Negative Log–Likelihood And Statistical Hypothesis Testing As The Basis Of Model Selection In IDEAs

(Full (Technical Report) Version)

Peter A.N. Bosman
peterb@cs.uu.nl

Dirk Thierens
Dirk.Thierens@cs.uu.nl

Department of Computer Science, Utrecht University
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands

August 2000

Abstract

In this paper, we analyze the most prominent features in model selection criteria that have been used so far in iterated density estimation evolutionary algorithms (IDEAs, EDAs, PMBGAs). These algorithms build probabilistic models and estimate probability densities based upon a selection of available points. We show that the negative log–likelihood is a basis of the inference features when the Kullback–Leibler divergence is used. We show how previously found to be problematic issues in the case of continuous random variables can be resolved by starting from the derived basics. By doing so, we have a probabilistic model search metric that can be justified through the use of statistical hypothesis tests. This in turn reduces the need for additional complexity penalties.

1 Introduction

The past few years, new probabilistic optimization algorithms inspired by evolutionary algorithms have appeared [17]. The algorithms that have been introduced in this field, apply search criteria to the space of probability distributions to find a suitable probabilistic model, given a vector of selected samples. The resulting probability distribution is subsequently used to draw more samples from, after which selection takes place again. Even though efficient algorithms have already been proposed in this field, the motivation of the search criteria for finding a suitable probabilistic model is often not thoroughly investigated. Our goal in this paper is to take a closer look at one such search criterion which is based on the Kullback–Leibler (KL) divergence. Different algorithms [2, 3, 4, 6, 7, 14] have been proposed that use this divergence. We want to emphasize the background of the KL divergence through its correspondence with likelihood maximization in probability theory. By doing so, we give a unifying background regarding these topics. Using the basic notion of likelihood, we aim to derive general probabilistic model selection criteria that in addition to previous approaches can be justified through the use of statistical hypothesis testing.

The remainder of this paper is organized as follows. First, we introduce some notation in section 2 and give a general framework for the algorithms themselves in section 3. Next, in section 4, we formalize the statistical tools that have been used in these algorithms and show how they together constitute some of the methods of induction used so far. In this same section, we unify these statistical tools by looking at their relation with a basic notion of fit in probability theory, being the likelihood. By taking a closer look at the negative log–likelihood, we come across a few subtle details. We go into these details and show how some of the previous algorithms can be simplified by statistical hypothesis testing on the basis of the negative log–likelihood. As a result, some problems that have been encountered in the expansion of discrete to continuous optimization algorithms [7] disappear. In section 5, we give some practical examples of the theoretic work described in this paper. Subsequently, in section 6, we summarize our results and discuss some topics of interest. Finally, we conclude this paper in section 7.

2 Some notation and statistics

We write $\boldsymbol{a} = (\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{|\boldsymbol{a}|-1})$ for a vector $\boldsymbol{a}$ of length $|\boldsymbol{a}|$. The ordering of the elements in a vector is *relevant*. We assume to have l random variables available, meaning that each sample point is an l dimensional vector. We introduce the notation $a\langle\boldsymbol{c}\rangle = (a_{\boldsymbol{c}_0}, a_{\boldsymbol{c}_1}, \ldots, a_{\boldsymbol{c}_{|\boldsymbol{c}|-1}})$. Let $\boldsymbol{\mathcal{L}} = (0, 1, \ldots, l-1)$ be a vector of l numbers and let $\boldsymbol{a} \sqsubseteq \boldsymbol{\mathcal{L}}$, meaning that $\boldsymbol{a}$ contains only elements of $\boldsymbol{\mathcal{L}}$. We assume the alphabet for each discrete random variable to be the same, just as we do for the range of each continuous random variable. We denote the alphabet for the discrete random variables by $\mathcal{A}$. We can now define the multivariate joint probability mass function (pmf):

$$p_{\boldsymbol{a}}(x\langle\boldsymbol{a}\rangle) = P(\forall_{i\in\boldsymbol{a}}\langle X_i = x_i\rangle) \quad \text{such that} \sum_{\boldsymbol{c}\in\mathcal{A}^{|\boldsymbol{a}|}} p_{\boldsymbol{a}}(\boldsymbol{c}) = 1, \text{ and } p_{\boldsymbol{a}}(\cdot) \geq 0 \tag{1}$$

We write $P(X\langle\boldsymbol{a}\rangle)(x\langle\boldsymbol{a}\rangle)$ for $p_{\boldsymbol{a}}(x\langle\boldsymbol{a}\rangle)$, making $P(X\langle\boldsymbol{a}\rangle)$ a pmf. In the continuous case, we cannot use a pmf. Let $dy\langle\boldsymbol{a}\rangle = \prod_{i=0}^{|\boldsymbol{a}|-1} dy_{\boldsymbol{a}_i}$ be shorthand notation for the multivariate derivative. Using the notation shorthand $\int$ for $\int_{-\infty}^{\infty}$, the multivariate joint probability density function (pdf) is the following:

$$\int_{b_0}^{c_0}\int_{b_1}^{c_1}\ldots\int_{b_{|\boldsymbol{a}|-1}}^{c_{|\boldsymbol{a}|-1}} f_{\boldsymbol{a}}(y\langle\boldsymbol{a}\rangle)dy\langle\boldsymbol{a}\rangle = P(Y\langle\boldsymbol{a}\rangle \in A) \tag{2}$$

$$\text{such that} \int\int\ldots\int f_{\boldsymbol{a}}(y\langle\boldsymbol{a}\rangle)dy\langle\boldsymbol{a}\rangle = 1, \quad f_{\boldsymbol{a}}(\cdot) \geq 0, \text{ and } A = \left(\prod_{i=0}^{|\boldsymbol{a}|-1}[b_i, c_i]\right) \subseteq \mathbb{R}^{|\boldsymbol{a}|}$$

We write $P(Y\langle\boldsymbol{a}\rangle)(y\langle\boldsymbol{a}\rangle)$ for $f_{\boldsymbol{a}}(y\langle\boldsymbol{a}\rangle)$, making $P(Y\langle\boldsymbol{a}\rangle)$ a pdf. We use Y to denote continuous random variables and X to denote discrete random variables. If we do not distinguish between these cases, we write Z.

We let $\boldsymbol{b} \sqsubseteq \boldsymbol{\mathcal{L}}$ and define $\boldsymbol{a} \sqcup \boldsymbol{b}$ to be the splicing of the two vectors so that the elements of $\boldsymbol{b}$ are placed behind the elements of $\boldsymbol{a}$, giving $|\boldsymbol{a} \sqcup \boldsymbol{b}| = |\boldsymbol{a}| + |\boldsymbol{b}|$. Using the definition of multivariate probability, we can define conditional probability by:

$$P(Z\langle\boldsymbol{a}\rangle|Z\langle\boldsymbol{b}\rangle) = \frac{P(Z\langle\boldsymbol{a}\sqcup\boldsymbol{b}\rangle)}{P(Z\langle\boldsymbol{b}\rangle)} \tag{3}$$

Shannon [20] has defined the multivariate entropy measure. In the case of discrete random variables, this measure is defined as:

$$H(P(X\langle\boldsymbol{a}\rangle)) = -\sum_{\boldsymbol{c}\in\mathcal{A}^{|\boldsymbol{a}|}} P(X\langle\boldsymbol{a}\rangle)(\boldsymbol{c})\ln(P(X\langle\boldsymbol{a}\rangle)(\boldsymbol{c})) \tag{4}$$

In the continuous case we use the term *differential entropy*, which is:

$$h(P(Y\langle\boldsymbol{a}\rangle)) = -\int\int\ldots\int P(Y\langle\boldsymbol{a}\rangle)(y\langle\boldsymbol{a}\rangle)\ln(P(Y\langle\boldsymbol{a}\rangle)(y\langle\boldsymbol{a}\rangle))dy\langle\boldsymbol{a}\rangle \tag{5}$$

Let $\boldsymbol{\mathcal{X}} = X\langle\boldsymbol{\mathcal{L}}\rangle$, $\boldsymbol{\mathcal{Y}} = Y\langle\boldsymbol{\mathcal{L}}\rangle$ and $\boldsymbol{\mathcal{Z}} = Z\langle\boldsymbol{\mathcal{L}}\rangle$. A well known distance metric from one probability distribution $P_0(\boldsymbol{\mathcal{Z}})$ to another probability distribution $P_1(\boldsymbol{\mathcal{Z}})$, is the Kullback–Leibler (KL) divergence. The KL divergence is also called relative entropy [13]. In the case of discrete random variables, this distance metric equals:

$$D(P_0(\boldsymbol{\mathcal{X}})||P_1(\boldsymbol{\mathcal{X}})) = \sum_{\boldsymbol{c}\in\mathcal{A}^l} P_0(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln\left(\frac{P_0(\boldsymbol{\mathcal{X}})(\boldsymbol{c})}{P_1(\boldsymbol{\mathcal{X}})(\boldsymbol{c})}\right) \tag{6}$$

In the continuous case, the KL metric is equal to:

$$d(P_0(\boldsymbol{\mathcal{Y}})||P_1(\boldsymbol{\mathcal{Y}})) = \int\int\ldots\int P_0(\boldsymbol{\mathcal{Y}})(y\langle\boldsymbol{\mathcal{L}}\rangle)\ln\left(\frac{P_0(\boldsymbol{\mathcal{Y}})(y\langle\boldsymbol{\mathcal{L}}\rangle)}{P_1(\boldsymbol{\mathcal{Y}})(y\langle\boldsymbol{\mathcal{L}}\rangle)}\right)dy\langle\boldsymbol{a}\rangle \tag{7}$$

