Eur. Phys. J. B **22**, 129–135 (2001)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences

Experimental modeling of physical laws

I. Grabec[a]

Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva 6, PP 394, 1001 Ljubljana, Slovenia

Received 24 February 2000 and Received in final form 21 February 2001

Abstract. A physical law is represented by the probability distribution of a measured variable. The probability density is described by measured data using an estimator whose kernel is the instrument scattering function. The experimental information and data redundancy are defined in terms of information entropy. The model cost function, comprised of data redundancy and estimation error, is minimized by the creation-annihilation process.

PACS. 06.20.DK Measurement and error theory – 02.50.-r Probability theory, stochastic processes, and statistics – 89.70.+c Information science

1 Introduction

Quantitative physical explorations of natural phenomena involve three basic tasks: performing experiments, processing data, and modeling physical laws [1]. The leading trend in the development of modern experimental systems is to automatize the first two tasks, while the solution of the critical problem of modeling is still left to intuition. In the recent literature there already appear attempts to program as well the modeling for execution on a computer, especially for data acquisition systems in industrial environments [2]. Since measurements are always subject to random influences [1], a statistical approach to modeling is needed. Here we consider the probability distribution as a general basis for modeling of a physical law. The first step of the modeling is an estimation of probability density function (PDF) from experimental data. The most widely applicable is non-parametric estimation as it requires no *a priori* assumptions about PDF [2,3].

From the estimated PDF the experimental physical law can be extracted using the conditional average [2]. This average represents a non-parametric regression which can be carried out simultaneously with the data acquisition by computer. The structure of the corresponding information processing system resembles a structure of the radial basis function neural network [2,4,5]. In addition to non-parametric regression, several other paradigms from the fields of artificial neural networks, such as multilayer perceptrons, can be interpreted as automatic modelers of physical laws [2,4,6]. Various algorithms for adapting a selected model to experimental data have already been described [2,4,6], but the development of fundamental principles for a specification of the model structure is still a subject of current research [7]. The problems stem from a significant contrast between the complexity of experimental data and the structure of physical laws. The information about the phenomenon explored is generally increased with the number of experimental data; hence instrumental science and technology tend to develop electronic devices with ever greater storage capacity. Contrary to this, the most prominent property of a physical law is its simplicity [1]. At present it is still not clear how an electronic modeler could automatically and optimally compress the overwhelming experimental data into a simple law, although the theory of algorithmic information has already prepared some fundamentals for the treatment of this problem [8–10].

A simple model of physical law can be obtained by minimizing a cost function which is composed of model error and complexity [8]. The theory of statistics offers well elaborated methods for the estimation of the error [3,4,6], while the description of the model complexity is physically less well established [11,12]. For this purpose the measure of algorithmic complexity is applicable [7–9], but this measure is derived from the program code that determines the average model performance. In the physical literature the complexity is usually considered as an intrinsic property of the phenomenon and should therefore be expressed directly in terms of measured values [11]. With this aim we define in the next section the experimental information provided by measurements with an instrument of limited accuracy. It turns out that experimental information is useful for the description of the excessive complexity of data which can be utilized for the introduction of the model cost function.

In order to avoid problems with joining the error and complexity of the model in the cost function, it is convenient to express both terms by a single quantity [8]. For this purpose we employ the entropy of information [12,13], since it is non-dimensional and provides a common basis for formulation of error and complexity.

[a] e-mail: igor.grabec@fs.uni-lj.si

2 Experimental information and redundancy of data

At the definition of the experimental information we consider a scalar-valued variable X since the generalization to a multivariate case is straightforward. For this variable we select a bounded continuous sample space $S_X = (-L, L)$, where $2L$ is the span of the instrument applied. We assume that an arbitrary number, say N, of statistically independent measurements has yielded the samples $x_1, \dots, x_N$. The non-parametric estimator of PDF is then expressed by the sample average [2,3]

$$f(x) \;=\; \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n). \tag{1}$$

This estimator, though unbiased, is not consistent [2]. As Parzen has shown [3,14], it can be made consistent by using as a kernel a smooth approximation of the delta function, such as the Gaussian

$$\mathrm{g}(x - x_n, \sigma) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x - x_n)^2}{2\sigma}\right] \tag{2}$$

