Twisting Statistics with Properties

B. Apolloni[1], D. Malchiodi[1]
[1] - Dip. di Scienze dell'Informazione Università degli Studi di Milano,
Via Comelico 39/41 20135 Milano, e-mail: {apolloni,malchiodi}@dsi.unimi.it

ABSTRACT

We give three steps in the direction of shifting probability from a descriptive tool of unpredictable events to a means for understanding them. At a very elementary level we state an operational definition of probability based solely on symmetry assumptions about observed data. This definition converges, however, on the Komogorov one within a special *large number law* that represents a first way of twisting features observed in the data with properties expected in subsequent observations. Within this probability meaning we fix a general *sampling mechanism* to generate random variables and extend our twisting device to computing probability distributions on population properties on the basis of the likelihood of the observed features. Here the randomness core translates from the above symmetry assumptions into a generator of unitary uniform random variables. The function mapping from these elementary to our more complex variables is exactly the object of our inference. Using this framework we revisit the basic linear regression problem; at the same time, however, we are capable of appreciating confidence intervals in the case of Gumbel or similar assumptions about the distribution law of measurement errors. At the other complexity extreme, we also give some initial directions for designing efficient learning algorithms on neural networks. Aiming to discover suitable features (which are classically defined as *sufficient statistics*), we refer directly to the notion of Kolmogorov complexity and *coding theorem* in particular. This is to connect the features to the inner structure of the observed data in terms of their concise codes. Thus we are able to shed some light on the current problem of splitting the learning task suitably into a subsymbolic part perfomed for instance by a neural network and a symbolic one done with symbolic models possibly complexer than but as clearly defined as the mentioned regression line.

Key words: Computational learning, Learning rules, Algorithmic inference, Kolmogorov complexity.

1 FIRST STEP

The estimate $\widehat{p}$ for the mean p of a random variable X distributed according to a Bernoulli law upon the presentation of a sample $\mathbf{X}_M = (X_1, \ldots, X_m)$ is usually derived counting the number $k = \sum_{i=1}^{m} X_i$ of 1's in the observed sample and then setting $\widehat{p} = k/m$. An alternative way for *appreciating* it can be found by considering that $\widehat{p}$, referring to the probability that the next observed bit X_{m+1} will assume the value 1, can be found through analysing the fact that for such X_{m+1} we would observe exactly $k+1$ ones within a $m+1$ sized sample. Assuming that the sequencing of zeroes and ones in the sample is inessential to the observed phenomenon, we appreciate $p_1 = \mathrm{P}[X_{m+1} = 1]$ as the ratio between the number of those permutations of the sampled values having the last element equal to one and the number of all permutations. Namely

$$\widehat{p}_1 = \frac{m!(k+1)}{(m+1)!} = \frac{k+1}{m+1}$$

Analogously, we appreciate $p_0 = \mathrm{P}[X_{m+1} = 0]$ through

$$\widehat{p}_0 = \frac{m!(m-k+1)}{(m+1)!} = \frac{m-k+1}{m+1}$$

Note that $\widehat{p}_0 + \widehat{p}_1 = \frac{m+2}{m+1}$; that is, the estimated probabilities do not sum up to 1: this is due to the fact that they refer to different probability spaces. To put this idea in a more rigorous form, first we introduce an incremental definition for the sample space:

Definition 1.1. *The* symmetric sample space *for a statistical experiment providing a sample* $\mathbf{X}_m = (X_1, \ldots, X_m)$ *is a pair* $(\Omega_m, \mathscr{B}_m)$ *where*

- Ω_m *is the set of all the* allowed *permutations of the* m*-tuplet* $(X_1, \ldots, X_m)$ [1];
- $\mathscr{B}_m$ *is a* σ*-field of subsets of* Ω_m.

In this context we define the probability measure P on $(\Omega_m, \mathscr{B}_m)$ per usual as the ratio between the number of interesting sample points and the number

[1] i.e. those permutations that do not change the meaning of the sequence.

of the possible ones. Given $\mathbf{X}_m$ as a sequence of independent bits (i.e. all permutations are allowed), in this framework the above alternative estimation of probability that the next observed bit will assume the value 1 has its natural environment in the probability space $(\Omega^1_{m+1}, \mathscr{B}_{m+1}, \mathrm{P})$, where Ω^1_{m+1} is the set of all the permutations of the $(m+1)$-tuplet $(X_1, \ldots, X_m, 1)$, while $\mathscr{B}_{m+1}$ and P are defined, as above described, over Ω^1_{m+1}.

Analogously, the probability that the next observed bit will assume the value 0 has to be calculated in the measurable space $(\Omega^0_{m+1}, \mathscr{B}_{m+1})$, where now Ω^0_{m+1} is the set of all the permutations of the $(m+1)$-tuplet $(X_1, \ldots, X_m, 0)$.