As a final statistical tool, we make use of the sample vector $\boldsymbol{\mathcal{S}}$. We let $\boldsymbol{\mathcal{S}} = (\boldsymbol{z^0}, \boldsymbol{z^1}, \ldots, \boldsymbol{z^{|\mathcal{S}|-1}})$ be the vector of sample points, so that $\boldsymbol{z^i} = (z_0^i, z_1^i, \ldots, z_{l-1}^i) = z^i\langle\boldsymbol{\mathcal{L}}\rangle$. The sample vector $\boldsymbol{\mathcal{S}}$ is taken to be a vector of independently drawn samples from a distribution $P(\boldsymbol{\mathcal{Z}})$. We denote this by $\boldsymbol{\mathcal{S}} \xleftarrow{s} P(\boldsymbol{\mathcal{Z}})$. The sample vector is used to find some probability distribution $\hat{P}(\boldsymbol{\mathcal{Z}})$ as an approximation to the true distribution $P(\boldsymbol{\mathcal{Z}})$. The likelihood that the samples were drawn from some given distribution $\hat{P}(\boldsymbol{\mathcal{Z}})$, is defined as:

$$\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{Z}})) = \prod_{i=0}^{|\boldsymbol{\mathcal{S}}|-1} \hat{P}(\boldsymbol{\mathcal{Z}})(\boldsymbol{z^i}) \tag{8}$$

It is common practise to use the negative logarithm of the likelihood measure as it is often computationally more convenient. The resulting expression is called the *negative log–likelihood*:

$$-\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{Z}}))) = -\sum_{i=0}^{|\boldsymbol{\mathcal{S}}|-1} \ln(\hat{P}(\boldsymbol{\mathcal{Z}})(\boldsymbol{z^i})) \tag{9}$$

3 Evolutionary optimization by iterated density estimation

Evolutionary optimization algorithms that make use of iterated density estimation, attempt to catch the probability distribution of a selected vector of samples. Using the estimated probability distribution, more samples are generated and mixed with the existing samples to get a new sample vector. Assume that we have an l dimensional cost function $C(z\langle\boldsymbol{\mathcal{L}}\rangle)$, which without loss of generality we seek to minimize. We let $P^\theta(\boldsymbol{\mathcal{Z}})$ be a probability distribution that is uniform over all vectors $z\langle\boldsymbol{\mathcal{L}}\rangle$ with $C(z\langle\boldsymbol{\mathcal{L}}\rangle) \leq \theta$ and 0 otherwise. Sampling from $P^\theta(\boldsymbol{\mathcal{Z}})$ gives more samples that evaluate to a value below θ. Moreover, if we know $\theta^* = \min_{z\langle\boldsymbol{\mathcal{L}}\rangle}\{C(z\langle\boldsymbol{\mathcal{L}}\rangle)\}$, sampling from $P^{\theta^*}(\boldsymbol{\mathcal{Z}})$ gives an optimal solution. This rationale has led to the definition of the $\mathbb{IDEA}$ (Iterated Density Estimation Evolutionary Algorithm) framework [4]. The greatest difference between other similar approaches and the $\mathbb{IDEA}$, is that the $\mathbb{IDEA}$ has mostly been used to focus on continuous optimization problems [4, 6, 7].

We cannot expect to efficiently solve every structured optimization problem with just any probabilistic model. This has been demonstrated on problems in which a multiple of variables interact [5, 18]. Therefore, a higher order probability distribution such as $\hat{P}(Z_0, Z_1)$ instead of $\hat{P}(Z_0)\hat{P}(Z_1)$ is sometimes required. However, the higher the order of complexity, the larger the amount of required computational resources are in order to be able to compute the probabilistic model. Therefore, the interactions between the problem variables are attempted to be *inferred* from the given sample vector to find a probabilistic model $\mathcal{M}$. The notion of a probabilistic model is used as a computational implementation of a probability distribution. A probabilistic model $\mathcal{M}$ uniquely corresponds to a probability distribution. It can be seen to consist of some structure ς and a vector of parameters $\boldsymbol{\theta}$ that defines the pdfs to be fit over each joint multivariate structure implied by ς. An example of a structure ς is the notion of a *factorization*, which we denote by $\mathfrak{f}$. A factorization *factors* the probability distribution over $\boldsymbol{\mathcal{Y}}$ to get a product of pdfs. An example in the case of $l = 3$ is $P(\boldsymbol{\mathcal{Y}}) = P(Y_0, Y_1)P(Y_2)$. Once a structure ς is given, the parameters that have to be estimated can be derived from the multivariate pdfs that have to be fit. Since the way in which the parameters $\boldsymbol{\theta}$ are fit, is predefined on beforehand together with the pdfs to fit, we denote the parameter vector that is obtained in this manner by $\boldsymbol{\theta} \xleftarrow{fit} \varsigma$. This implies that whereas we formally define a probabilistic model to be $\mathcal{M} = (\varsigma, \boldsymbol{\theta})$, in our practical case, we can write $\mathcal{M} = (\varsigma, \boldsymbol{\theta} \xleftarrow{fit} \varsigma)$. As a probability distribution can therefore be identified using only the structure ς, we denote it by $P_\varsigma(\boldsymbol{\mathcal{Y}})$. The definition of the $\mathbb{IDEA}$ framework can now be given as follows:

$\mathbb{IDEA}(n, \tau, m, sel(), rep(), ter(), sea(), est(), sam())$	
Initialize an empty vector of samples	$\boldsymbol{\mathcal{P}} \leftarrow ()$
Add and evaluate n random samples	***for*** $i \leftarrow 0$ ***to*** $n-1$ ***do***
	$\boldsymbol{\mathcal{P}} \leftarrow \boldsymbol{\mathcal{P}} \sqcup \text{NewRandomVector}()$
	$c[\boldsymbol{\mathcal{P}}_i] \leftarrow C(\boldsymbol{\mathcal{P}}_i)$
Initialize the iteration counter	$t \leftarrow 0$
Iterate until termination	***while*** $\neg ter()$ ***do***
Select $\lfloor \tau n \rfloor$ samples	$(z^0\langle\boldsymbol{\mathcal{L}}\rangle, z^1\langle\boldsymbol{\mathcal{L}}\rangle, \ldots, z^{\lfloor \tau n \rfloor -1}\langle\boldsymbol{\mathcal{L}}\rangle) \leftarrow sel()$
Set θ_t to the worst selected cost	$\theta_t \leftarrow c[z^k\langle\boldsymbol{\mathcal{L}}\rangle]$ such that
	$\forall_{i \in \boldsymbol{\mathcal{N}}_{\boldsymbol{\tau}}} \langle c[z^i\langle\boldsymbol{\mathcal{L}}\rangle] \leq c[z^k\langle\boldsymbol{\mathcal{L}}\rangle]\rangle$
Search for a structure ς	$\varsigma \leftarrow sea()$
Estimate the parameters $\boldsymbol{\theta} \xleftarrow{fit} \varsigma$	$\boldsymbol{\theta} \leftarrow est()$
Create an empty vector of new samples	$\boldsymbol{\mathcal{O}} \leftarrow ()$
Sample m new samples from $\hat{P}_\varsigma(\boldsymbol{\mathcal{Z}})$	***for*** $i \leftarrow 0$ ***to*** $m-1$ ***do***
	$\boldsymbol{\mathcal{O}} \leftarrow \boldsymbol{\mathcal{O}} \sqcup sam()$
Replace a part of $\boldsymbol{\mathcal{P}}$ with a part of $\boldsymbol{\mathcal{O}}$	$rep()$
Evaluate the new samples in $\boldsymbol{\mathcal{P}}$	***for each*** unevaluated $\boldsymbol{\mathcal{P}}_i$ ***do***
	$c[\boldsymbol{\mathcal{P}}_i] \leftarrow C(\boldsymbol{\mathcal{P}}_i)$
Update the generation counter	$t \leftarrow t+1$
Denote the required iterations by t_{end}	$t_{\text{end}} \leftarrow t$

In the $\mathbb{IDEA}$ framework, we have that $\boldsymbol{\mathcal{N}}_{\boldsymbol{\tau}} = (0, 1, \ldots, \lfloor \tau n \rfloor - 1)$, $\tau \in [\frac{1}{n}, 1]$, $sel()$ is the selection operator, $rep()$ replaces a subset of $\boldsymbol{\mathcal{P}}$ with a subset of $\boldsymbol{\mathcal{O}}$, $ter()$ is the termination condition, $sea()$ is a model structure search algorithm, $est()$ estimates the parameters that are implied by both the structure ς as well as predefined pdfs and $sam()$ generates a single sample using the estimated densities. The evolutionary algorithm characteristic of the $\mathbb{IDEA}$ lies in the fact that a population of individuals is used from which individuals are selected to generate new offspring with. Using these offspring along with the parent individuals and the current population, a new population is constructed.

Once ς has been selected, the actual pdfs are computed. During the course of deriving ς however, the required pdfs will often have already been computed [4]. In whatever way, using equation 3, computing the pdfs can functionally be restricted to computing multivariate joint pdfs. In the discrete case, there has been only one way in which these functions are computed, which is the most straightforward way. The probabilities equal the frequency in which some combination of values appears in the sample vector[1], divided by the total amount of samples:

$$\hat{P}(X\langle\boldsymbol{a}\rangle)(c\langle\boldsymbol{a}\rangle) = \frac{1}{|\boldsymbol{\mathcal{S}}|}\sum_{j=0}^{|\boldsymbol{\mathcal{S}}|-1}\begin{cases} 1 \text{ if } \forall_{i \in \boldsymbol{a}}\langle c_i = x_i^j\rangle \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

In the case of continuous random variables, such a most straightforward way to compute the probabilities does not exist. In general, some model is fit to the data as good as possible by specifying its parameters. For instance, a multivariate normal pdf can be used, which is fully specified by its covariance matrix and its mean vector. This discrepancy between discrete and continuous random variables is important in certain model selection techniques as we shall see.

4 Negative log–likelihood as the basis of statistical model selection

The algorithms in the $\mathbb{IDEA}$ field that have been proposed so far, can roughly be divided into three categories. One category consists of the methods that use a univariate distribution in which none of the variables interact with other variables. We make no statement of any kind on these algorithms as they perform no induction on the model since ς is fixed. The methods in another category use

[1]Note that in the $\mathbb{IDEA}$ framework we have that the samples we are referring to are the result of the selection operation $\boldsymbol{\mathcal{S}} \leftarrow sel()$.

a higher order structure and for instance use the KL divergence with the full joint probability distribution to guide the search. For factorizations, it has been noted before [6] that using the KL divergence should be accompanied by strong restrictions on $\mathfrak{f}$ because otherwise minimizing the divergence will give just the full joint probability distribution. In case of a factorization, the heavy restrictions are usually embodied by a parameter κ which denotes the amount of variables any one variable is allowed to interact with. To do away with κ, the methods in a third category have been invented. These methods add a penalty term to the search metric in order to penalize more complex models. The amount of penelization is however usually parameterized again.