with some standard deviation σ dependent on N. Parzen's estimator

$$f(x) \;=\; \frac{1}{N} \sum_{n=1}^{N} \mathrm{g}(x - x_n, \sigma(N)) \tag{3}$$

is therefore biased [2], but the bias asymptotically vanishes if $\sigma(N)$ properly decreases towards 0 with increasing N [3,14]. The samples $x_1, \dots, x_N$ themselves are convenient parameters of the PDF model, but unfortunately their number must increase without limit and the smoothing parameter $\sigma(N)$ is introduced arbitrarily [4]. Since measurements are subject to instrumental scattering, the requirement that $\sigma(N)$ vanishes is in conflict with a correct physical presentation of measured quantities [2]. Consequently, we want to replace Parzen's method by a finite procedure, which would be more in tune with properties of experiments and would from the very beginning incorporate the measurement inaccuracy in the PDF estimator. For this purpose we turn first to the description of the instrumental scattering and interpretation of measured data.

A strict mathematical analysis of the performance of various PDF estimators has attracted much attention and in more advanced publications on this subject the information entropy is utilized as a common analytical tool [15–17]. However, an exhaustive mathematical analysis of estimator performance appears too cumbersome for experimentalist which often want to estimate the performance of the estimator already during execution of experiments. Consequently we still utilize the kernel estimator but contrary to Parzen take into account at the description of the kernel the scattering of data caused by measurement procedure and describe the estimator performance by the entropy of information.

An acquisition of a measured datum can generally be considered as a measurement process in which the measured object generates the instrument output x. Common to all measurements is that there exists an agreement by which the units for the observed variable are selected. Hence we assume that a set of objects is available by which the units $\{U_k; k = 1, \dots\}$ are determined. Using these objects we can perform a calibration of our instrument. The next common property of measurements is that the outputs of instruments are fluctuating even when calibration is performed. We assume that this property can be characterized by determining the density of the probability distribution of the instrument output at each selected unit. We denote the density of this distribution by $\psi(x|U_k)$. Its mean value $u_k = \mathrm{E}[x|U_k]$ and standard deviation σ are usually used to denote the kth element of the scale and scattering of instrument output at the calibration. For the sake of simplicity we further consider the cases where the output scattering does not depend on the position on the scale and can be expressed as a function of $x - u_k$ and σ alone: $\psi(x|U_k) = \psi(x - u_k, \sigma)$. Most commonly a Gaussian scattering function $\psi(x - u_k, \sigma) = \mathrm{g}(x - u_k, \sigma)$ is observed. Its parameters u_k and σ can be estimated by the sample mean and variance of values measured during repeated calibration. At given calibration unit we consider the output scattering as a result of inherent fluctuations of measurement procedure and the standard deviation σ as the parameter that describes the quality of the instrument.

Consider a single observation of variable X over a calibrated instrument which yields a particular value x_1. A crucial step of our approach to experimental modeling is then performed by assuming that this value represents the center of the probability distribution $\psi(x|X) = \mathrm{g}(x - x_1, \sigma)$ that corresponds to the state of the observed phenomenon and instrument during measurement. According to this interpretation the value x_1 appears at the instrument output, since it is the most probable at given conditions of observation.

Let us next consider a set of N measurements of variable X which yield the set of data $\{x_i, ; i = 1, \dots, N\}$. According to the interpretation of a single datum we adapt to these data the distributions $\{\psi(x - x_i, \sigma); i = 1, \dots, N\}$. When measured data $x_1, \dots, x_N$ are spaced for more than σ apart, we interpret their scattering as the consequence of the external variation of instrument input in repeated measurements. In this case we consider the input X as a random variable and describe its PDF by the mean over the set of experimentally obtained distributions $\{\psi(x - x_i, \sigma); i = 1, \dots, N\}$. The corresponding mixture model [3]

$$f_N(x) \;=\; \frac{1}{N} \sum_{i=1}^{N} \psi(x - x_i, \sigma), \tag{4}$$

resembles Parzen's estimator (3), but here σ is a constant given by instrument calibration that is independent of N. Therefore we also omit in the following text σ from ψ. If the true probability distribution of variable X is given, then the general properties of this estimator can be

analyzed following the methods developed by the other authors [15–17]. However, we rather proceed to the definition of experimental information and demonstration of its applicability for the estimation of an optimal number of experiments needed for the specification of the PDF. With this aim we first describe the indeterminacy of variable X in terms of the entropy of information [12]. For a discrete random variable that assumes N states with probabilities p_i Shannon introduced the entropy of information by [13]

