(c_1, \dots, c_N)$ of probability measures on D that are consistent with the constraints. Very rarely will $\Delta(c_1, \dots, c_N)$ consist of a single probability distribution. Instead, it will either contain more than one element, or be empty (when the constraints are inconsistent). A fundamental problem in probabilistic reasoning then is to select from the admissible set $\Delta(c_1, \dots, c_N)$ a single distribution $\boldsymbol{p} =: \mathit{sel}(c_1, \dots, c_N)$ as the best guess for the true distribution the constraints describe.

This 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 $\Delta(c_1, \dots, c_N)$ [**?**; **?**; **?**; **?**; **?**; **?**]. A more general class of constraints is considered by Drudzel and van der Gaag [**?**] 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 [**?**; **?**; **?**], 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 unsymmetrical. 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 principal 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 thus seen as randomly sampled pieces of information. The distribution of this constraint data is (in part) determined by the true distribution on the domain, which we want to determine (note that we are here talking about two different probability distributions: one on the domain D, and one on constraints. The latter depends on the former). 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 that 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 postulate 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 even when it is the wrong model.

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 by Jaeger [**?**], 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 [**?**], 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

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

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 (**??**) can be transformed into a unique normal form (**??**), we can also identify constraints with points $\boldsymbol{s} \in S^{n-1}$. Taking S^{n-1} as our sample space, we model randomly observed constraints by probability distributions on S^{n-1}.

In the binomial case ($n = 2$), a constraint (**??**) 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 *sign*(r) as $-1, 0$ or 1, depending on whether $r < 0$, $r = 0$, or $r > 0$. The sign vector *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, four-way distinction: a constraint $\boldsymbol{s}$ is *vacuous* iff *sign*$(s_i) \neq 1$ for all i (a vacuous constraint is satisfied by all $\boldsymbol{p} \in \Delta^n$); $\boldsymbol{s}$ is *unsatisfiable* iff *sign*$(s_i) = 1$ for all i; $\boldsymbol{s}$ is a *support constraint* iff *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 *sign*$(s_i) = 1$ and *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 **??** illustrates constraints from different sectors. Shown in the figure is the polytope Δ^3 with its 3 vertices corresponding to probability distributions that assign unit mass to one of the states in D. Six different constraints are represented by the halfplanes of points satisfying the constraint. In the figure halfplanes are shown by their boundary line, and a shading that indicates to which side of the boundary the halfplane extends. The two constraints drawn with a solid boundary line belong to the sector $S^{(-1,1,1)}$, those with a dashed boundary to the sector $S^{(1,-1,0)}$, and those with a dotted boundary to the (non-proper) sector $S^{(1,0,1)}$.

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 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 that set of support.

Figure 1: Constraints from different sectors

Our inference problem now is the following: given a sample $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ of proper constraints, determine

$$\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) \subseteq \mathit{int}\,\Delta^n \tag{4}$$

consisting of those $\boldsymbol{p}$ most likely to produce the sample. Note that we do not necessarily require that *sel*$(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ consists of a single point. Of course, one principal objective in the design of particular selection rules *sel* is to ensure that *sel*$(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ is a unique point in as many cases as possible. However, one needs to take the possibility into account that no principled statistical method can guarantee unique solutions in all cases. An even more unfamiliar aspect of (**??**) is that it is not demanded that *sel*$(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) \subseteq \Delta(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$. Such a demand, which is natural from the constraints as knowledge perspective, actually does not make much sense 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 distribution does not actually belong to $\Delta(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$.

3 Equivariance

To define concrete examples of maximum likelihood selection rules we have to assign to every parameter $\boldsymbol{p} \in \mathit{int}\,\Delta^n$ a probability distribution $F_{\boldsymbol{p}}$ on S^{n-1}, so that the likelihood of $\boldsymbol{p}$ given the sample is defined under the assumption that the constraints are independent:

$$L(\boldsymbol{p} \mid \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) := \prod_{i=1}^{N} f_{\boldsymbol{p}}(\boldsymbol{s}_i), \tag{5}$$

where $f_{\boldsymbol{p}}$ is the density of $F_{\boldsymbol{p}}$. As discussed in the introduction, we usually will not really know the underlying parametric family $(F_{\boldsymbol{p}})_{\boldsymbol{p} \in \mathit{int}\,\Delta^n}$ responsible for the constraint generation. While yet avoiding to make any specific assumptions on the form of an individual $F_{\boldsymbol{p}}$, we will show in this section that we can make a very reasonable assumption on the structure of the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$, i.e. on how two distributions $F_{\boldsymbol{p}}$ and $F_{\boldsymbol{p}'}$ are related. This will be the assumption of G-invariance.