4.1 Negative log–likelihood and sample entropy

To start out, we first note that there is a correspondence between equations 4 and 9 as well as between equations 5 and 9. We first focus on the discrete case, which is the most simple.

We split up $\boldsymbol{\mathcal{S}}$ into $|\mathcal{A}|^l$ *mutually disjoint* subvectors so that each subvector contains only one type of truncated sample point. If we define $\boldsymbol{a} \sqcap \boldsymbol{b}$ to be the order dependent intersection of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that each element of $\boldsymbol{a}$ is preserved if it occurs in $\boldsymbol{b}$, we can rewrite equation 9 by summing over all possible subvectors of $\boldsymbol{\mathcal{S}}$ as they together are *exactly* equal to $\boldsymbol{\mathcal{S}}$:

$$-\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{X}}))) = -\sum_{\boldsymbol{c}\in\mathcal{A}^l} \sum_{i=0}^{|\boldsymbol{\mathcal{S}}\sqcap\boldsymbol{c}|-1} \ln(\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) = -\sum_{\boldsymbol{c}\in\mathcal{A}^l} |\boldsymbol{\mathcal{S}} \sqcap \boldsymbol{c}| \ln(\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) \tag{11}$$

Note that $|\boldsymbol{\mathcal{S}} \sqcap \boldsymbol{c}|$ is just the amount of times that vector $\boldsymbol{c}$ occurs in the sample vector. Using equation 10, this means that we have $|\boldsymbol{\mathcal{S}} \sqcap \boldsymbol{c}| = |\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})$, so we may conclude:

$$-\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{X}}))) = -\sum_{\boldsymbol{c}\in\mathcal{A}^l} |\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) = |\boldsymbol{\mathcal{S}}|H(\hat{P}(\boldsymbol{\mathcal{X}})) \tag{12}$$

In the discrete case we thus have that the negative log–likelihood equals $|\boldsymbol{\mathcal{S}}|$ times the multivariate entropy of the estimated model. The reason for this is that the estimated model in the discrete case fits over the sample points with a maximum likelihood. The entropy measures are exact in the sense that they go over the probabilities of *every* domain element. Alternatively, we can define *sample* entropies that compute the measure given a vector of samples $|\boldsymbol{\mathcal{S}}'| \overset{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{Z}})$. Note that it is *essential* that the samples were sampled from the probability distribution for which we are computing the entropy. As the sample entropy is the same in the discrete as well as the continuous case, we have only a single definition for it:

$$\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}(\boldsymbol{\mathcal{Z}})) = \frac{-1}{|\boldsymbol{\mathcal{S}}'|} \sum_{i=0}^{|\boldsymbol{\mathcal{S}}'|-1} \ln(\hat{P}(\boldsymbol{\mathcal{Z}})(\boldsymbol{\mathcal{S}}'_i)) \;\text{ such that } \boldsymbol{\mathcal{S}}' \overset{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{Z}}) \tag{13}$$

Note that the definition of the sample entropy is closely related to that of the negative log–likelihood. To be exact, $-\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}'|\hat{P}(\boldsymbol{\mathcal{Z}}))) = |\boldsymbol{\mathcal{S}}'|\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}(\boldsymbol{\mathcal{Z}}))$. The difference is that the definition of sample entropy is only valid if $\boldsymbol{\mathcal{S}}' \overset{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{Z}})$, which we do not require in the case of the negative log–likelihood.

In the derivation of equation 12 from 11, we have used that $|\boldsymbol{\mathcal{S}} \sqcap \boldsymbol{c}| = |\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})$. But, $\boldsymbol{\mathcal{S}} \overset{s}{\leftarrow} P(\boldsymbol{\mathcal{X}})$ and *not* $\boldsymbol{\mathcal{S}} \overset{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{X}})$. Therefore it should be noted that *only* in the limit of $|\boldsymbol{\mathcal{S}}'| \to \infty$ we have that $\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}(\boldsymbol{\mathcal{X}})) \to H(\hat{P}(\boldsymbol{\mathcal{X}}))$, $|\boldsymbol{\mathcal{S}}'| \overset{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{X}})$. Furthermore, in the limit of $|\boldsymbol{\mathcal{S}}| \to \infty$, we have that $|\boldsymbol{\mathcal{S}} \sqcap \boldsymbol{c}| = |\boldsymbol{\mathcal{S}}|P(\boldsymbol{\mathcal{X}})(\boldsymbol{c})$, since then $\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c}) \to P(\boldsymbol{\mathcal{X}})(\boldsymbol{c})$ and thus in the additional limit of $|\boldsymbol{\mathcal{S}}'| \to \infty, |\boldsymbol{\mathcal{S}}'| \overset{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{X}})$, we have that $\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}(\boldsymbol{\mathcal{X}})) \to H(P(\boldsymbol{\mathcal{X}}))$. We therefore have:

$$\lim_{|\boldsymbol{\mathcal{S}}| \to \infty} -\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{X}}))) = -|\boldsymbol{\mathcal{S}}|\sum_{\boldsymbol{c}\in\mathcal{A}^{|\boldsymbol{a}|}} P(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) = \tag{14}$$

$$-|\boldsymbol{\mathcal{S}}|\sum_{\boldsymbol{c}\in\mathcal{A}^{|\boldsymbol{a}|}} P(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(P(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) = -|\boldsymbol{\mathcal{S}}|H(P(\boldsymbol{\mathcal{X}}))$$

In the continuous case, things are more complex. We discretize each variable so that it consists of equidistant intervals of length m and let $M = \langle \ldots, [-2m, -m\rangle, [-m, 0\rangle, [0, m\rangle, [m, 2m\rangle, \ldots \rangle$. Note that $(\boldsymbol{\mathcal{S}} \sqcap A)_i$ is the i–th sample from the sample vector that falls into area A. Recalling that $\boldsymbol{\mathcal{S}} \stackrel{s}{\leftarrow} P(\boldsymbol{\mathcal{Y}})$, equation 9 becomes:

$$-\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{Y}}))) = -\sum_{A \in M^l} \sum_{i=0}^{|\boldsymbol{\mathcal{S}} \sqcap A|-1} \ln(\hat{P}(\boldsymbol{\mathcal{Y}})((\boldsymbol{\mathcal{S}} \sqcap A)_i)) \tag{15}$$

Now in the limit of $|\boldsymbol{\mathcal{S}}| \to \infty$, we have that $|\boldsymbol{\mathcal{S}} \sqcap A| \to |\boldsymbol{\mathcal{S}}|P(\boldsymbol{\mathcal{Y}} \in A)$. We may thus write:

$$\lim_{|\boldsymbol{\mathcal{S}}| \to \infty} - \ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{Y}}))) = -\sum_{A \in M^l} \sum_{i=0}^{|\boldsymbol{\mathcal{S}}|P(\boldsymbol{\mathcal{Y}} \in A)-1} \ln(\hat{P}(\boldsymbol{\mathcal{Y}})((\boldsymbol{\mathcal{S}} \sqcap A)_i)) \tag{16}$$

If we now let $m \to 0$, the area A becomes infinitely small and is therefore shrunk to a single point. The discrete sum over all areas then becomes an integral over all points:

$$\lim_{|\boldsymbol{\mathcal{S}}| \to \infty} - \ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{Y}}))) = -\int\!\!\int\!\ldots\!\int \sum_{i=0}^{|\boldsymbol{\mathcal{S}}|P(\boldsymbol{\mathcal{Y}})(y\langle \boldsymbol{\mathcal{L}} \rangle)-1} \ln(\hat{P}(\boldsymbol{\mathcal{Y}})(y\langle \boldsymbol{\mathcal{L}} \rangle))dy\langle \boldsymbol{\mathcal{L}} \rangle = \tag{17}$$

$$-|\boldsymbol{\mathcal{S}}|\int\!\!\int\!\ldots\!\int P(\boldsymbol{\mathcal{Y}})(y\langle \boldsymbol{\mathcal{L}} \rangle)\ln(\hat{P}(\boldsymbol{\mathcal{Y}})(y\langle \boldsymbol{\mathcal{L}} \rangle))dy\langle \boldsymbol{\mathcal{L}} \rangle$$

From equation 17 it follows that in the limit of $|\boldsymbol{\mathcal{S}}'| \to \infty$, $\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}(\boldsymbol{\mathcal{Y}})) \to h(\hat{P}(\boldsymbol{\mathcal{Y}}))$ *if and only if* $\boldsymbol{\mathcal{S}}' \stackrel{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{Y}})$. Note that we cannot enforce the same further relation as we derived for the discrete case. The relation that we have in the discrete case that in the limits of $|\boldsymbol{\mathcal{S}}| \to \infty$, $|\boldsymbol{\mathcal{S}}'| \to \infty$, the sample entropy over $\hat{P}(\boldsymbol{\mathcal{X}})$ becomes equal to the entropy over the actual underlying probability distribution $P(\boldsymbol{\mathcal{X}})$, does not hold in general in the continuous case. Actually, it does not hold in general in the discrete case either, but we have assumed that in the discrete case we always approximate probabilities by using definition 10. As the definition in equation 10 becomes equal to $P(\boldsymbol{\mathcal{X}})$ in the limit of $|\boldsymbol{\mathcal{S}}| \to \infty$, this is clearly true. However, in the continuous case, we hardly ever have that in the limit of $|\boldsymbol{\mathcal{S}}| \to \infty$ we get $\hat{P}(\boldsymbol{\mathcal{Y}}) \to P(\boldsymbol{\mathcal{Y}})$. So we can neither conclude that the entropy of the true distribution becomes equal to the entropy of the estimated distribution. In order for that to hold in general, we require a special pdf:

$$\int_{a_0}^{b_0}\int_{a_1}^{b_1}\ldots\int_{a_{l-1}}^{b_{l-1}} \hat{P}(\boldsymbol{\mathcal{Y}})(y\langle \boldsymbol{\mathcal{L}} \rangle)dy\langle \boldsymbol{\mathcal{L}} \rangle = \frac{\boldsymbol{\mathcal{S}} \sqcap \prod_{i=0}^{l-1}[a_i, b_i]}{|\boldsymbol{\mathcal{S}}|} \tag{18}$$

Clearly, the pdf so constructed is too specific with respect to the sample vector and is therefore said to *overfit* the data. This is a very important aspect of density estimation and is termed *generalization*. As in a practical application we shall use a more generalizing model for $\hat{P}(\boldsymbol{\mathcal{Y}})$, we cannot enforce that in the limits of $|\boldsymbol{\mathcal{S}}| \to \infty$, $|\boldsymbol{\mathcal{S}}'| \to \infty$, $\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}(\boldsymbol{\mathcal{Y}})) \to h(P(\boldsymbol{\mathcal{Y}}))$, $\boldsymbol{\mathcal{S}}' \stackrel{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{Y}})$.