$$H = -\sum_{i=1}^{N} p_i \log p_i. \tag{5}$$

It is always between 0 and $\log N$ and attains its maximal value when all probabilities are equal: $p_i = 1/N$. For a continuous random variable with PDF $f(x)$ the indeterminacy must be defined relative to some given reference probability density function $\rho(x)$ as [18]

$$H = -\int_{S_X} f(x) \log \frac{f(x)}{\rho(x)}\,\mathrm{d}x. \tag{6}$$

The negative value of this variable is usually called the relative entropy of f with respect to ρ [12]. We will use as the reference the uniform density $\rho(x) = 1/2L$ over the instrument range for which we get

$$H = -\int_{-L}^{L} f(x) \log f(x)\,\mathrm{d}x - \log 2L. \tag{7}$$

With this formula we first express the uncertainty of the instrument calibration as

$$H_\mathrm{u} = -\int_{-L}^{L} \psi(x,u) \log \psi(x,u)\,\mathrm{d}x - \log 2L. \tag{8}$$

For $\sigma \ll L$ we obtain from the Gaussian scattering function $\psi(x, x_i) = \mathrm{g}(x - x_i, \sigma)$ the approximation

$$H_\mathrm{u} \approx \log \frac{\sigma}{L} + \frac{1}{2}\Big[1 + \log \frac{\pi}{2}\Big], \tag{9}$$

which shows that the uncertainty of calibration depends only upon the ratio of scattering width 2σ and the instrument span $2L$. The number $\log(\sigma/L)$ determines the lowest possible uncertainty of measurement on the given instrument, as achieved at its calibration.

The indeterminacy of the random variable X, which characterizes the scattering of experimental data, is defined by

$$H_\mathrm{e} = -\int_{-L}^{L} f_N(x) \log f_N(x)\,\mathrm{d}x - \log 2L \tag{10}$$

and is generally greater than the uncertainty of calibration described by H_u. Since H_u denotes the lowest possible indeterminacy of observation carried out over given instrument, we define the experimental information about X by the difference

$$I = H_\mathrm{e} - H_\mathrm{u} = -\int_{-L}^{L} f_N(x) \log f_N(x)\,\mathrm{d}x + \int_{-L}^{L} \psi(x,u) \log \psi(x,u)\,\mathrm{d}x. \tag{11}$$

With this definition we exclude the reference uniform density $\rho(x)$ from the specification of the experimental information which depends on $f_N(x)$ and $\psi(x,u)$ only. For a measurement that yields a single sample x_1 the probability density is given by $f_1(x) = \psi(x, x_1)$, both integrals in equation (11) are equal, and the experimental information I is zero. For a measurement which yields multiple samples $x_1, \ldots, x_N$ that are mutually separated by several σ, the distributions $\psi(x, x_i) = \mathrm{g}(x - x_i, \sigma)$ are non-overlapping and the first integral on right of equation (11) can be approximated as

$$-\frac{1}{N}\sum_{i=1}^{N}\int_{-L}^{L} \psi(x,x_i) \log\Big[\frac{1}{N}\sum_{i=1}^{N}\psi(x,x_i)\Big]\,\mathrm{d}x \approx \log N - \int_{-L}^{L} \psi(x,x_1) \log \psi(x,x_1)\,\mathrm{d}x \tag{12}$$

and this yields $I \approx \log N$. When distributions $\psi(x,x_i)$ are overlapping, but not concentrated at a single point, the inequality $0 \le I \le \log N$ holds. As the same relation is characteristic of the entropy of information for a discrete random variable, the experimental information has a similar meaning to that of the entropy of information for a discrete case. It describes how much information is provided by a series of N experiments performed by an instrument with the density of scattering distribution $\psi(x, x_i)$. We thus interpret I as a measure of the complexity of experimental data.

According to the above analysis N repeated experiments can at most provide $I_\mathrm{max} = \log N$ of information and this happens when the distributions $\psi(x, x_i)$ are non-overlapping. Since some overlapping normally takes place, the actual experimental information I is smaller than I_max. In such a case the measurements do not give the maximal possible information, which means that characterization of the probability distribution by N experiments is to some extent redundant. Accordingly, we define the redundancy of experimental observation by the difference

$$R = I_\mathrm{max} - I. \tag{13}$$

This definition is based only on available experimental data, therefore R can be determined experimentally at each step of data acquisition. It should be pointed out that our definition differs from the common definition of the redundancy in terms of mutual information which requires specification of joint probability distribution of variables that describe the data samples [12, 15–17].