More in general, the probability that K of the following M bits will assume the value 1 will be estimated through

$$\widehat{p}_{(K;M,k,m)} = \frac{\binom{m}{k}\binom{M}{K}}{\binom{m+M}{k+K}} = \frac{\binom{k+K}{k}\binom{m-k+M-K}{m-k}}{\binom{m+M}{m}} \quad (1)$$

computed in the space $(\Omega_{m+M}, \mathscr{B}_{m+M}, \mathrm{P})$, where Ω_{m+M} is the set of all the permutations of the $(m+M)$-tuplet $(X_1, \ldots, X_m, X_{m+1} \equiv 1, \ldots X_{m+K} \equiv 1, X_{m+K+1} \equiv 0 \ldots, X_{m+M} \equiv 0)$.

The links between this inference of the empirical probability and the Kolmogorov one [17] arise asymptotically:

- When $m \to \infty$, the Kolmogorov sample space is the set of all possible values which can be assumed by the elements of the strings which constitute our sample space.
- When $m \to \infty$, the two definitions of estimated and appreciated probability converge to the same value.
- When $M, K \to \infty$, $\widehat{p}_{(K;M,k,m)} = p_{(k;m,M,K)}$ tends to the Binomial distribution law $p_{(k;m,K/M)}$ with the sample size and the asymptotic 1's frequency in the population for parameters.

To appreciate the relationship between the two definitions for small sample and small population (as usual we denote by this term the subsequent M bits) we generated a variety of pairs (sample, population) from a Bernoulli variable with p ranging from 0 to 1 under the constraint of both $k+K$ and $m+M$ as in (1) being constant. In Figure 1 we see that, as expectable, probabilities appreciated through (1) go around, but do not coincide with those classically estimated through frequency. The reasons why we prefer this new inferential way of finding the empirical probability are the following:

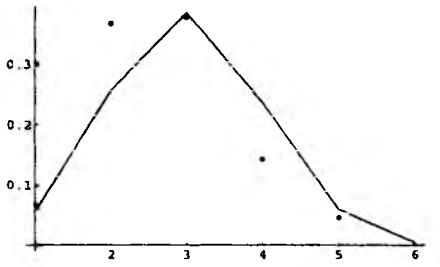

Figure 1. Relationship between appreciated and estimated probability in 60 families of (sample, population), each obtained by sampling from a Bernoulli variable where p rises from 0 to 1 with step 0.01. With reference to Equality 1, $m = 20$, $M = 5$, $k + K = 9$. Horizontal axis: K, vertical axis: both frequencies (bullets) and joined values of $\widehat{p}_{M,K}$ (line).

- We don't need to suppose the existence of an intrinsical probability, but we can perform our inferential method only on the basis of the observed data;
- it makes sense also for small size samples;
- future and past play the same role because the above defined space is made of the global strings of data.

2 Second step

The typical inference framework is met when $M \to \infty$ and m is small. In this case the object of our inference is a (possibly infinite) string of data X that we partition in a prefix which we assume to be known at the current time (and therefore call sample), and an infinite suffix of unknown data which concerns the future that we call population (see Figure 2). All these data share the feature of being independent observations of a same phenomenon. Therefore, in the limit of the convergence of the probability to the target of large number law, without loss of generality we assume these data as the output of a same function g_ϑ having input from a set of independent random variables U uniformly distributed in the unitary interval – effectively, the most essential source of randomness[2].

[2]Such a g_ϑ always exists by the probability integral transformation theorem [12]. By default capital letters will denote random variables and small letters their corresponding realizations.

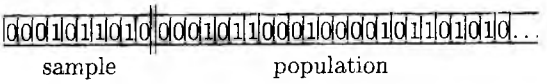

Figure 2. Sample and population of random bits.

We will refer to $\mathscr{M} = (U, g_\vartheta)$ as a *sampling mechanism* and to g_ϑ as an *explaining function*, and this function is precisely the object of our inference. Let us consider, for instance, the sample mechanism $\mathscr{M} = (U, g_p)$, where

$$g_p(u) = \begin{cases} 1 & \text{if } u \leq p \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

explains sample and population distributed according to a Bernoulli law of mean p like in Figure 2. As shown in Figure 3, for a given sequence of U's we obtain different binary strings depending on the height of the threshold line corresponding to p. Thus it is easy to desume the following implication chain

$$(K_{\tilde{p}} \geq k) \Leftarrow (p < \tilde{p}) \Leftarrow (K_{\tilde{p}} \geq k+1) \qquad (3)$$

and the consequent bound on the probability

$$P[K_{\tilde{p}} \geq k] \geq P[p < \tilde{p}] \geq P[K_{\tilde{p}} \geq k+1] \qquad (4)$$

which characterizes the cumulative distribution function (c.d.f.) F_p of the parameter p. In our statistical framework indeed, the unknown p is a random variable in $[0, 1]$ representing the asymptotic frequency of 1 in the populations that are compatible, as a function of U suffix of the sample, with the number k of actually observed 1. Here $K_{\tilde{p}}$ denotes the random variable counting the number of 1's in the sample if the threshold in the explaining function switches to $\tilde{p}$ *for the same realizations* of U.

Note the asymmetry in the implications. It derives from the fact that:

- raising the threshold parameter in g_p cannot decrease the number of 1 in the observed sample, but
- we can recognize that such a rise occurred only if we really see a number of ones in the sample greater than k.