There is however one final important issue. In the discrete case, we found that the entropy of the estimated distribution equals $|\boldsymbol{\mathcal{S}}|$ times the negative log–likelihood over the sample vector (equation 12). We have however not established this in the continuous case. In our derivation, we have directly derived that given an infinite amount of samples, the sample entropy becomes equal to the actual differential entropy over the estimated model. As a matter of fact, we do not in general have the same relation between the entropy and the negative log–likelihood in the continuous case as we do in the discrete case. The reason why it does hold in the discrete case, is because equation 10 is a maximum likelihood estimate of the sample vector. As we have not assumed that our fit in the continuous case is a maximum likelihood fit unless we use equation 18, we have not enforced this relation in the continuous case. However, Kullback [13] has shown that *if and only if* $\hat{P}(\boldsymbol{\mathcal{Y}})$ is a maximum likelihood estimate over $\boldsymbol{\mathcal{S}}$, the equality holds both in the discrete case as well as the continuous case. Let $\boldsymbol{\vartheta}$ be the vector of parameters for $\hat{P}(\boldsymbol{\mathcal{Z}})$ and $\boldsymbol{\Theta}$ be the vector space of all possible values for the parameters $\boldsymbol{\vartheta}$, we can then formalize this by:

$$\hat{P}(\boldsymbol{\mathcal{Z}}) = arg\ \max\{\mathfrak{L}(\boldsymbol{\mathcal{S}}|\mathcal{P}) \mid \mathcal{P} \in \{\hat{P}(\boldsymbol{\mathcal{Z}}|\boldsymbol{\vartheta}) \mid \boldsymbol{\vartheta} \in \boldsymbol{\Theta}\}\} \Longleftrightarrow -\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}(\boldsymbol{\mathcal{Z}})) = |\boldsymbol{\mathcal{S}}|h(\hat{P}(\boldsymbol{\mathcal{Z}})) \quad (19)$$

4.2 KL divergence and negative log–likelihood model selection

In the methods in the second mentioned category, model selection is performed subject to certain constraints. Let $\mathfrak{C}_\varsigma$ be the set of all model structures that satisfy the constraints, $\varsigma \in \mathfrak{C}_\varsigma$. The constraints on the model have varied over the past from none [1, 19] to those of chain dependencies [3], tree dependencies [2], conditional dependencies with a maximum of κ conditionals per variable [4, 6, 7, 14, 15, 16] and marginal product models [11]. Prior to the I$\mathbb{DE}$A framework, all of the higher order models were regarded in the case of discrete (binary) random variables.

The basic idea is then to minimize the distance between the true probability distribution and the estimated probability distribution. As a distance measure, the KL divergence between two probability distributions can be used in which one probability distribution is the most complex probability distribution $P(\boldsymbol{\mathcal{Z}})$ and the other is the estimate $\hat{P}_\varsigma(\boldsymbol{\mathcal{Z}})$.

In the case of factorizations, an algorithm to approximately find the best factorization can start from the complete univariate factorization in which each variable is taken independently and incrementally build a more complex factorization. Each step in the incremental algorithm is based upon the change that brings about the largest decrease in the KL divergence to $\hat{P}(\boldsymbol{\mathcal{Z}})$. In the case of a model structure in general, this means that if we let $\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{Z}})$ be the current model, we test it against a (more complex) model $\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{Z}})$ by observing the change in the KL divergence to the most complex probability distribution. The model that brings about the greatest decrease is selected as the new current model. The metric that has been used in the discrete case, has only been applied to factorizations and is computed with respect to the full joint factorization:

$$\begin{aligned} &D(\hat{P}(\boldsymbol{\mathcal{X}})||\hat{P}_{\mathfrak{f}^0}(\boldsymbol{\mathcal{X}})) - D(\hat{P}(\boldsymbol{\mathcal{X}})||\hat{P}_{\mathfrak{f}^1}(\boldsymbol{\mathcal{X}})) = \\ &\sum_{\boldsymbol{c}\in\mathcal{A}^l}\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(\hat{P}_{\mathfrak{f}^1}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) - \sum_{\boldsymbol{c}\in\mathcal{A}^l}\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(\hat{P}_{\mathfrak{f}^0}(\boldsymbol{\mathcal{X}})(\boldsymbol{c}))\end{aligned} \quad (20)$$

We note that by using conditional probabilities, we can specify any factorization of the probability distribution using multivariate joint pdfs as its elemental building blocks. With equation 3, we therefore have that both sums in equation 20 are actually sums over sums over certain vectors $\boldsymbol{a}$ of $\hat{P}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(\hat{P}(X\langle\boldsymbol{a}\rangle)(c\langle\boldsymbol{a}\rangle))$ with $\boldsymbol{c} \in \mathcal{A}^l$. As given some vector $\boldsymbol{k} \sqsupseteq \boldsymbol{a}$, the theorem of total probability gives us that $\sum_{\boldsymbol{c}\in\mathcal{A}^{|\boldsymbol{k}|}}P(X\langle\boldsymbol{k}\rangle)(\boldsymbol{c}) = \sum_{\boldsymbol{c}\in\mathcal{A}^{|\boldsymbol{a}|}}P(X\langle\boldsymbol{a}\rangle)(\boldsymbol{c})$, equation 20 transforms into:

$$D(\hat{P}(\boldsymbol{\mathcal{X}})||\hat{P}_{\mathfrak{f}^0}(\boldsymbol{\mathcal{X}})) - D(\hat{P}(\boldsymbol{\mathcal{X}})||\hat{P}_{\mathfrak{f}^1}(\boldsymbol{\mathcal{X}})) = H(\hat{P}_{\mathfrak{f}^0}(\boldsymbol{\mathcal{X}})) - H(\hat{P}_{\mathfrak{f}^1}(\boldsymbol{\mathcal{X}})) \quad (21)$$

In the continuous case, we can use similar arguments to show that we have:

$$d(\hat{P}(\boldsymbol{\mathcal{Y}})||\hat{P}_{\mathfrak{f}^0}(\boldsymbol{\mathcal{Y}})) - d(\hat{P}(\boldsymbol{\mathcal{Y}})||\hat{P}_{\mathfrak{f}^1}(\boldsymbol{\mathcal{Y}})) = h(\hat{P}_{\mathfrak{f}^0}(\boldsymbol{\mathcal{Y}})) - h(\hat{P}_{\mathfrak{f}^1}(\boldsymbol{\mathcal{Y}})) \quad (22)$$

The resulting expression in equation 22 has been used so far in continuous I$\mathbb{DE}$As [4, 6, 7] as a logical expansion over the discrete case. This requires however the computation of a multivariate integral over a given pdf. For the normal pdf [4], this expression can be derived analytically. However, in the case of the normal kernels pdf [7], this is no longer possible. The same is the case for the use of a normal mixture pdf. The latter two pdfs are however very useful in the I$\mathbb{DE}$A approach. In order to use them anyhow, the integrals may be computed numerically. There are many ways to do this and mostly this requires an exponential amount of time in the dimensionality of the integral. However, we have seen in this paper that we can use an approximation. If we can efficiently sample points from the estimation $\hat{P}_{\mathfrak{f}}(\boldsymbol{\mathcal{Y}})$, equation 17 tells us that we may use $\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}_{\mathfrak{f}}(\boldsymbol{\mathcal{Y}}))$ with $\boldsymbol{\mathcal{S}}' \overset{s}{\leftarrow} \hat{P}_{\mathfrak{f}}(\boldsymbol{\mathcal{Y}})$, as an approximation to the true entropy over the estimated model. This resolves the problems regarding the computation of the entropy in the search metric as encountered so far in continuous I$\mathbb{DE}$As [6, 7]. Summarizing, for finding $\mathfrak{f}$, the methods using the KL divergence have used:

$$\begin{cases} \min_{\mathfrak{f}\in\mathfrak{C}_\mathfrak{f}}\{H(\hat{P}_\mathfrak{f}(\boldsymbol{\mathcal{X}}))\} & \text{(Discrete case)} \\ \min_{\mathfrak{f}\in\mathfrak{C}_\mathfrak{f}}\{h(\hat{P}_\mathfrak{f}(\boldsymbol{\mathcal{Y}}))\} & \text{(Continuous case)} \end{cases} \tag{23}$$

As an approximation, the sample entropy can be used:

$$\min_{\mathfrak{f}\in\mathfrak{C}_\mathfrak{f}}\{\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}_\mathfrak{f}(\boldsymbol{\mathcal{Z}}))\} \tag{24}$$