If the standard deviation σ of scattering is decreased by improving the experiment, the redundancy is reduced and tends to 0 along with σ. With an increasing number of samples the overlapping of distributions $\psi(x, x_i)$ on the average increases and due to this overlapping I increases more slowly than $I_\mathrm{max} = \log N$ and tends to a certain value I_∞ with increasing N. Consequently, the redundancy increases on the average with the number of samples. Accordingly, the experimental information I can be interpreted as a characteristic which determines the

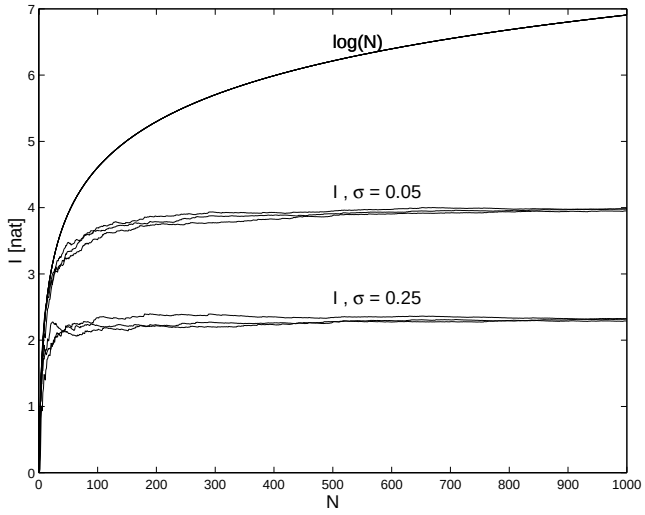

Fig. 1. Dependence of experimental information I on number of samples N for a normally distributed variable X with $s = 2.5$ and various instrumental scattering widths σ.

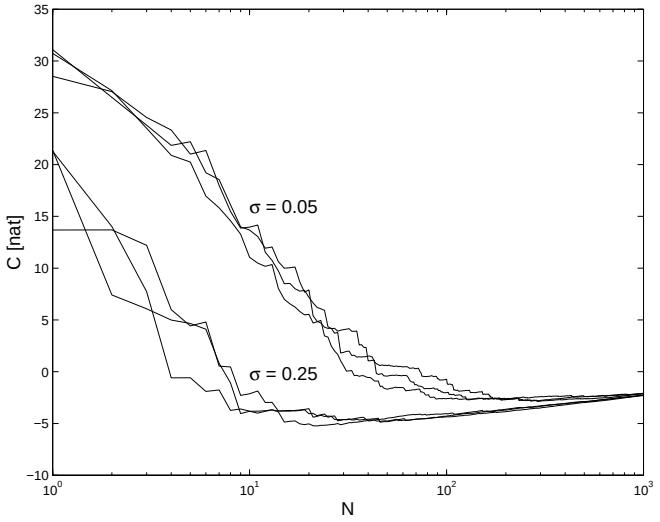

Fig. 2. Dependence of the cost function C on number of samples N for a normally distributed variable X with $s = 2.5$ and $T = 10^4$.

number K of non-overlapping distributions that could represent the experimental observation. This number is defined by

$$K = \mathrm{e}^I \tag{14}$$

and can be determined from experimental data and the scattering function ψ. Asymptotically K tends to a value K_∞, a characteristic, which can be estimated quite accurately from a finite number of experiments.

We illustrate the above–mentioned properties by using a normal random variable X with standard deviation $s = 2.5$. In order to render possible a simple setting of its properties in illustrated examples, the samples x_i were generated by a computer. Figure 1 shows the dependence of the experimental information on the number of samples for two cases of Gaussian instrument scattering with $\sigma = 0.05$ and 0.25. The results obtained with three different sample sets demonstrate the statistical variation of empirical information. In both cases the convergence of experimental information to a fixed value is observed and the limits $K_\infty \approx 50$ and $K_\infty \approx 10$ are approximately estimated. As could be expected, for both cases they are equal to the ratio s/σ. Similar results were also observed for the uniform PDF and for mixtures of normal PDFs. The displacement between the maximal possible experimental information $I_{\max} = \log N$ and other curves in Figure 1 is the redundancy of observation.