We will refer to every expression similar to (3) as a *twisting argument*, since it allows us to exchange events on parameters with events on statistics.

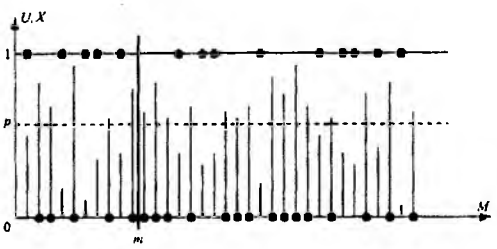

Figure 3. Generating a Bernoullian sample. Horizontal axis: index of the U realizations; vertical axis: both U (lines) and X (bullets) values. The threshold line p realizes a mapping from U to X through (2).

Twisting sample with population properties is our approach to statistical inference, which we call *algorithmic inference*. Its general framework is depicted in Figure 4. For any sampling mechanism, we have on the one hand the *world of hypotheses* about

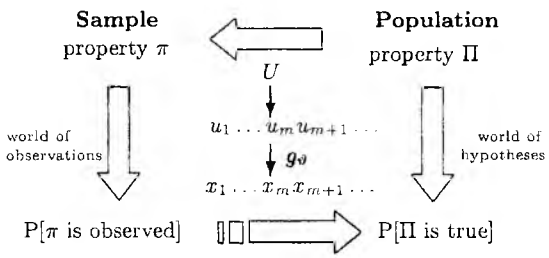

Figure 4. Twisting properties between sample and population.

g_ϑ that results in special properties of the population, which we call Π: on the other, the *world of actual observations* where - as g_ϑ is the same - the above hypotheses result in corresponding properties π on the sample. So we can use the likelihood of the actual sample in respect to π, a quantity that in principle can be easily computed when the hypotheses are fully specified, to get the probability that the corresponding Π are satisfied.

The theorem below states that, under weak regularity conditions on the data (see footnote below), the general form of a twisting argument must be grounded on sufficient statistics [17]. Let us start characterizing sufficiency as follows:

Definition 2.1. *For a given parameter set Θ, and explaining function g_ϑ with $\vartheta \in \Theta$, let X be the generated random variable, $f_X(.;\vartheta)$ its probability density, and $\mathbf{S}_m$ a sample of size m drawn from X within a space $\mathfrak{X}$. A statistic $T : \mathfrak{X}^m \mapsto \mathbb{R}$, inducing the partition $\mathfrak{U}(T)$ on $\mathfrak{X}^m$, is said to be* sufficient *with reference to the parameter ϑ if the ratio*

$f_{\mathbf{S}_m}(\mathbf{x}^1;\vartheta)/f_{\mathbf{S}_m}(\mathbf{x}^2;\vartheta)$ *does not depend on* ϑ *when* $\mathbf{x}^1$ *and* $\mathbf{x}^2$ *belong to a same element of* $\mathfrak{U}(T)$*;*

Theorem 2.1. *[7] Let* $\mathscr{P} = \{\mathrm{P}_\vartheta, \vartheta \in \Theta\}$ *be a regular family of probability measures on a random variable* X, $\mathbf{S}_m = (X_1, \ldots, X_m)$ *a sample drawn from* $\mathrm{P}_\vartheta \in \mathscr{P}$ *with sample space* $\left(D_X^\vartheta\right)^m$, *and* $T = T(\mathbf{S}_m)$ *a statistic. With reference to a fixed sampling mechanism* (U, g_ϑ) *and denoting with* $T_{f_A}^{-1}(z)$ *the set of* $a \in A$ *such that* $T(f_a) = z$ *(eventually* $f_a \equiv a$*)*

- *the twisting argument*

$$\forall m \in \mathbb{N}, \forall \mathbf{u} = (u_1, \ldots, u_m) \in [0,1]^m \text{ such that } g_\vartheta(\mathbf{u}) = (g_\vartheta(u_1), \ldots, g_\vartheta(u_m)) = (x_1, \ldots, x_m) \text{ for a proper } \vartheta, \text{ and } T(x_1, \ldots, x_m) = t$$

$$(T(g_{\widetilde{\vartheta}}(\mathbf{u})) \geq t) \Leftarrow (\vartheta < \widetilde{\vartheta}) \Leftarrow (T(g_{\widetilde{\vartheta}}(\mathbf{u})) \geq t') \quad \text{for almost every } \mathbf{u} \quad (5)$$

- *provided that, i) for the mentioned* $\mathbf{u}$ *and corresponding* t, *for each* $\vartheta \in \Theta$ *either* $T_{\mathbf{S}_m}^{-1}(t) \subseteq \left(D_X^\vartheta\right)^m$ *or* $T_{\mathbf{S}_m}^{-1}(t) \subseteq \overline{\left(D_X^\vartheta\right)^m}$ *holds, ii)* $T^{-1}_{g_\vartheta(\mathbf{u})}(t') \neq \emptyset$, *and iii)* $t \leq t' \leq t + l$, *for* l *suitable (namely, for each* $r \in (0,1)$, *there exists a sample size* m_0 *such that for each sample* $\mathbf{u}$ *of size* $m > m_0$ l *divided by the range of* T *with* ϑ *is uniformly bounded by* r*),*

- *can be stated only if* T *is a function of a sufficient statistic for* Θ [3].