Note that even though equation 23 consists of exact definitions, the estimated distributions are based upon samples. Therefore, both optimization problems are inherently sample based. The divergence that we have minimized so far in the discrete and continuous case however, is a divergence between some probability distribution and the most complex probability distribution that we have *estimated*. This only works if the most complex distribution of our model can be fit to the sample vector to be a good estimator. It is however far from likely that this will always be the case. The problem lies not in the use of the KL divergence, but in the use of the most complex probability distribution to which we are computing the KL divergence. What we *should* be using as a distance metric is the distance between our estimated model and the *actual* probability distribution $P(\boldsymbol{\mathcal{Z}})$. The reason for this is that we want our estimated model to resemble the true distribution as good as possible. By doing so, using the KL divergence becomes a maximum likelihood optimization of our estimated model, which is exactly what we want. At this point, we no longer solely have to regard factorizations as the model structure. In the discrete case, we thus want to look at the following KL divergence difference:

$$\begin{gathered} D(P(\boldsymbol{\mathcal{X}})||\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{X}})) - D(P(\boldsymbol{\mathcal{X}})||\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{X}})) = \\ \sum_{\boldsymbol{c}\in\mathcal{A}^l} P(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) - \sum_{\boldsymbol{c}\in\mathcal{A}^l} P(\boldsymbol{\mathcal{X}})(\boldsymbol{c})\ln(\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{X}})(\boldsymbol{c})) \end{gathered} \tag{25}$$

We now face sums over a probability distribution $P(\boldsymbol{\mathcal{X}})$ that we do not know. However, equation 14 tells us that we can use the negative log–likelihood as a sample estimator to the expressions in the true KL divergence difference to obtain:

$$D(P(\boldsymbol{\mathcal{X}})||\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{X}})) - D(P(\boldsymbol{\mathcal{X}})||\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{X}})) \approx \frac{1}{|\boldsymbol{\mathcal{S}}|}(\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{X}}))) - \ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{X}})))) \tag{26}$$

In the continuous case, the true KL difference is:

$$\begin{gathered} d(P(\boldsymbol{\mathcal{Y}})||\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{Y}})) - d(P(\boldsymbol{\mathcal{Y}})||\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{Y}})) = \\ \int\!\!\int\!\ldots\!\int P(\boldsymbol{\mathcal{Y}})(y\langle\boldsymbol{\mathcal{L}}\rangle)\ln(\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{Y}})(y\langle\boldsymbol{\mathcal{L}}\rangle))dy\langle\boldsymbol{\mathcal{L}}\rangle - \\ \int\!\!\int\!\ldots\!\int P(\boldsymbol{\mathcal{Y}})(y\langle\boldsymbol{\mathcal{L}}\rangle)\ln(\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{Y}})(y\langle\boldsymbol{\mathcal{L}}\rangle))dy\langle\boldsymbol{\mathcal{L}}\rangle \end{gathered} \tag{27}$$

We again face explicit multivariate integrals as we implicitly did in equation 22 because of the differential entropies. The additional problem with equation 27 is that, just as we had in the discrete case, we have to integrate over a probability distribution $P(\boldsymbol{\mathcal{Y}})$ that we do not know. However, using equation 17, we get an approximation that is similar to the one in the discrete case based on the negative log–likelihood:

$$d(P(\boldsymbol{\mathcal{Y}})||\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{Y}})) - d(P(\boldsymbol{\mathcal{Y}})||\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{Y}})) \approx \frac{1}{|\boldsymbol{\mathcal{S}}|}(\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{Y}}))) - \ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{Y}})))) \tag{28}$$

We have thus simplified the approaches so far in continuous IDEAs [4, 6, 7]. Even though using equation 24, we already no longer had the need to explicitly evaluate the integrals, we still had to be able to draw samples from the distribution. By using equation 28 however, we can directly use the given sample points and evaluate the estimated pdf at those points. The use of the KL divergence leads to the testing of the appropriateness of two different models through the use of the negative log–likelihood. Consequently, the search for good model structures is guided by the maximum likelihood of the sample vector, given some probabilistic model. The best fitting model is taken to be the one with the minimum negative log–likelihood. From now on, we call this procedure *minimum log–likelihood model selection*. In effect, the approaches using this selection attempt to solve:

$$\min_{\varsigma \in \mathfrak{C}_\varsigma}\{-\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}_\varsigma(\boldsymbol{\mathcal{Z}})))\} \tag{29}$$

At this point, we have three different optimization problems as defined in equations 23, 24 and 29. We have already argued that in the limit of $|\boldsymbol{\mathcal{S}}'| \to \infty, |\boldsymbol{\mathcal{S}}'| \overset{s}{\leftarrow} \hat{P}(\boldsymbol{\mathcal{Z}})$, optimization problems 23 and 24 are identical. The previous approaches using the KL divergence have all used the optimization problem in equation 23. However, because of equation 19, the newly proposed optimization problem in equation 29 is identical to the so far used optimization problem under the condition that the estimated model is a maximum likelihood estimate. So in effect, the methods so far have *also* been attempting to maximize the likelihood of the probabilistic model with respect to $|\boldsymbol{\mathcal{S}}|$ but only because maximum likelihood pmf or pdf estimates were used.

4.3 Justifying sample based model selection

We have defined minimum log–likelihood model selection as discriminating between probabilistic models on the basis of the log–likelihood value in equation 9. This value is based on samples just as is the sample entropy. It has gone unmentioned that this is also the case for using the exact entropy over the estimation as defined in equation 23, since the estimation itself is based on samples. Because of the fact that samples are involved, a random variable can be identified for the resulting value that we are using.

If we have some random variable R along with N values $r_0, r_1, \ldots, r_{N-1}$ for it, its sample mean $\overline{R}$ and its unbiased sample standard deviation $\tilde{s}_R$ are:

$$\overline{R} = \frac{1}{N}\sum_{i=0}^{N-1} r_i, \qquad \tilde{s}_R = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(r_i - \overline{R})^2} \tag{30}$$

Because of the central limit theorem, the sample mean is approximately normally distributed. Now observe the following statistic:

$$T = \frac{\sqrt{N}(\overline{R} - \mathfrak{h})}{\tilde{s}_R} \tag{31}$$

It can be shown (see for instance [12]) that the T statistic is distributed according to Student's T distribution. Given some value for $\overline{R}$, when the first $N-1$ values are set, the value for r_{N-1} cannot be chosen freely anymore and must be fixed if the definitions in equation 30 are to be obeyed. Therefore, we say that the T statistic has $\delta = N - 1$ degrees of freedom. It can be shown (see for instance [12]) that the T statistic has the following pdf with δ degrees of freedom:

$$f_{\mathcal{T}}(\delta, y) = \frac{\Gamma(\frac{\delta+1}{2})}{\Gamma(\frac{\delta}{2})\sqrt{\delta\pi}}\left(1 + \frac{y^2}{\delta}\right)^{-\frac{\delta+1}{2}} \tag{32}$$

In the definition of $f_{\mathcal{T}}$ we have used Euler's Gamma function $\Gamma(y)$:

$$\Gamma(y) = \int_0^\infty x^{y-1}e^{-x}\ dx, \quad y > 0 \tag{33}$$

To evaluate $\Gamma(y)$ efficiently, we can use a result by Feller [10]:

$$\Gamma(y) \approx y^{y-\frac{1}{2}}e^{-y}\sqrt{2\pi}\left(1 + \frac{1}{12y}\right) \tag{34}$$

We can now state that given some hypothesized value $\mathfrak{h}$ for the actual mean μ_R of R, the value of T can be computed and it can be tested at the significance level α whether or not the hypothesized value is a valid assumption or not. Since the T distribution is symmetric, this can be done by finding the value $t_{1-\frac{\alpha}{2}}$ such that $\int_{t_{1-\frac{\alpha}{2}}}^{\infty} f_T(N-1,y)dy = \frac{\alpha}{2}$. The hypothesis is then rejected if $|T| > t_{1-\frac{\alpha}{2}}$ and accepted otherwise. If we now assume to have two random variables R_0 and R_1, for which we both have N samples, we observe:

$$\mathbb{T} = \frac{\sqrt{N}(\overline{R_0} - \overline{R_1} - \mathfrak{h})}{\sqrt{\tilde{s}^2_{R_0} + \tilde{s}^2_{R_1}}} \tag{35}$$

It can be shown that the $\mathbb{T}$ statistic (see for instance [12]) is also distributed according to Student's T distribution with $\delta = 2(N-1)$ degrees of freedom, where $\mathfrak{h}$ is now an hypothesis for the value of the difference between μ_{R_0} and μ_{R_1}.

If we let R_i stand for the negative log–likelihood of the samples under model $\hat{P}_{\varsigma^i}(\boldsymbol{\mathcal{Z}})$, $i \in \{0,1\}$, we can use the $\mathbb{T}$ statistic to test whether the average value of the negative log–likelihood of $\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{Z}})$ is significantly smaller than $\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{Z}})$. If this is so, $\hat{P}_{\varsigma^1}(\boldsymbol{\mathcal{Z}})$ is selected in favor of $\hat{P}_{\varsigma^0}(\boldsymbol{\mathcal{Z}})$. In order to do this, we set the hypothesized value $\mathfrak{h}$ to 0 and perform a right sided test. This means that we are testing whether the expected value of $R_0 - R_1$ is greater than 0 and thus whether $\overline{R_0} > \overline{R_1}$. The amount of degrees of freedom in our case is $\delta = 2(|\boldsymbol{\mathcal{S}}|-1) - |\boldsymbol{\theta}|$. For each parameter $\boldsymbol{\theta}_i$ that is to be estimated, we lose one degree of freedom. The fact that we use the *average* value, means that we are minimizing $1/|\boldsymbol{\mathcal{S}}|$ times the negative log–likelihood, but as this factor is the same over all models, the structure of the minimization problem remains unaltered.

5 Examples and experiments

In this section, we give some examples in both the discrete as well as the continuous case that bring the theoretical derivations into practice. Furthermore, we test the results on real optimization functions using continuous ID$\mathbb{E}$As.