3 Cost function and an optimal number of samples

With an increasing number of experimental samples the empirically estimated PDF converges to a function

$$f_\infty(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathrm{g}(x - q_i, \sigma), \tag{15}$$

which we consider as the hypothetical PDF of variable X. Since it can not be determined by repetition of experiments, we must decide when to stop the experimentation. From the analysis of the properties of Parzen's estimator (see Eq. (4.19) in Ref. [1]) we obtain the estimate for the variance $\mathrm{Var}[f_N(x)] \le [\sup \mathrm{g}(x)]^2/N$, which is applicable if the accuracy of estimation is prescribed. When the accuracy is not prescribed, the inequality only indicates that N should be increased in order to decrease the variance; but with increasing N the redundancy increases and we should consider both properties when deciding about a proper number of samples N. With this aim we utilize two estimators comprising N and T samples. The estimator with T samples is introduced as a reference by which we estimate the prediction error of the estimator with N samples. Consequently, f_T should estimate f_∞ with much greater accuracy than f_N and therefore we take $T \gg N$. We then describe the estimation error by the Kullback-Leibner information divergence [2]

$$D = \int_{-L}^{L} \Big[f_N(x) - f_T(x) \Big] \log \frac{f_N(x)}{f_T(x)} \,\mathrm{d}x \tag{16}$$

and define the information cost of f_N relative to f_T by

$$C = D + R_N - R_T. \tag{17}$$

The dependence of C on N with $T = 10^4$ is shown in Figure 2 for the same data as in the case of Figure 1. The number N_0 at which the cost C is minimal is to be considered as the proper number of samples for modeling of PDF. It depends on the samples used in estimator f_N and we statistically determined $N_0 = 35 \pm 20$ for $\sigma = 0.25$, and $N_0 = 218 \pm 64$ for $\sigma = 0.05$. The relatively large statistical scattering of N_0 is a consequence of the very slow, approximately logarithmic divergence of the redundancy. The number N_0 also depends on the sample set used in estimator f_T, but if T is much greater than N_0, its influence is negligible in comparison with statistical scattering.

Figure 3 shows an example of the estimated probability density f_{N_0} for $\sigma = 0.25$, $N_0 = 46$ and $C(N_0) = -4.8\,\mathrm{nat}$. For the purpose of comparison, f_T is also shown in Figure 3. Our examples show that the proper number N_0 is several times greater than K_∞. Since K_∞ can be simply calculated, N_0 can be roughly estimated also without calculation of the cost function.

Figure 3 shows that f_{N_0} is a rather coarse estimator of probability density. The reason for this property can be explained if the variation of the estimation error and redundancy term in the cost function with increasing N is considered. When N increases the minimum of C is achieved at a low number of samples because of increasing redundancy; hence the estimation error need not be negligible but just properly counter balanced by the redundancy. This further means that a low number of functions $\mathrm{g}(x - q_i, \sigma)$ with a small σ cannot very accurately represent a broad and smooth function $f_T(x)$.

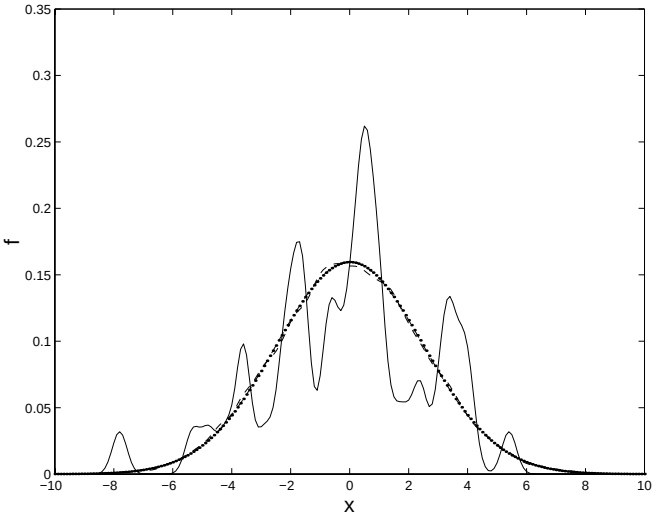

Fig. 3. Probability density functions f_{N_0} (solid line), f_T (dashed line), and f_∞ (dotted line).