In this inferential approach we recover the key notion of $1-\delta$ confidence interval for the parameter ϑ, intended as the pair of values (L_i, L_s) such that:

$$\mathrm{P}[L_i < \vartheta < L_s] \geq 1 - \delta$$

In particular:

- From (4), to compute confidence intervals for p we choose (L_i, L_s) such that

$$\sum_{i=k+1}^{m} \binom{m}{i} L_s^i (1-L_s)^{m-i} = 1 - \frac{\delta}{2} \quad (6)$$

$$\sum_{i=k}^{m} \binom{m}{i} L_i^i (1-L_i)^{m-i} = \frac{\delta}{2} \quad (7)$$

[3] See [7] for a more complete statement of the claim including both necessary and sufficient conditions. In the same paper details can be found on regularity conditions as well.

Here the random variable is exactly p, and the confidence refers to the possible suffix of a given sample observed on X. This is highlighted in Figure 5, where we considered a string of 20+200 unitary uniform variables representing, respectively, the randomness source of a sample and a population of Bernoulli variables. Then, according to the explaining function (2) we computed a sequence of Bernoullian 220 bits long vectors with p rising from 0 to 1. The pairs $k/20$ and $h/200$, computing the frequency of ones in the sample and in the population respectively, are reported along one fret line in the figure. We repeated this experiment 20 times (using different vectors of uniform variables). Then we drew on the same graph the solutions of Equations 6 and 7 with respect to p with varying k and $\delta = 0.1$. As we can see, for a given value of k the intercepts of the above curves with a vertical line with abscissa $k/20$ determine an interval containing almost all intercepts of the frets with the same line. A more intensive experiment would show that, in the approximation of $h/200$ with the asymptotic frequency of ones in the suffixes of the first 20 sampled values, on all samples, and even for each sample if we draw many suffixes of the same one, almost $100(1-\delta)$ percent of the frets fall within the analytically computed curves.

- If the sampling mechanism induces a linear relation between the x and y components of the observed data, a sample $\mathbf{S}_m$ can be described as follows:

$$\mathbf{S}_m = \{(x_i, y_i) | y_i = a + b(x_i - x) + \varepsilon_i, i = 1, ..., m\} \quad (8)$$

where $\bar{x}$ denotes the sample mean, a and b are specifications of two random variables, which we call respectively A and B, not depending on the single observation, while ε_i is the random noise moving the coordinate pairs far from the regression line. In this case, denoting with $\widetilde{y}_i$ the value assumed by the observation y_i when unknown parameters shift from a and b to $\widetilde{a}$ and $\widetilde{b}$ respectively in the sampling mechanism, we can easily check that

$$(a \leq \widetilde{a}) \Leftrightarrow \left(\sum_{i=1}^{m} y_i \leq \sum_{i=1}^{m} \widetilde{y}_i\right) \quad (9)$$

$$\left(b \leq \widetilde{b}\right) \Leftrightarrow \left(\sum_{i=1}^{m} y_i (x_i - \bar{x}) \leq \sum_{i=1}^{m} \widetilde{y}_i (x_i - \bar{x})\right) \quad (10)$$

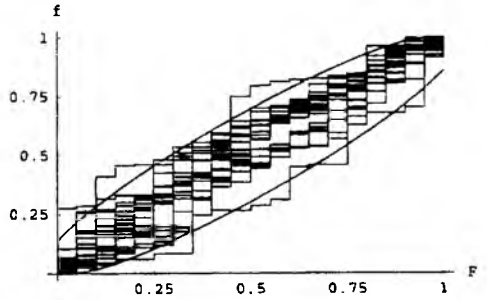

Figure 5. Generating 0.9 confidence intervals for the mean p of a Bernoulli random variable with population and sample of $n = 200$ and $m = 20$ elements, respectively.
$F = k/m$ = frequency of ones in the sample; $f = h/n$ = frequency of ones in the population.
Fret lines: trajectories described by the number of ones in sample and population when p ranges from 0 to 1, for different sets of initial uniform random variables. Curves: trajectories described by the confidence interval extremes when the observed number k of 1 in the sample ranges from 0 to m.