5.1 Examples

In all of our examples, we regard only two random variables Z_0 and Z_1 and factorizations $\mathfrak{f}$ for the model structure ς. We attempt to infer whether it is more beneficial to use $\hat{P}_{\mathfrak{f}^1} = \hat{P}(Z_0, Z_1)$ instead of $\hat{P}_{\mathfrak{f}^0} = \hat{P}(Z_0)\hat{P}(Z_1)$. We start out with discrete random variables and restrict ourselves to the elementary case of binary random variables with alphabet $\mathcal{A} = \{0,1\}$. Lets assume that our underlying pmf $P(Z_0, Z_1)$ is completely random, meaning that $\forall_{(c,d)\in\mathcal{A}^2}\langle P(X_0,X_1)(c,d) = \frac{1}{4}\rangle$. Now lets assume that we have 41 samples. Instead of drawing them randomly from the true distribution, we let the first 40 samples be 10 groups of 4 samples of all possible combinations for X_0 and X_1. Over the first 40 samples, the empirical probabilities computed according to equation 10 are thus equal to the true distribution. Whatever we let the 41–st sample be, it will distort the approximation. We choose to let the 41–st sample to be $(0,0)$. We can now compute the discrete approximation from equation 10, the entropy of the estimated pmf by using equation 4, the sample entropy of the estimated pmf by using equation 13 and drawing 100 random samples from the approximated pmf, and the negative log–likelihood of the estimated pmf by using equation 9. The results for the two different models are the following[2]:

$\mathfrak{f}$	$\hat{P}_{\mathfrak{f}}$	$H(\hat{P}_{\mathfrak{f}})$	$\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}_{\mathfrak{f}})$	$-\frac{1}{\lvert\boldsymbol{\mathcal{S}}\rvert}\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}} \mid \hat{P}_{\mathfrak{f}}))$
$\mathfrak{f}^0$	$\frac{441}{1681}, \frac{420}{1681}, \frac{420}{1681}, \frac{400}{1681}$	1.385699	1.387865	1.385699
$\mathfrak{f}^1$	$\frac{11}{41}, \frac{10}{41}, \frac{10}{41}, \frac{10}{41}$	1.385416	1.388113	1.385416

[2]The probabilities in the table are for the respective combinations $(0,0), (0,1), (1,0), (1,1)$.

As expected from our derivations in equations 12 and 19, the negative log–likelihood is equal to the entropy over the discrete estimated pmf in equation 10 and the sample entropy is close this value. Based on the KL divergence, the approach so far has been to use equation 23 and select the model with the minimal entropy value. In this case, even though the difference between the two models is only very slight, the absolute difference leads to select $P_{\mathfrak{f}^1}(X_0, X_1)$ over $P_{\mathfrak{f}^0}(X_0, X_1)$. If we now compute the statistics as proposed in section 4.3, we get:

$\overline{R_{\mathfrak{f}^0}}$	$\tilde{s}_{\mathfrak{f}^0}$	$\overline{R_{\mathfrak{f}^1}}$	$\tilde{s}_{\mathfrak{f}^1}$	$\mathbb{T}$
1.385699	0.035331	1.385416	0.042754	0.032726

At the significance level of $\alpha = 5\%$, the critical value of Student's T distribution is 1.668 at $\delta = 2(41-1)-2-3 = 75$ degrees of freedom. This implies that if we start with the simple model $P_{\mathfrak{f}^0}(X_0, X_1)$ and want to decide whether $P_{\mathfrak{f}^1}(X_0, X_1)$ is a better choice, using our new approach, we would reject this decision. Note that this is of course what we want, since the source is just a univariate probability distribution that does not need to be described by higher order complex models.

As a second example, we take the pmf that equals 0 over combinations $(0,1)$ and $(1,0)$ and equals $\frac{1}{2}$ over combinations $(0,0)$ and $(1,1)$. As a vector of 41 samples, we take 20 occurrences of $(0,0)$ and $(1,1)$ each and let the 41–st sample be $(1,0)$ as some very slight noise on the input. Clearly, we cannot describe this distribution satisfactorily with the product model $P_{\mathfrak{f}^0}(X_0, X_1)$. In this case, we want our approach to agree with the previous approaches and select model $P_{\mathfrak{f}^1}(X_0, X_1)$ instead at the significance level of $\alpha = 5\%$. We therefore first compute the entropy, sample entropy and negative log–likelihood:

$\mathfrak{f}$	$\hat{P}_{\mathfrak{f}}$	$H(\hat{P}_{\mathfrak{f}})$	$\mathfrak{H}(\boldsymbol{S}', \hat{P}_{\mathfrak{f}})$	$-\frac{1}{\lvert\boldsymbol{S}\rvert}\ln(\mathfrak{L}(\boldsymbol{S}\lvert\hat{P}_{\mathfrak{f}}))$
$\mathfrak{f}^0$	$\frac{420}{1681}, \frac{400}{1681}, \frac{441}{1681}, \frac{400}{1681}$	1.385699	1.384450	1.385699
$\mathfrak{f}^1$	$\frac{20}{41}, 0, \frac{1}{41}, \frac{20}{41}$	0.790906	0.807712	0.790906

Again, we observe that using merely the entropy difference, we would prefer $P_{\mathfrak{f}^1}(X_0, X_1)$ over $P_{\mathfrak{f}^0}(X_0, X_1)$. The required statistics are the following:

$\overline{R_{\mathfrak{f}^0}}$	$\tilde{s}_{\mathfrak{f}^0}$	$\overline{R_{\mathfrak{f}^1}}$	$\tilde{s}_{\mathfrak{f}^1}$	$\mathbb{T}$
1.385699	0.007620	0.790906	0.467858	8.139338

This time, the $\mathbb{T}$ statistic is larger than the critical value of 1.668. This implies that using our approach too, the joint model would be preferred.

Moving to the case of continuous random variables, we take the domain of the variables to be $[-1, 1]$. Again, we work with 41 samples. Using two different cases, we want to show that for continuous random variables, we get similar results. The first sample vector is a set of 40 points from a uniform distribution such that the points are placed on a grid of 8×5. The 41–st sample distorts the uniform characteristic of the sample vector slightly and is placed at $(-0.9, 0.9)$. The second sample vector is a joint sample vector and consists of points that are equidistantly placed along two lines that are parallel to $Y_1 = Y_0$. The 41–st sample lies in the center and equals $(0,0)$. To keep computations simple, we use a parametric approximation to the sample vector. Our parametric function is the normal pdf. A maximum likelihood fit for the normal pdf over a sample vector given random variables $Y\langle\boldsymbol{a}\rangle$, can be found by computing the sample mean and sample covariances as described in equation 36 below. The resulting estimations are shown in figure 1.

$$f(y\langle\boldsymbol{a}\rangle) = \frac{(2\pi)^{-\frac{|\boldsymbol{a}|}{2}}}{(\det \boldsymbol{S}(\boldsymbol{a}))^{\frac{1}{2}}} e^{-\frac{1}{2}(y\langle\boldsymbol{a}\rangle - \overline{Y}\langle\boldsymbol{a}\rangle)^T(\boldsymbol{S}(\boldsymbol{a}))^{-1}(y\langle\boldsymbol{a}\rangle - \overline{Y}\langle\boldsymbol{a}\rangle)} \tag{36}$$

$$\text{where} \quad \overline{Y}\langle\boldsymbol{a}\rangle = \frac{1}{|\boldsymbol{S}|}\sum_{i=0}^{|\boldsymbol{S}|-1} y^i\langle\boldsymbol{a}\rangle, \quad \boldsymbol{S}(\boldsymbol{a}) = \frac{1}{|\boldsymbol{S}|}\sum_{i=0}^{|\boldsymbol{S}|-1}(y^i\langle\boldsymbol{a}\rangle - \overline{Y}\langle\boldsymbol{a}\rangle)(y^i\langle\boldsymbol{a}\rangle - \overline{Y}\langle\boldsymbol{a}\rangle)^T$$

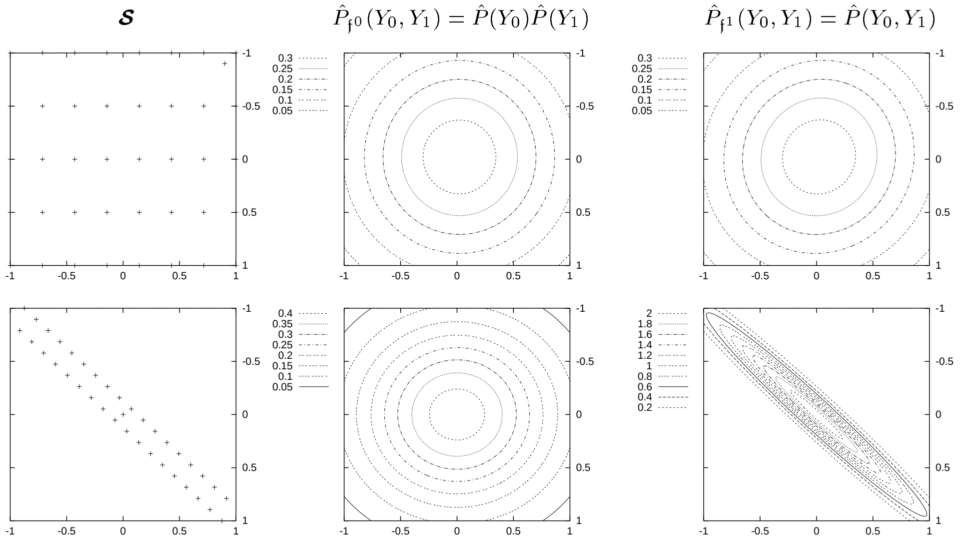

Figure 1: Density contours of maximum likelihood normal pdf estimates using a univariate product model (second column) and a multivariate joint model (third column). The results are shown over two different sample vectors (first column).