4 Generalized PDF model

If we want to improve the representation of the PDF by a small number of functions we evidently may not keep σ fixed. For this purpose we change the estimator of f_{N_0} into a general mixture model

$$f_M(x) = \sum_{i=1}^{M} p_i\, \psi_i(x) \tag{18}$$

by using M basis functions $\psi_i(x) = g(x - q_i, \sigma_i)$ and adjustable parameters q_i, σ_i and p_i. We define here the entropy of basis functions and the information content of the model as means over probabilities p_i

$$H_\mathrm{b} = -\int_{-L}^{L} \sum_{i=1}^{M} p_i\, \psi_i(x)\, \log \psi_i(x)\, \mathrm{d}x - \log 2L, \tag{19}$$

$$I_M = H_M - H_\mathrm{b} \tag{20}$$

$$= -\int_{-L}^{L} \sum_{i=1}^{M} p_i\, \psi_i(x)\, \log\Big[\frac{\sum_{j=1}^{M} p_j\, \psi_j(x)}{\psi_i(x)}\Big]\, \mathrm{d}x.$$

The model redundancy end estimation error are then

$$R_M = \log M - I_M, \tag{21}$$

$$D_M = \int_{-L}^{L} \Big[f_M(x) - f_T(x)\Big] \log \frac{f_M(x)}{f_T(x)}\, \mathrm{d}x. \tag{22}$$

With these characteristics we define the information cost of the model relative to experimentally estimated $f_T(x)$ as

$$C_M = D_M + R_M - R_T. \tag{23}$$

If we want to adapt the model equation (18) to experimental data we must specify the number M and parameters q_i, σ_i, p_i of basis functions [2]. We cannot achieve this by the variation method since M is an integer number [3]. Various methods of growing and pruning have been developed for this purpose in the field of neural networks [6,7,19]. The growing methods are mainly utilized when the model is adapted to an increasing number of experimental samples, while pruning is used when a large number of experimental samples is compressed to a smaller number of representative data. In any case a decision about the creation or annihilation of model terms must be reached, based upon some criterion. In the literature various criteria have already been proposed, ranging from purely heueristical to strictly theoretical ones, but at present there is still no generally accepted method [7]. In our treatment we decide to change the number of basis functions in the model if the cost function C_M is decreased by such action. With this criterion we tested first the annihilation process and then a combined creation-annihilation process, which are described in the following subsections.

4.1 Model optimization by annihilation of terms

Consider the case when the function f_T is determined by an extensive set of redundant experimental data. We start the adaptation of the model (18) to these data by selecting $M = T$ and assigning the values $q_i = x_i$, $\sigma_i = \sigma$, $p_i = 1/T$ to parameters of basis functions. After that we consider a model with $M = T - 1$ terms. If we try to determine the parameters of the compressed model by a strict mathematical procedure based on minimization of the cost function C_{T-1}, we obtain a set of non-linear equations that is difficult for further treatment. Less rigorously, but physically more sensibly, we proceed by assuming that an improved model can be obtained by compressing ith and kth term determined by p_i, q_i, σ_i and p_k, q_k, σ_k into single jth term with parameters $p_j = p_i + p_k$, $q_j = (p_i q_i + p_k q_k)/p_j$, and $\sigma_j = [\sigma_j^2\, p_i/p_j + \sigma_k^2\, p_k/p_j + (q_i - q_k)^2\, p_i p_k/p_j]^{1/2}$, that represent the common probability, center of gravity and standard deviation, respectively; consequently, the total

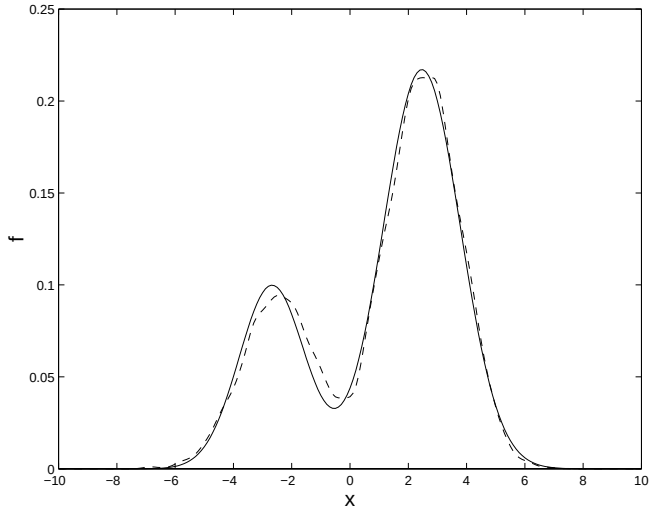

Fig. 4. Probability density function f_M (solid line) adapted to f_T (dashed line) by the compression of basis functions in the model.