According to Theorem 2.1, logical relations (9) and (10) represent a twisting argument if $\sum_{i=1}^{m} y_i$ and $\sum_{i=1}^{m} y_i(x_i - \overline{x})$ are weak minimal sufficient statistics. This happens when for instance ε is assumed Gaussian. In this case, after introducing the random variables $S_E = \sum_{i=1}^{m} \varepsilon_i$, and $S'_E = \sum_{i=1}^{m} \varepsilon_i(x_i - \overline{x})$, we have

$$F_A(\tilde{a}) = 1 - F_{S_E}\left(\sum_{i=1}^{m} y_i - m\tilde{a}\right) \qquad (11)$$

$$F_B(\tilde{b}) = 1 - F_{S'_E}\left(\sum_{i=1}^{m} y_i(x_i - \overline{x}) - \tilde{b}\sum_{i=1}^{m}(x_i - \overline{x})^2\right)$$

Moreover, since

$$\left(a + b(x - \overline{x}) \le \tilde{a} + \tilde{b}(x - \overline{x}),\ \forall x \ge \overline{x}\right) \Leftrightarrow \left(a \le \tilde{a} \wedge b \le \tilde{b}\right)$$

and

$$\left(\exists a', b' \text{ s.t. } (a' + b' \le \tilde{k}) \wedge (a + b(x - \overline{x}) \le a' + b'(x - \overline{x})\, \forall x \ge \overline{x})\right) \Leftrightarrow \left(a + b \le \tilde{k}\right)$$

after having introduced the random variable

$$S''_E = \varepsilon_i\left(1 + m\frac{x_i - \overline{x}}{\sum_{i=1}^{m}(x_i - \overline{x})^2}\right)$$

the following relation holds for $I^* = \{(a^*, b^*)\}_I = \{\arg\sup\{a' + b', (a', b') \text{ s.t. } a' + b' \le k\}, a' \in I\}$, where I is an assigned interval, ruled by 11:

$$\mathrm{P}[A + B(x - \overline{x}) \le a^* + b^*(x - \overline{x}) \quad \forall x \ge \overline{x} \text{ for some } (a^*, b^*) \in I^*] = 1 - F_{S''_E}\left(\sum_{i=1}^{m} y_i + m\frac{\sum_{i=1}^{m} y_i(x_i - \overline{x})}{\sum_{i=1}^{m}(x_i - \overline{x})^2}\right) \qquad (12)$$

Coupling this with the analogous equation for $x \le \overline{x}$, we obtain a confidence interval for the whole regression line as in Figure 6. Finally, considering the further shift of the single observed point from the regression line, we obtain the larger butterfly region in the figure representing the envelope of the confidence intervals for these points. Comparing the two pairs of (conventional and our approach) regions in Figure 6, we note that the algorithmic inner region is designed to contain the original regression line in full, while the standard counterpart is the union of separate confidence intervals drawn for each x. The quadratic shape of the outcoming borders may induce some shadow zones in this region, such that no line fully contained in the confidence region passes through them and, in any case, may promote the border crossing by the source line as in the figure.

- Consider the non homogeneously exponential random variable T [4] with associated c.d.f.

$$F_T(t) = 1 - \mathrm{e}^{-\frac{t}{\beta(t)}} \qquad (13)$$

[4] Related to the distribution of survival data in breast cancer treating [9].

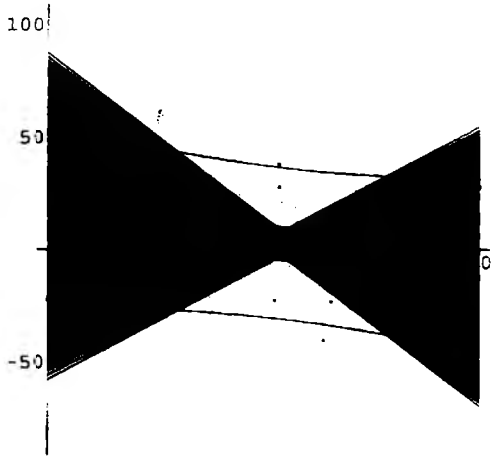

Figure 6. Algorithmic confidence regions for the Gaussian driven regression sample $\{(x_i, y_i)$ such that $y_i = 5 + 5(x_i - \overline{x}) + \varepsilon_i$, $i = 1, \ldots, 20\}$.
with $\sigma = 20$ and x_i uniformly drawn in $[0, 20]$.
Horizontal axis: x values; vertical axes: y values.
Bold line: source regression line.
Dark shadow region: 90% algorithmic confidence region for regression line.
Light shadow region: 90% algorithmic confidence region for random points.
Dark curve: 90% standard confidence region for regression line.
Light curve: 90% standard confidence region for random points.

where $\beta(t) = \beta_0 \beta_1^{-\log t}$, $\beta_1 > e^{-1}$ and $\beta_0 > 0$. For a sample $(t_1, \ldots, t_m)$ drawn from T, we can express $\beta(t)$ as a function of

$$\overline{\log t} = \frac{1}{m} \sum_{i=1}^{m} \log t_i$$

through the form

$$\beta(t) = \overline{\beta}_0 \beta_1^{-(\log t - \overline{\log t})},$$

where $\overline{\beta}_0 = \beta_0 \beta_1^{-\overline{\log t}}$ replaces β_0 as one of the two objects of our inference. The benefit of this representation comes from the fact that, according to the inverse transform algorithm [13], an explaining function $g_{\overline{\beta}_0, \beta_1}$ for T associates to every seed u_i the solution of

$$\overline{\beta}_0 \beta_1^{-(\log t_i - \overline{\log t})}(-\log(1 - u_i)) = t_i$$

with respect to the unknown t_i. Thus focusing on the statistic $L = \sum_{i=1}^{m} \log T_i$ and its realization l, we can state the following twisting argument for $\overline{\beta}_0$

$$\left(\overline{\beta}_0 < \widetilde{\overline{\beta}}_0\right) \Leftrightarrow \left(L_{\widetilde{\overline{\beta}}_0} > l\right)$$

where $L_{\widetilde{\overline{\beta}}_0}$ denotes the value assumed by L if the parameter value shifts from $\overline{\beta}_0$ to $\widetilde{\overline{\beta}}_0$. The distribution of L is derived through the following steps:

1. $\log T$ has the c.d.f.

$$F_{\log T}(t) = 1 - F_{G(a,b)}(2a - t)$$

where $a = \frac{\log \beta_0}{1 + \log \beta_1}$, $b = \frac{1}{1 + \log \beta_1}$ and $F_{G(a,b)}$ denotes the c.d.f. of a Gumbel distribution with parameters a and b.

2. The sum of Gumbel distributions is well approximated by a Gaussian law, even for small values of m.

Analogously, the twisting argument $(\beta_1 < \widetilde{\beta}_1) \Leftrightarrow (H_{\widetilde{\beta}_1} \prec h)$ is based on the statistic

$$H_{\beta_1}(T_1, \ldots, T_m) = \frac{\sum_{i=1}^{m} (\log T_i - \overline{\log T}) \log(-\log(1 - U_i))}{\sum_{i=1}^{m} (\log T_i - \overline{\log T})^2} - 1$$

where $(U_1, \ldots, U_m)$ are the uniform random variables used to describe T through the sampling mechanism and $\prec$ is one of the two usual order relations $<$ or $>$[5].

Grouping these results with those in the previous section, we can derive a confidence region for $\beta(t)$ taking note of the following:

1. The distribution of H, though not analytically known in principle, can be approximated by an empirical cumulative distribution obtained after a suitable simulation process.
2. This simulation as well as computations on F_L need the prior knowledge of β_1, which has been approximated through its *maximum likelihood* estimator [17] .

[5] Actually we cannot prove the sufficiency of H; however empirical simulations show the essentially monotone behavior (either increasing or decreasing) of the statistic values when the model parameters increase.

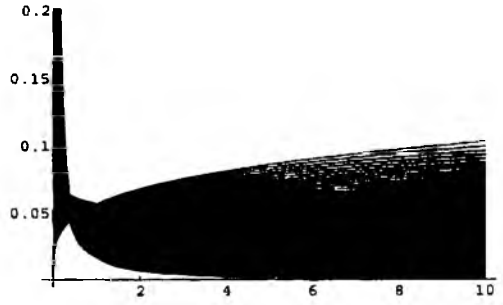

Figure 7. 0.9 confidece region for $\beta(t)$ computed basing on a sample of size 21 drawn from a clinical record, assuming F_T as in 13. Dark area: confidence region; light plot: shape of $\beta(t)$ for ML estimates $\widehat{\beta}_0 = 0.04$ and $\widehat{\beta}_1 = 1.45$.

3. The inference for the whole $\beta(t)$ is derived focusing on the quantity $\log \boldsymbol{\beta(t)}$, which turns out to be a corresponding function of the regression line in the usual frameworks, depending on the parameters $\log \beta_0$ and $\log \beta_1$.

Figure 7 shows such a derived confidence region.

- At a subsymbolic level, when working with neural networks, with reference to relation (5) we check inequalities on statistics just by changing the parameters of the network. A main problem is to check whether the error function which drives the visit of the parameter space is a sufficient statistic or not. If it is, we can assume this statistic as a reliable indicator of the closeness of the current parametrization of the neural network to the function to be learnt. Otherwise, we incur the usual drawbacks such as relative minima of the statistic, overfitting and so on. This is the reason why Boltzmann machines [1] and statistics consisting of the Kullback relative entropy are gerenally preferred to learn probability distributions.

3 Third step

The twisting argument leaves us with the crucial problem of finding sufficient statistics. This problem finds a straighforward solution through the factorization lemma [18] when we work with easy probabilistic models. On the contrary, the same lemma does not enjoy manageable results when the distribution law in hands is explained by complex functions, for instance computing the solution of NP-hard problems [6]. In this section we give a dual issue of the factorization lemma in our statistical framework. It sheds light on new methods for finding sufficient statistics or approximations of them. The defintion of sufficiency simply says that, when looking at a useful sample property (a statistic indeed) we must focus on properties that remain unchanged on samples having the same occurrence probability of the observed one. If we do not know this probability we can estimate it through a maximum likelihood principle as follows.

Consider the following lemma putting in relation probability with complexity of a string.