The entropy of a joint multivariate normal pdf can be shown [8] to be:

$$h(y\langle \boldsymbol{a}\rangle) = \frac{1}{2}(|\boldsymbol{a}| + \ln((2\pi)^{|\boldsymbol{a}|}(\det\ \boldsymbol{S}(\boldsymbol{a})))) \tag{37}$$

Using this information, we can compute the actual entropy of the estimated distribution as well as the sample entropy over 100 samples and the negative log–likelihood over the sample vector. The results are the following:

Univariate sample vector				
$\mathfrak{f}$	$\hat{P}_{\mathfrak{f}}$ descriptors	$h(\hat{P}_{\mathfrak{f}})$	$\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}_{\mathfrak{f}})$	$-\frac{1}{\lvert\boldsymbol{\mathcal{S}}\rvert}\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}\lvert\hat{P}_{\mathfrak{f}}))$
$\mathfrak{f}^0$	$\overline{Y_0} = -0.021951, \boldsymbol{S}((0)) = [0.507079]$ $\overline{Y_1} = 0.021951, \boldsymbol{S}((1)) = [0.437393]$	2.084871	2.086227	2.084871
$\mathfrak{f}^1$	$(\overline{Y_0}, \overline{Y_1}) = (-0.021951, 0.021951)$ $\boldsymbol{S}((0,1)) = \begin{bmatrix} 0.507079 & -0.019274 \\ -0.019274 & 0.437393 \end{bmatrix}$	2.084033	2.071505	2.084033

Joint sample vector				
$\mathfrak{f}$	$\hat{P}_{\mathfrak{f}}$ descriptors	$h(\hat{P}_{\mathfrak{f}})$	$\mathfrak{H}(\boldsymbol{\mathcal{S}}', \hat{P}_{\mathfrak{f}})$	$-\frac{1}{\lvert\boldsymbol{\mathcal{S}}\rvert}\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}\lvert\hat{P}_{\mathfrak{f}}))$
$\mathfrak{f}^0$	$\overline{Y_0} = 0.000000, \boldsymbol{S}((0)) = [0.359435]$ $\overline{Y_1} = 0.000000, \boldsymbol{S}((1)) = [0.374679]$	1.835424	1.781525	1.835424
$\mathfrak{f}^1$	$(\overline{Y_0}, \overline{Y_1}) = (0.000000, 0.000000)$ $\boldsymbol{S}((0,1)) = \begin{bmatrix} 0.359435 & 0.359435 \\ 0.359435 & 0.374679 \end{bmatrix}$	0.234478	0.215413	0.234478

Because of equation 19 and the fact that we have used a maximum likelihood estimation, we find that $-\ln(\mathfrak{L}(\boldsymbol{\mathcal{S}}|\hat{P}_{\mathfrak{f}}(\boldsymbol{\mathcal{Y}}))) = |\boldsymbol{\mathcal{S}}|h(\hat{P}_{\mathfrak{f}}(\boldsymbol{\mathcal{Y}}))$, just as we did in the discrete case. In both cases we would prefer the two dimensional joint probability distribution if we decide solely on the entropy over $\hat{P}_{\mathfrak{f}^i}(Y_0, Y_1)$. However, we again see that this absolute difference is very small in the uniform sample vector. To see if this difference is significant, we compute the required statistics for Student's T difference test and find:

HT	$\overline{C_0}$	n	RT	$\overline{C_1}$	n	RT	$\overline{C_2}$	n	RT
No	9999999.96	350	130.17	7.50	275	44.32	27.73	550	4.95
Yes	9999999.86	250	1251.56	6.32	225	1299.03	18.75	350	244.71

Figure 2: Results of the experiments over C_0, C_1 and C_2.

Univariate sample vector				
$\overline{R_{\mathfrak{f}0}}$	$\tilde{s}_{\mathfrak{f}0}$	$\overline{R_{\mathfrak{f}1}}$	$\tilde{s}_{\mathfrak{f}1}$	$\mathbb{T}$
2.084871	0.604633	2.084033	0.605166	0.006274

Joint sample vector				
$\overline{R_{\mathfrak{f}0}}$	$\tilde{s}_{\mathfrak{f}0}$	$\overline{R_{\mathfrak{f}1}}$	$\tilde{s}_{\mathfrak{f}1}$	$\mathbb{T}$
1.835424	0.931245	0.234478	0.483911	9.767834

At the significance level of $\alpha = 5\%$ and the accompanying critical value of Student's T distribution of 1.669 at $2(41-1)-4-5 = 71$ degrees of freedom, the joint probabilistic model is only selected in the case of the joint sample vector as expected. In the continuous case, we can thus also conclude that our approach results in desirable behavior of the probabilistic model selection process.

5.2 Experiments

In addition to the examples in the previous section, we give the results of some experiments to demonstrate the use of the statistical testing in practice. We have used the following continuous function maximization problems:

C_0	$\gamma_i = \frac{24}{1000}(i+2) - y_i$	$y\langle\mathcal{L}\rangle \in [-3,3]^l$
C_1	$\gamma_0 = y_0, \gamma_i = y_i + \gamma_{i-1}$	$y\langle\mathcal{L}\rangle \in [-3,3]^l$
C_2	$\gamma_0 = y_0, \gamma_i = y_i + \sin(\gamma_{i-1})$	$y\langle\mathcal{L}\rangle \in [-3,3]^l$

$$C_i = \frac{100}{10^{-5} + \sum_{i=0}^{l-1} |\gamma_i|}$$

To compare our experiments with prior results [6] where no statistical hypothesis testing was used, we use the same settings. We have $l = 100$, a maximum of 200000 function evaluations and we average the results over 20 runs. The pdf we use is the normal pdf and we use the special case conditional graph search algorithm based on Edmonds algorithm [9] for optimum branchings. The arcs that were available to Edmonds algorithm were those that resulted in a $\mathbb{T}$ value over the critical Student's T value at the $\alpha = 25\%$ significance level. Repeating earlier reported results [6], figure 2 shows that our new approaches obtain comparable results. The tables contain whether hypothesis testing was used, the average cost $\overline{C_i}$, the best value for n, and the relative time RT. Let n_e be the required amount of function evaluations, $\mathrm{FT}(x)$ the time to perform x random function evaluations and TT the total algorithm time including the n_e function evaluations. Then, $\mathrm{RT} = (\mathrm{TT} - \mathrm{FT}(n_e))/\mathrm{FT}(n_e)$.

Note that the approaches that result by using the statistical hypothesis tests are quite a lot slower. The reason for this is that for every arc that has to be justified, $\mathcal{O}(\tau n)$ additional involved computations have to be done, whereas in the case of using the normal pdf and the entropy value, this is $\mathcal{O}(1)$. Since there are $\mathcal{O}(l^2)$ such arcs, the running times are strongly influenced. Furthermore, the obtained results are not as good as the results that were obtained by the previous approach. The function that we tested has a very high order structure, namely the full joint probability distribution. Therefore, incorporating more dependencies will result in a better problem value. However, the normal pdf is quite unsensitive to the hypothesis test. Since the normal pdf is very general, the samples are almost always underfit. Furthermore, the population size is kept small in our example since only a fixed amount of evaluations is allowed. As a result, the structure of the problem is harder to find. Therefore, a lower significance value should be used so as to be less conservative towards possible dependencies.

There are two more important issues to note. First of all, if we only use the normal pdf for simplicity, the issue of inferring the structure of the problem is not as important as in the case of discrete random variables or the histogram pdf for continuous random variables. In the latter two cases, the amount of parameters that has to be determined grows exponentially with the amount of incorporated dependencies. For the full joint distribution, the amount of parameters in the case of binary random variables is $2^l - 1$. However, in the case of the normal pdf, the amount of parameters for the full joint distribution is $\frac{1}{2}l^2 + \frac{3}{2}l$. The I$\mathbb{DE}$A instance that results if the normal pdf and the full joint distribution are used, indeed runs in polynomial time [4]. It can therefore be argued that finding structure in the case of continuous random variables is less important. However, since the polynomial time is bounded only by $\mathcal{O}(l^4)$, it is still beneficial to use lower order structures if possible from a practical point of view.

Another issue to point out is that using only the normal pdf is too insensitive to be truly effective on a wide range of problems. Higher order interactions such as clusters that cannot be fit well by the normal pdf, will cause a higher order fit to be only slightly better than multiple lower order normal pdfs. Therefore, adding clustering methods to remove non–linear interactions will improve the effectiveness of the approaches proposed so far in the case of continuous random variables. Furthermore, the statistical hypothesis test as we introduced in this paper will then also contribute more to the search. An illustrative example is given in figure 3. A symmetric sample vector shows non–linear behaviour that cannot be fit well by a single multivariate normal pdf. The obtained result is almost identical to the product of two univariate normal pdfs. However, if two clusters are identified, the symmetry over the horizontal axis is removed and two multivariate normal pdfs can be fit over these individual sample vectors. The result is a well fitting distribution as can be seen in figure 3. Note that the negative log–likelihood hypothesis test over each cluster separately would have resulted in the displayed fit because of the strong correlation that is observed within each cluster separately.

6 Summary and discussion

We have shown that previous approaches using the KL divergence have attempted to find a factorization $\mathfrak{f}$ by minimizing the entropy of the estimated distribution $\hat{P}_{\mathfrak{f}}(\boldsymbol{\mathcal{Z}})$ (equation 23).

In the continuous case, problems have been encountered to compute the differential entropy because of the integrals. We have shown that as an approximation, a model may be found from minimizing the sample entropy of the estimated distribution $\hat{P}_{\varsigma}(\boldsymbol{\mathcal{Z}})$ over $\boldsymbol{\mathcal{S}}' \overset{s}{\leftarrow} \hat{P}_{\varsigma}(\boldsymbol{\mathcal{Z}})$ (equation 24).

We have shown that in the limit of $|\boldsymbol{\mathcal{S}}'| \rightarrow \infty$, the two optimization problems are identical. We have also shown that both optimization problems are guided by a metric that is a distance from some probabilistic model to the *estimated* most complex probability model. This is however not the desired distance to the *actual* most complex probability model. Alternatively, we have shown that the best descriptive model can be found from minimizing the negative log–likelihood of $\boldsymbol{\mathcal{S}} \overset{s}{\leftarrow} P(\boldsymbol{\mathcal{Z}})$ given the estimated distribution $\hat{P}_{\varsigma}(\boldsymbol{\mathcal{Z}})$ (equation 29).

All approaches using the KL divergence so far, have essentially been attempting to minimize the entropy over the estimated distribution that has been fit using maximum likelihood estimate pmfs or pdfs. From equation 19, we have that these approaches have been doing the same as minimizing the negative log–likelihood over the estimated distribution.