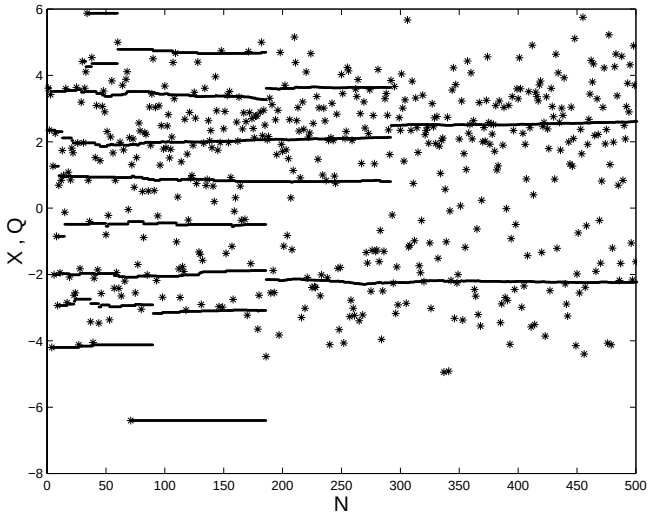

Fig. 5. Scheme of the creation-annihilation process. ($*$ experimental samples, $\bullet$ centers of model basis functions).

probability and the first two moments of the probability distribution are preserved. The terms are actually compressed only if the cost function is decreased. In the case of just two terms with equal probabilities and widths it was found numerically that they are compressed only if their centers are separated by less than approximately 3σ. The procedure is then iterated on all terms of the model until all possible compressions are carried out. Figure 4 shows a result of this procedure for a bi-modal PDF. From the function f_T, determined by 10^4 experimental data, we obtain, after compression, the model with just two basis functions and significantly reduced redundancy. The agreement between the experimentally estimated f_T and the model function f_M is determined by the prediction error $D_M = 0.01\,\mathrm{nat}$, while $C_M = 0.15\,\mathrm{nat}$ describes the information cost of such a representation. In this case the cost is mainly determined by the redundancy $R_M = 0.14\,\mathrm{nat}$, which is a consequence of the overlapping of model basis functions.

4.2 Model adaptation by the creation-annihilation process

Although model optimization by annihilation is simple, its weak point is that all the experimental data must be acquired before the start of adaptation. But it is often convenient to form the model simultaneously with acquisition of experimental data. In this case the compression method could still be performed after each acquisition step, but for this purpose all previously acquired data must be stored. We therefore propose a more economical method whereby less numerous model parameters are stored. At $T = 1$ we start modeling by setting $f_1(x) = \mathrm{g}(x - x_1, \sigma)$. After each acquisition step we then create a new term with the parameters x_T, $\sigma_T = \sigma$, $p_T = 1/T$ and include it in the previous model function by using weighted average $f_T(x) = \mathrm{g}(x - x_T, \sigma)/T + f_{T-1}(x)(T-1)/T$. On this function the compression is then performed. The created term is either annihilated, if the acquired sample x_T falls close to the center of one of the basis functions that comprise the model, or is preserved as an additional term of the model. With increasing T the modification of the PDF by new experimental samples is less and less pronounced. When we perform this procedure with the samples that were used in the preparation of Figure 4, the resulting model PDF agrees with the function, which was obtained by the annihilation process. In the annihilation process the model function is compared with f_T as determined from a large number of samples, while in the creation-annihilation process two successive model functions are compared. Since comparison of model functions can lead to accumulation of errors, one could generally expect smaller modeling error when using the annihilation process.

On average the number of model terms in the creation-annihilation process initially increases and subsequently decreases with the number of acquired experimental samples. Therefore, it is instructive to follow the development of the model with an increasing number of samples. Figure 5 shows the result obtained during the adaptation of the model to bi-modal PDF of Figure 4. At each acquisition time T the position of the sample x_T is marked by a star, while the centers of basis functions q_i are marked by bullets which may merge into lines. In the initial phase of the model adaptation several basis functions are created and in the later phase some of them are annihilated until ultimately an optimal model structure is established. After that the parameters of the model are less and less influenced by new experimental data. Annihilation of model terms generally keeps the number M of model functions below the number T of samples. Consequently, for large T the storage of model parameters usually requires significantly less memory space than the storage of all the experimental data, and the resulting parameters of the model can often be related to basic processes underlying the investigated phenomenon.