Definition 3.1. *[6] Let $\mathfrak{X}$ be the set of all binary strings and $|x|$ the length of the string x. Denote with $f(x) < \infty$ the fact that f is defined on x. A partial recursive function (prf) $\phi : \mathfrak{X}^* \to \mathbb{N}$ is said prefix* [7] *if $\phi(x) < \infty$ and $\phi(y) < \infty$ implies that x is not a proper prefix of y. Fixed a universal prefix prf U* [8]*, the conditional Prefix (or Levin's) Complexity $K(x|y)$ of x given y is defined as*

$$K(x|y) = \min_{p \in \mathfrak{X}} \{|p| \text{ such that } U(p, y) = x\}, \qquad (14)$$

and the unconditional Prefix Complexity $K(x)$ of x as

$$K(x) = K(x|\lambda). \qquad (15)$$

where λ is the empty string.

Lemma 3.1. *[6] The probability measure* P *of any string $x \in \mathfrak{X}$ explained by the function $g_\vartheta(x)$ is related to the prefix complexity K of x and g_ϑ through the following equation:*

$$\mathrm{P}[x] \leq 2^{-K(x) 2^{K(g_\vartheta)}} \qquad (16)$$

The lemma comes from the fact that $-\log(\mathrm{P}[x])$ can be used as a prefix code of x in a prefix machinery having the description of g_ϑ in its library, and this machinery can be simulated by a universal prefix machinery U by running a code of length $K(g_\vartheta)$. Thus, a sequence of length $-\log(\mathrm{P}[x]) + K(g_\vartheta)$ can be used to code x in the reference machinery U of Definition 3.1. Of course, in respect to this machinery the shortest code of x has a length $K(x)$ no greater than the above.

Though both $K(x)$ and $K(g_\vartheta)$ are not computable in general by definition [9] [6], we will use

[6] This is the case of the distribution law of pairs of random instance and solution of a knapsack problem [8].

[7] Where $\mathfrak{X}^*$ denotes the set of words obtainable concatenating symbols from the alphabet $\mathfrak{X}$.

[8] i.e. a machinery capable of computing any computable function according to the Church thesis [3].

[9] This negative result is a variant of the well-know Turing machine *halting lemma* [11].

the upper bound in Equation 16 as a ML estimate of P[x]. Therefore, our problem of finding a consistent statistic approximately coincides just with the one of reading the upper bound on the probability of a sample and identifying the function of the sampled data on which the upper bound depends. Disregarding for a moment the term $2^{K(g_\vartheta)}$, we isolate within the other upper bound factor the part we assume to be independent on the unknown aspects of the population (synthetized by the parameter ϑ) from the part depending on them (the wanted statistics). Of course, samples with this same statistic have the same probability, apart from coefficients independent on ϑ; hence these statistics result sufficient. Therefore a second approximation (an estimate indeed) consists in writing the second member of (16) in such a readable form. As mentioned before, we cannot write the minimal codes ϕ underlying $K(x)$; however we can look for very efficient programs π as estimates $\widehat{\phi}$ of ϕ that we split as follows:

$$\phi_x = \pi_{\widetilde{g}}((\pi_{h(x_1)}, ..., \pi_{h(x_m)}), \pi_{t(x_1,...x_m)}) \qquad (17)$$

with $g_{\vartheta(x)} \equiv \widetilde{g}(x;\vartheta)$, where the allotment of the computational tasks is aimed at minimizing the total π_is' length. Namely, we recognize in the efficient compression of property t of the sample the sufficient statistic evoking a general property of the whole population, while the remnant part h of x_i must be computed singularly on each variable. $\widetilde{g}$ is the part of the envisaged population property that we already know. It is a cognitive constraint that generally makes $\widehat{\phi}_x$ longer, but also a useful help in devising it. We can easily recognize in the first term of the sequence the $-\log$ of the first factor of the likelihood factorization when a sufficient statistic exists:

$$\mathrm{P}(x) = f_1(x_1, ..., x_m) f_2(t(x_1, ..., x_m), \vartheta) \qquad (18)$$

Here we further split f_2, thus allowing a balancing of description complexities of statistics, constraints, and residual unknown parts of a sample (which looks for an enlarged issue of the structural risk minimization principle introduced in [16]).

Summing up, Equation 16 reads:

$$\widehat{P}(x) = 2^{-\sum_{i=1}^{m} \widehat{K}(h(x_i)) - \widehat{K}(t(x_1,...,x_m)) - \widehat{K}(\widetilde{g})} 2^{K(g_\vartheta)} \qquad (19)$$

$2^{K(g_\vartheta)}$ is a sort of rewarding factor allowing us to assume great probability in case of complex explaining functions. However neither the true ϑ nor the true complexity value is known; thus the maximum likelihood principle requires us to give a very short global description of the sample by minimizing the total length of π as in Equation 17. In line with current thread on hybrid systems [2, 5, 15] we may imagine fulfilling this task in a subsymbolic and a symbolic step. The former accounts for what we formally know about the string sampling mechanism. The subsymbolic part must supply what still remains unkown. This a typical job of a neural network for instance. In this case a subsidiary inference task arises to estimate the parameters of this device. Thus another (hopefully sufficient) statistic joins the previous one; in other words, we realize that the global inference problem needs a pair of sufficient statistics. Learning a neural network is a non easy problem supported by an actually poor theory. In the previous section we got some insights from the twisting argument theory, but we can enjoy still poorer intuition about the joint estimation of the pair of statistics. Rather, still in the aim of minimizing our sample description, we enunciate the following "don't cheat" principle:

Principle 1. *For suitably describing a function on a training set, a formula beats a neural network only if its description length, including observed statistics for free parameters, is shorter than the neural network's.*

Example 3.1. *In force of the above principle,*

1. *The symbolic description of the XOR function, for instance through the formula "$1 - x_1x_2 - (1 - x_1)(1 - x_2)$", beats its description through a neural network described by a 2-2-1 MLP, namely "$\sigma(5.52(\sigma(-1.49x_1 + 1.48x_2 - 0.53)) + 5.52(\sigma(-1.48x_1 - 1.49x_2 - 0.53)) - 3.27$" where σ denotes a sigmoidal activation function, learnt from the sample $\{(1,1,0),(1,0,1),(0,1,0),(0,0,0)\}$ through the usual backpropagation algorithm [14].*
2. *In classifying emotions in a phonetic database, a C4.5[10] decison tree consisting of 64 IF-THEN-ELSE rules on 74 features is definitely beaten by a Support Vector Machine [16] with linear kernel on the same variables [4].*

4 Conclusions

We moved from a first model based on a random bit generator where the sole knowledge we can extract is the frequency of ones in the next bits, to a

second model where the random source is coupled with a computing machinery and we observe properties about the latter, up to the last model where the random source disappears in favor of a prefix machinery coupled with the unfeasible job of computing the shortest descriptions. We realize the inference goal is to compute properties which could be ascertained in terms of occurrence frequencies in the future observations of a phenomenon, and conclude that these properties have nothing to do with the mysterious operation of tossing a coin; rather, they merely represent a correct synthesis of what we have already observed or know about the phenomenon. This approcah, whose ultimate randomness source lies in some uncomputable strings, ϕ's in the last section, allows us to deepen some crucial inference task at both symbolic (regression curves) and subsymbolic (neural networks) levels. Moreover, a further release of the minimal structural risk minimization principle sheds some light on the designing of hybrid subsymbolic-symbolic learning paradigms.

References

[1] E. Aarts and J. Korst. *Simulated annealing and Boltzmann machines : a stochastic approach to combinatorial optimization and neural computing.* Wiley-Interscience series in discrete mathematics and optimization. John Wiley, Chichester, 1989.

[2] B. Apolloni, D. Malchiodi, C. Orovas, and G. Palmas. From synapses to rules. In *Foundations of Connectionist-symbolic Integration: Representation, Paradigms, and Algorithms - Proceedings of the 14th European Conference on Artificial Intelligence*, 2000.

[3] A. Church. *Introduction to Mathematical Logic I*, volume 13 of *Annals of Mathematics Studies.* Princeton University Press, Princeton, NJ, 1944.

[4] W. A. Fellenz, G. J. Taylor, R. Cowie, E. Douglas-Cowie, F. Piat, S. Kollias, C. Orovas, and B. Apolloni. On emotion recognition of faces and speech using neural networks, fuzzy logic and the assess system. In S. Amari, C. Lee Giles, M. Gori, and V. Piuri, editors, *Proceeding of the IEEE-INNS-ENNS International Joint Conference on Neural Networks - IJCNN 2000*, pages II–93,II–98, Los Alamitos, 2000. IEEE Computer Society.

[5] M. Hilario. An overview of strategies for neurosymbolic integration. In F. Alexandre, editor, *Connectionist-Symbolic Processing: From Unified to Hybrid Approaches.* Lawrence Erlbaum, 1998.

[6] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer, Berlin, 1993.

[7] D. Malchiodi. *Algorithmic approach to the statistical inference of non-Boolean function classes.* Ph.D. dissertation, Università degli studi di Milano, 2000.

[8] S. Martello and P. Toth. The 0–1 knapsack problem. In *Combinatorial Optimization*, pages 237–279. Wiley, 1979.

[9] E. Marubini and M. Valsecchi. *Analysing Survival Data from Clinical Trials and Observational Studies.* John Wiley & Sons, Chichester, UK, 19995.

[10] J. Quinlan. Comparing connectionist and symbolic learning methods. In *Computational Learning Theory and Natural Learning Systems. Volume I. Constraints and Prospects*, pages 445–456. MIT Press, Cambridge, 1994.

[11] H. Roger. *Theory of recoursive functions and effective computability.* Mc Graw-Hill, 1967.

[12] V. K. Rohatgi. *An Introduction to Probability Theory and Mathematical Statistics.* Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, 1976.

[13] S. Ross. *Simulation.* Statistical Modeling and Decision Science. Academic press, San Diego, second edition, 1997.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing, Vol. 1.* MIT Press, Cambridge, Massachusstes, 1986.

[15] R. Sun. *Integrating rules and connectionism for robust commonsense reasoning.* Wiley, New York, 1994.

[16] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[17] S. S. Wilks. *Mathematical Statistics.* Wiley Publications in Statistics. John Wiley. New York, 1965.

[18] S. Zacks. *The Theory of Statistical Inference.* Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, 1971.