The basic intuition underlying the concept of G-invariance is that the random mechanism that produces the constraints is uniform for all $\boldsymbol{p}$. In example **??** for instance, there is some random mechanism at work that presents us with constraints on the outcome of a mayoral election. The constraints we get to observe come as the result of a long chain of chance events: we accidentally overhear the two strangers talking as they happen to exchange the best lower bounds they happen to know for the votes for Jones and Smith. There is no reason to believe that the random events here involved depend on the actual outcome of the election, i.e. the true values of $P(\mathit{Jones})$ and $P(\mathit{Smith})$. Different outcomes will only lead to different numerical values of the bounds observed through the same sequence of chance events.

In example **??** the answers given by the various experts also are in part the product of a number of random events that do not depend on the true values of P_1 and P_2: the event that a particular expert really knows anything about polycarpia, and therefore feels qualified to state any bounds, that he knows about recent research on polycarpia, and therefore his bounds are fairly accurate, etc.

The basic assumption on the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ then is that the basic underlying random constraint generating mechanism is the same for all $\boldsymbol{p}$. The same chance sequence of events, that in the case that the true distribution is $\hat{\boldsymbol{p}}$ generates constraint $\hat{\boldsymbol{s}}$, will generate a *corresponding* constraint $\boldsymbol{s}^*$ when the true distribution is $\boldsymbol{p}^*$. In particular, $f_{\hat{\boldsymbol{p}}}(\hat{\boldsymbol{s}}) = f_{\boldsymbol{p}^*}(\boldsymbol{s}^*)$. But what constraint $\boldsymbol{s}^*$ corresponds to $\hat{\boldsymbol{s}}$ when we move from $\hat{\boldsymbol{p}}$ to $\boldsymbol{p}^*$? What we are looking for is 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 $\boldsymbol{p}^*$. This transformation should have the following two properties.

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

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

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

Both properties express a preservation of elementary qualitative properties of constraints under the correspondence expressed by g. The intuition behind sector preservation is that a given sequence of chance events will always lead to the observation of constraints of the same qualitative type, irrespective of the true distribution. Thus, pairs of corresponding constraints should belong to the same sector. Implication preservation says that logical relationships between constraints should be preserved . This means that it does not depend on the true distribution whether the constraint $\boldsymbol{s}_k$ generated by a certain sequence of chance events is redundant given previous observations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{k-1}$.

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

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

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 (**??**), without the restriction to solutions $\boldsymbol{p} \in \Delta^n$. In analogy to (**??**) we can then define *global implication preservation* of g by the condition

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

With condition (**??**) we look at constraints as defining sets of real numbers, not sets of probability distributions. In our context condition (**??**) seems to be the more pertinent one. We nevertheless here introduce the global version (**??**), 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 (**??**). A proof of this modified theorem appears to be considerably harder, however.

In light of theorem **??** we see the transformations $g_{\boldsymbol{r}} \in G_n$ as the adequate realizations of the concept of correspondence of constraints. Dual to G_n we define transformations on Δ^n.

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

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

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 $\boldsymbol{p}$,$\boldsymbol{s}$

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

Figure **??** 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.

The intuition that the transformations $g_{\boldsymbol{r}}, \bar{g}_{\boldsymbol{r}}$ formalize a canonical concept of correspondence of constraints relative to different probability distributions in Δ^n now leads to the requirement for a selection rule to be G-equivariant in the sense of the following definition.

/PSfrag wherepop(transf1)[[0(Bl)1 0]](transf2)[[1(Bl)1 0]](transf3)[[2(Bl)1 0]]3 0 1/Begin PSfraguserdict /PSfragpopputifelse

/End PSfrag Sector transformation /Unhide PSfra

Figure 2: Transformed constraints

Definition 3.4 A selection rule *sel* is called *G-equivariant* iff for samples $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$ of constraints, and every $g_{\boldsymbol{r}} \in G_n$

$$\mathit{sel}(g_{\boldsymbol{r}}(\boldsymbol{s}_1), \ldots, g_{\boldsymbol{r}}(\boldsymbol{s}_N)) = \bar{g}_{\boldsymbol{r}}(\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)). \quad (9)$$

The condition of G-equivariance formalizes the intuition that a "monotone shift" of the constraints should lead to a similar shift of the selected measure (cf. example **??**). In figure **??** 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. The condition of G-equivariance is not directly tied to the constraints as data perspective; it makes equally sense under the constraints as knowledge perspective. The constraints as data perspective, however, leads very easily to concrete G-equivariant selection rules in the form of maximum likelihood selection for G-invariant families of probability distributions. In the following definition we write $g_{\boldsymbol{r}}(F_{\boldsymbol{p}})$ for the transformation of the distribution $F_{\boldsymbol{p}}$ induced by $g_{\boldsymbol{r}}$.