What has gone unmentioned in all previous work however, is that all three optimization problems are based upon samples. In the first optimization problem this is implicitly the case whereas in the other two optimization problems, this is explicitly so. Deciding between probabilistic models during optimization should therefore be justified by using a statistical hypothesis test. We have shown that Student's T difference test on the average negative log–likelihood over the given samples $\boldsymbol{\mathcal{S}}$ is the required justification approach.

We might now argue that there is no more need for additional constraints on the probabilistic model. The use of the KL divergence in the first optimization problem and no constraints leads to a search algorithm that in a greedy fashion tries to incorporate as many dependencies as possible. Therefore, additional constraints are needed to find the best matching constrained model with

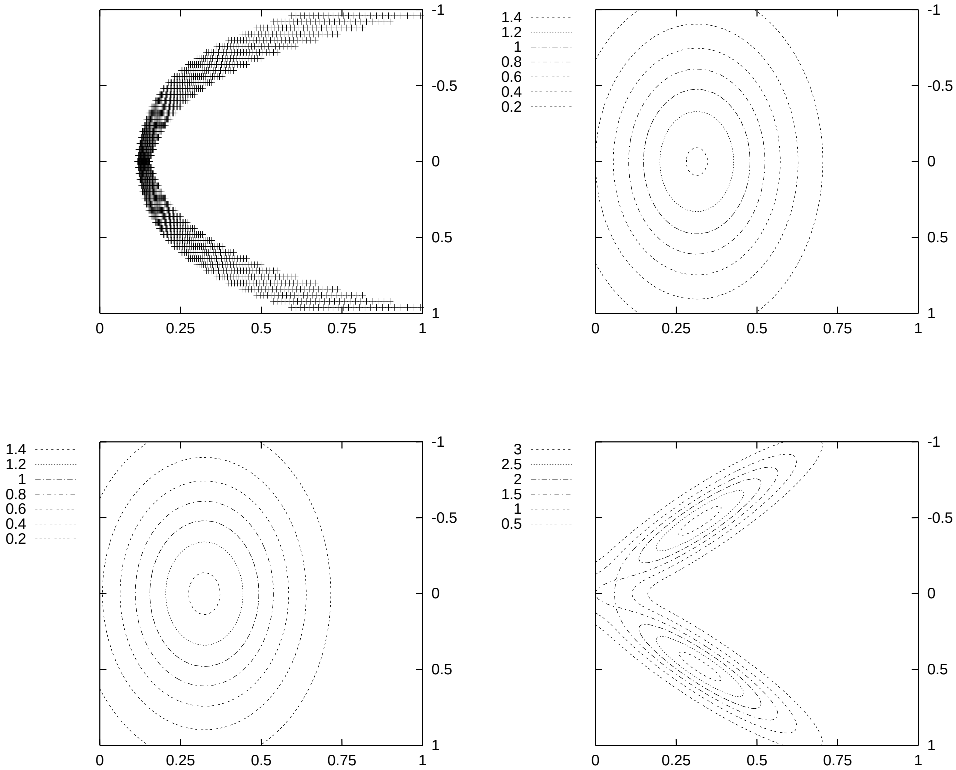

Figure 3: Density contours of maximum likelihood normal pdf estimates on a non–linear sample vector (top left). The density contours are shown for the product of two one dimensional normal pdfs (top right), a single two dimensional normal pdf (bottom left) and two normal pdfs in two dimensions that have been fit after the sample vector was clustered (bottom right).

respect to the most complex model. In the last optimization algorithm however, each higher order model has to be justified by being a better fit of the sample points. Hence, we can argue that we no longer require additional constraints as the dependencies that will be incorporated in the estimated model will have been justified. Still, penalizing more complex models can be useful from a computational point of view.

An important remark is that in the continuous case, using Student's T test in combination with only a normal pdf, is very sensitive to outliers. As such outliers are expected to occur frequently in optimization by iterated density estimation, either more complex models, such as normal mixture models, or means of clustering are required. By performing all of the steps in clusters of $\boldsymbol{\mathcal{S}}$, the tests will be more reliable if the amount of samples in each cluster does not become too small and the form of the clusters does not deviate too much from the possible density contours of a normal pdf. An advantage of this approach is that clusters can efficiently break up non–linear behavior in the original sample vector that cannot be fit well by a single normal pdf. Also, simple but effective clustering methods are computationally efficient as well.

Because of the results of this paper, we can now easily use an approach in which we additionally penalize more complex models. This leads to metrics such as the MDL [11] and the BIC [14] that have been used previously in iterated density estimation evolutionary algorithms. Using our results, the general idea is to use the negative log–likelihood as a measure of goodness of fit and use some function of the amount of parameters in the probabilistic model, which inherently determines its complexity, as an additional term. For instance, we might use a constant times an exponential in the amount of parameters of our distribution estimation. In the case of a multivariate normal pdf in N dimensions, this equals $\frac{1}{2}N^2 + \frac{3}{2}N$ for the unique parameters in the covariance matrix and the mean vector.

7 Conclusions

So far, algorithms based on the KL divergence have guided the search for a probabilistic model by computing the (differential) entropy over the estimated distribution. By attempting to minimize the negative log–likelihood over the estimated distribution instead, we do away with the problematic issue of computing the integrals in the differential entropy. However, under the assumption that the pmf or pdf estimates are of maximum likelihood, these minimization problems are identical. Furthermore, the metric that is computed using the negative log–likelihood, is a definition of how well a probabilistic model describes a certain vector of samples. By justifying the preference for one model over another on the basis of Student's T difference test, we have a robust scheme that infers the dependencies between the variables from the maximum likelihood. As a result, there is arguably no more need for additional constraints on the probabilistic model as the dependencies are attempted to be inferred through the use of justifiable statistics.

References

[1] S. Baluja and R. Caruana. Removing the genetics from the standard genetic algorithm. In A. Prieditis and S. Russell, editors, *Proceedings of the twelfth International Conference on Machine Learning*, pages 38–46. Morgan Kauffman publishers, 1995

[2] S. Baluja and S. Davies. Using optimal dependency–trees for combinatorial optimization: Learning the structure of the search space. In D.H. Fisher, editor, *Proc. of the 1997 Int. Conf. on Mach. Lear.* Morgan Kauffman publishers, 1997

[3] J.S. De Bonet, C. Isbell, and P. Viola. Mimic: Finding optima by estimating probability densities. *Advances in Neural Information Processing*, 9, 1996

[4] P.A.N. Bosman and D. Thierens. An algorithmic framework for density estimation based evolutionary algorithms. Utrecht University Technical Report UU–CS–1999–46. ftp://ftp.cs.uu.nl/pub/RUU/CS/techreps/CS-1999/1999-46.ps.gz, 1999

[5] P.A.N. Bosman and D. Thierens. Linkage information processing in distribution estimation algorithms. In W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, and R.E. Smith, editors, *Proceedings of the GECCO–1999 Genetic and Evolutionary Computation Conf.*, pages 60–67.

Morgan Kaufmann Publishers, 1999

[6] P.A.N. Bosman and D. Thierens. Continuous iterated density estimation evolutionary algorithms within the IDEA framework. In M. Pelikan, H. Mühlenbein, and A.O. Rodriquez, editors, *Proceedings of the Optimization by Building and Using Probabilistic Models OBUPM Workshop at the Genetic and Evolutionary Computation Conference GECCO–2000.* Morgan Kaufmann Publishers, 2000

[7] P.A.N. Bosman and D. Thierens. IDEAs based on the normal kernels probability density function. Utrecht University Technical Report UU–CS–2000–11. ftp://ftp.cs.uu.nl/pub/RUU/CS/techreps/CS-2000/2000-11.ps.gz, 2000

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* John Wiley & Sons Inc., 1991

[9] J. Edmonds. Optimum branchings. *J. Res. Nat. Bur. Standards*, 71B:233–240, 1967. Reprinted in Math. of the Decision Sciences, *Amer. Math. Soc. Lectures in Appl. Math.*, 11:335–345, 1968

[10] W. Feller. *An Introduction To Probability Theory And Its Applications, Volume 1.* Wiley, 1968

[11] G. Harik. Linkage learning via probabilistic modeling in the ECGA. IlliGAL Tech. Rep. 99010. ftp://ftp-illigal.ge.uiuc.edu/pub/papers/IlliGALs/ 99010.ps.Z, 1999

[12] M.G. Kendall and A. Stuart. *The Advanced Theory Of Statistics, Volume 2, Inference And Relationship.* Charles Griffin & Company Limited, 1967

[13] S. Kullback. *Information Theory And Statistics.* New York: Dover, 1968

[14] P. Larrañaga, R. Etxeberria, J.A. Lozano, and J.M. Peña. Optimization by learning and simulation of bayesian and gaussian networks. University of the Basque Country Technical Report EHU-KZAA-IK-4/99. http://www.sc.ehu.es/ccwbayes/postscript/kzaa-ik-04-99.ps, 1999.

[15] H. Mühlenbein and T. Mahnig. FDA – a scalable evolutionary algorithm for the optimization of additively decomposed functions. *Evol. Comp.*, 7:353–376, 1999

[16] M. Pelikan, D.E. Goldberg, and E. Cantú-Paz. BOA: The bayesian optimization algorithm. In W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, and R.E. Smith, editors, *Proc. of the GECCO–1999 Genetic and Evolutionary Computation Conf.*, pages 525–532. M.K. Publishers, 1999

[17] M. Pelikan, D.E. Goldberg, and F. Lobo. A survey of optimization by building and using probabilistic models. IlliGAL Technical Report 99018. ftp://ftp-illigal.ge.uiuc.edu/pub/papers/IlliGALs/99018.ps.Z, 1999

[18] M. Pelikan and H. Mühlenbein. The bivariate marginal distribution algorithm. In R. Roy, T. Furuhashi, K. Chawdry, and K. Pravir, editors, *Advances in Soft Computing – Engineering Design and Manufacturing.* Springer–Verlag, 1999

[19] M. Sebag and A. Ducoulombier. Extending population–based incremental learning to continuous search spaces. In A.E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature – PPSN V*, pages 418–427. Springer, 1998

[20] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948