The general mixture model quite often exhibits significantly lower redundancy than the experimental model of equation (4). For example, after compression of the experimental data which determine f_T of Figure 3, we obtained just one term with q_1 and σ_1 determined by the sample mean and standard deviation of variable X. These represent a non-redundant optimal model of the hypothetical PDF. A similar conclusion holds for the model of the bimodal PDF of Figure 4.

5 Conclusions

We have shown how the PDF of a scalar variable can be estimated non-parametrically by taking into account the inaccuracy of measurements. By the properties of the PDF estimator we have defined the experimental information and redundancy of data. Even though the same definition can be performed with a multivariate variable, the analysis is less comprehensible since the number of parameters in the scattering function increases. We have not specified the form of the scattering function based on fundamental principles, but the central limit theorem of probability indicates that for this purpose a normal distribution could be a proper choice, unless some other is suggested by experiment. The most essential terms of the model cost function are the estimation error and the redundancy. During cost minimization the estimation error provides for a proper adaptation of the model to experimental data, while the redundancy prevents excessive growth of complexity. The search for the cost function minimum yields an estimate of the proper number of the acquisition system data storage cells. The proper number of data cells can be surprisingly low since the redundancy and the divergence are even-handedly treated in the cost function. If the width of basis functions is determined by experimental scattering only, then the model yields a rather coarse estimate of PDF. The quality of the estimate can often be significantly improved by using the generalized mixture model. The adaptation of the mixture model leads to an effective PDF estimator that is applicable in automatic measurement systems. The creation-annihilation process described also represents a new approach to modeling of artificial neural networks [2]. In this case the modeler represents a dynamic system with adaptable parameters which are influenced by the experimental data. Evolution of the model terms by creation and annihilation resembles condensation processes in vapors or evolution of grains in alloys and is a typically non-linear, self-organized phenomenon. This analogy indicates the possibility of optimal modeler description by statistical physics and synergetics.

The research was supported by the Ministry for Science and Technology of Slovenia and the Volkswagen Foundation in Germany. Prof. T. Klinc from the Faculty of Mechanical Engineering, University of Ljubljana contributed to the preparation of this article by valuable suggestions and critical comments.

References

1. R. Feynman, *The Character of Physical Law* (The MIT Press, Cambridge, MA, 1994).
2. I. Grabec, W. Sachse, *Synergetics of Measurement, Prediction and Control* (Springer-Verlag, Berlin, 1997).
3. R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis* (J. Wiley and Sons, New York, 1973), Chap. 4.
4. V. Cherkassky, F. Mulier, *Learning from Data* (John Wiley & Sons, Inc., New York, 1998).
5. I. Grabec, W. Sachse, J. Appl. Phys. **69**, 6233 (1991).
6. S. Haykin, *Neural Networks, A Comprehensive Foundation* (Mcmillan College Publishing Company, New York, 1994).
7. A. Leonardis, H. Bischof, Neural Networks **11**, 963 (1998).
8. J. Rissanen, in *Complexity, Entropy, and the Physics of Information*, edited by W.H. Zurek (Addison-Wesley, 1990), pp. 117–125.
9. J. Rissanen, IEEE Trans. Inf. Theory **42**, 40 (1996).
10. M. Li, P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications* (Springer-Verlag, New York, 1993).
11. P. Grassberger, Helvetica Physica Acta **62**, 489 (1989).
12. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 1991).
13. C.E. Shannon, Bell. Syst. Tech. J. **27**, 379 (1948).
14. E. Parzen, Ann. Math. Stat. **35**, 1065 (1962).
15. B.S. Clarke, A.R. Barron, IEEE Trans. Inf. Theory **36**, 453 (1990).
16. D. Haussler, M. Opper, Ann. Stat. **25**, 2451 (1997).
17. D. Haussler, IEEE Trans. Inform. Theory **43**, 1276 (1997).
18. A.N. Kolmogorov, IEEE Trans. Inf. Theory **IT-2**, 102 (1956).
19. S.W. Stepniewski, A.J. Keane, Neural Comput. & Applic. **5**, 76 (1997)