Definition 3.5 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 $g_{\boldsymbol{r}} \in G$

$$g_{\boldsymbol{r}}(F_{\boldsymbol{p}}) = F_{\bar{g}_{\boldsymbol{r}}(\boldsymbol{p})}. \quad (10)$$

G-invariance of the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ finally captures our intuition that the random process that generates the constraints is uniform for all $\boldsymbol{p}$, and that when a certain random sequence of events leads to the observation of constraint $\boldsymbol{s}$ when the true distribution is $\boldsymbol{p}$, then the same sequence of events will lead to observation of $g_{\boldsymbol{r}}(\boldsymbol{s})$ when the true distribution is $\bar{g}_{\boldsymbol{r}}(\boldsymbol{p})$. By representing the measures $F_{\boldsymbol{p}}$ via densities $f_{\boldsymbol{p}}$ with respect to a suitable underlying measure on S^{n-1}, condition (**??**) becomes equivalent to

$$f_{\boldsymbol{p}}(\boldsymbol{s}) = f_{\bar{g}_{\boldsymbol{r}}(\boldsymbol{p})}(g_{\boldsymbol{r}}(\boldsymbol{s})) \quad (11)$$

for all $\boldsymbol{r}, \boldsymbol{p}, \boldsymbol{s}$. When (**??**) holds, then

$$\mathit{sel}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) := \{\hat{\boldsymbol{p}} \mid L(\hat{\boldsymbol{p}} \mid \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N) =_{\boldsymbol{p}} L(\boldsymbol{p} \mid \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)\}. \quad (12)$$

is a G-equivariant selection rule.

4 Robust Estimation

In the previous section we have seen that our conceptual model of constraints being generated by a random mechanism that works uniformly for all $\boldsymbol{p}$ leads to the structural assumption of G-invariance for the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$. This is not enough to derive a concrete selection rule, as for (**??**) we still need to fix particular densities $f_{\boldsymbol{p}}$. Because of condition (**??**) this is equivalent to fixing the density $f_{\boldsymbol{p}}$ for a single distribution $\boldsymbol{p}$, e.g. for the uniform distribution $\boldsymbol{p} = \boldsymbol{u} := (1/n, \ldots, 1/n)$.

Unfortunately, there do not seem to be many reasonable assumptions that one could make about $F_{\boldsymbol{u}}$ in the absence of any specific information about the constraint generating mechanism. One assumption one would actually be compelled to make because of the absence of specific information, however, is that $F_{\boldsymbol{u}}$ is permutation invariant, i.e. symmetric with regard to all states of the domain.

We shall now introduce a particular distribution $L_{\boldsymbol{u}}$, and propose to conduct measure selection by maximum likelihood inference for the G-invariant family $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$ defined by $L_{\boldsymbol{u}}$. The motivation for using this particular family is not derived from a particularly strong conviction that these $L_{\boldsymbol{p}}$ capture very accurately the actual constraint generating mechanism. The reason for using this family lies in the robustness property of the ensuing selection rule, which is formulated in theorem **??** below. This robustness property says that even when the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$ according to which constraints are generated is different from $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$, inferences based on $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$ will still be correct in the long run.

To define $L_{\boldsymbol{u}}$, we begin with the introduction of a metric on sectors.

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

$$d^\zeta(\boldsymbol{s}, \boldsymbol{s}') := \left(\sum_{i,j:\, \zeta_i \neq 0, \zeta_j \neq 0} \mathit{Log}^2\left(\frac{s'_i s_j}{s'_j s_i}\right) \right)^{1/2} \quad (13)$$

The density $l_{\boldsymbol{u}}(\boldsymbol{s})$ now is defined for $\boldsymbol{s} \in S^\zeta$ as a function of the distance between $\boldsymbol{s}$ and a reference constraints $m(\zeta) \in S^\zeta$. The constraint $m(\zeta)$ then is going to be the constraint with maximal likelihood in sector S^ζ, and can be thought of as the expected constraint in sector S^ζ. To make the resulting distribution $L_{\boldsymbol{u}}$ permutation invariant, the constraint $m(\zeta)$ must be defined by two constants $m^+ > 0$ and $m^- < 0$, so that $m(\zeta)_i = m^+$ if $\zeta_i = 1$, $m(\zeta)_i = m^-$ if $\zeta_i = -1$, and $m(\zeta)_i = 0$ if $\zeta_i = 0$.

For the purpose of the present paper no additional restrictions on the $m(\zeta)$ are necessary. Particular choices for the $m(\zeta)$ will affect the behavior of the resulting selection rule on small samples, but do not affect the asymptotic robustness result.

Thus, let $m(\zeta)$ be given for every sign vector ζ, and define

$$l_{\boldsymbol{u}}(\boldsymbol{s}) := \mathit{exp}(-d^\zeta(\boldsymbol{s}, m(\zeta))) \qquad (\boldsymbol{s} \in S^\zeta) \quad (14)$$

One can show that the functions (**??**) define a probability density on S^{n-1} with regard to the same reference measure that was already needed for condition (**??**). We shall not go into the details here, but only mention that what we have defined here are essentially Laplace distributions on every sector, which are combined into a distribution on the sphere.

Defining $L_{\boldsymbol{u}}$ via (**??**) provides us with a G-invariant family $(L_{\boldsymbol{p}})_{\boldsymbol{p}\in \mathit{int}\,\Delta^n}$ with densities $l_{\boldsymbol{p}}$. With these densities we can finally define by (**??**) and (**??**) a concrete selection rule realizing the maximum likelihood approach. We denote it by *sel*$_{ml}$.

The first question about *sel*$_{ml}$ we have to address, is under what conditions *sel*$_{ml}(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N)$ will be a unique point. The answer to this question is a little bit involved, and we will here only give a rough sketch of what it looks like. By a suitable parameterization, constraints $\boldsymbol{s} \in S^{n-1}$ and distributions $\boldsymbol{p} \in \Delta^n$ can both be identified with points in $n-1$-dimensional space. The distributions $\boldsymbol{p}$ that maximize $L(\boldsymbol{p} \mid \boldsymbol{s}_1,\ldots,\boldsymbol{s}_N)$ then correspond to the points that minimize the sum of the Euclidean distances to the points corresponding to the $\boldsymbol{s}_i$. This sum is minimized by a unique point in $n-1$-space provided that $n \geq 3$ and that the points defined by the constraints are not all collinear. A sufficient condition for *sel*$_{ml}(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N)$ to be uniquely defined therefore is that $n \geq 3$, and that $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N$ satisfy a certain “richness” condition which precludes collinearity in the new parameterization.

The main benefit of working with the family $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$ lies in the robustness of the resulting selection rule.

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

$$F^{\infty}_{\boldsymbol{p}}(\mathit{lim}_{N\rightarrow\infty}\mathit{sel}_{ml}(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N) = \boldsymbol{p}) = 1. \qquad (15)$$

The conditions $n \geq 3$ and $F_{\boldsymbol{u}}(S^{\zeta}) > 0$ make sure that with probability 1 *sel*$_{ml}(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_N)$ will be a unique point for all sufficiently large N. A result similar to theorem **??** can also be obtained for $n = 2$, but this requires an additional condition on $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$. In statistical terminology, (**??**) says that the estimator *sel*$_{ml}$ is *consistent* for the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$. Consistency properties typically hold for maximum likelihood estimators. The remarkable point of (**??**) is that *sel*$_{ml}$ is defined by maximizing a likelihood function derived from the family $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$, whereas the samples are generated by a member from the family $(F_{\boldsymbol{p}})_{\boldsymbol{p}}$. The proof of theorem **??** follows the proof of a general robustness result given as theorem 1 in [**?**].

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. We proposed maximum likelihood inference for the family $(L_{\boldsymbol{p}})_{\boldsymbol{p}}$ as one particular such method. While the robustness property of this method is a very attractive feature, there is no reason to believe that it is the only selection rule that has such a property. Moreover, asymptotic robustness provides no guarantee that the selection rule shows a sensible behavior on small samples. The question of what ultimately will prove to be the best selection rule under the constraints as data perspective therefore is still wide open at this point. To answer this question (possibly in the negative by realizing that no best rule exists), we first need to find additional useful criteria by which to judge the performance of a selection rule.