Estimation of experimental data redundancy and related statistics

Igor Grabec

Faculty of Mechanical Engineering, University of Ljubljana,

Aškerčeva 6, PP 394, 1001 Ljubljana, Slovenia,

Tel: +386 01 4771 605, Fax: +386 01 4253 135, E-mail: igor.grabec@fs.uni-lj.si

Abstract. Redundancy of experimental data is the basic statistic from which the complexity of a natural phenomenon and the proper number of experiments needed for its exploration can be estimated. The redundancy is expressed by the entropy of information pertaining to the probability density function of experimental variables. Since the calculation of entropy is inconvenient due to integration over a range of variables, an approximate expression for redundancy is derived that includes only a sum over the set of experimental data about these variables. The approximation makes feasible an efficient estimation of the redundancy of data along with the related experimental information and information cost function. From the experimental information the complexity of the phenomenon can be simply estimated, while the proper number of experiments needed for its exploration can be determined from the minimum of the cost function. The performance of the approximate estimation of these statistics is demonstrated on two–dimensional normally distributed random data.

1 Introduction

The basic task of experimental physical exploration of natural phenomena is to provide quantitative data on measured variables and, from them extract physical laws [1]. Related to this task, experimenters must decide how many experiments to perform in order to provide proper exper-

imental data. We know that it is reasonable to repeat experiments as long as they yield essentially new data, and to stop repetition when the data become redundant. In order to describe this concept objectively, we have introduced in previous articles [2,3] two statistics called experimental information I and redundancy R of experimental data based on the entropy of information [4]. Their difference $C = R - I$ can be interpreted as the information cost function of the experimental exploration. From the cost function minimum, the proper number $N_\circ$ of experiments can be determined in an objective way. The entropy of information is defined by the integral of a nonlinear function of the probability density function of experimental data, and consequently its calculation is numerically demanding. This property represents a serious obstacle, especially when treating multivariate data. Therefore, our aim is to show how this obstacle can be effectively avoided by estimating data redundancy without integration. For this purpose we first briefly repeat the route to the definition of redundancy [2,3] and subsequently show how the integral in the corresponding expression can be approximated. The performance of the derived approximate method of calculation is demonstrated using two–dimensional normally distributed random data.

2 Redundancy of experimental data

Let us consider a phenomenon characterized by N measurements of a variable x using an instrument with span $S_x = (-L, L)$. Properties of the instrument are specified by calibration on a unit u. The probability density function (PDF) of the instrument's output scattering during calibration is described by the scattering function $\psi(x, u)$. When the scattering is caused by mutually independent disturbances in the experimental system, the scattering function is Gaussian [1,4] :

$$\psi(x, u) = g(x - x_i, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x - u)^2}{2\sigma^2}\right]. \quad (1)$$

We apply this function in our further treatment. The mean value u and standard deviation σ can be estimated statistically by repetition of calibration.

Let x_i denote the most probable instrument output in the i–th experiment. Using $\psi(x, x_i)$ we describe the properties of the explored phenomenon during the i–th experiment. Similarly, the properties in a series of N repeated experiments, which yield the basic data set $\{x_i;\ i = 1, \ldots, N\}$, are described by the experimentally estimated PDF:

$$f_N(x) = \frac{1}{N} \sum_{i=1}^{N} \psi(x, x_i). \quad (2)$$

In addition, we introduce a uniform reference PDF $\rho(x) = 1/(2L)$ indicating that all outcomes of the experiment are hypothetically equally probable before executing the experiments.

Based upon functions $f_N(x)$ and $\rho(x)$ we describe the indeterminacy of variable x by the negative value of the relative entropy [5,6,7]:

$$H_x = -\int_{S_x} f(x) \log\left(\frac{f_N(x)}{\rho(x)}\right) dx. \quad (3)$$

Similarly, we describe the uncertainty H_u of calibration performed on a unit u by:

$$H_u = -\int_{S_x} \psi(x, u) \log\left(\frac{\psi(x, u)}{\rho(x)}\right) dx. \quad (4)$$

Using the difference of these statistics we define the experimental information:

$$I = H_x - H_u$$
$$= -\int_{S_x} f(x) \log(f_N(x)) \, dx$$
$$+ \int_{S_x} \psi(x, u) \log(\psi(x, u)) \, dx. \tag{5}$$

Using Eq. 2 in this expression we get:

$$I = \log(N) - \frac{1}{N} \sum_{i=1}^{N} \int_{S_x} \psi(x, x_i) \log\left(\sum_{j=1}^{N} \psi(x, x_j)\right) dx$$
$$- \int_{S_x} \psi(x, u) \log\left(\psi(x, u)\right) dx. \tag{6}$$

If we express the logarithm in the second term as:

$$\log\left(\sum_{j=1}^{N} \psi(x, x_j)\right) = \log \psi(x, x_i) + \log\left(1 + \sum_{j \# i}^{N} \frac{\psi(x, x_j)}{\psi(x, x_i)}\right) \tag{7}$$

we obtain:

$$I = \log(N) + \frac{1}{N} \sum_{i=1}^{N} \int_{S_x} \psi(x, x_i) \log\left(\psi(x, x_i)\right) dx$$
$$- \frac{1}{N} \sum_{i=1}^{N} \int_{S_x} \psi(x, x_i) \log\left(1 + \sum_{j \# i}^{N} \frac{\psi(x, x_j)}{\psi(x, x_i)}\right) dx$$
$$- \int_{S_x} \psi(x, u) \log\left(\psi(x, u)\right) dx. \tag{8}$$

The second and the fourth term on the right side of this equation yield 0 and we get:

$$I = \log(N) - \frac{1}{N} \sum_{i=1}^{N} \int_{S_x} \psi(x, x_i) \log\left(1 + \sum_{j \# i}^{N} \frac{\psi(x, x_j)}{\psi(x, x_i)}\right) dx. \tag{9}$$

With the last term we introduce the statistic called redundancy of data:

$$R = \frac{1}{N} \sum_{i=1}^{N} \int_{S_x} \psi(x, x_i) \log\left(1 + \sum_{j \# i}^{N} \frac{\psi(x, x_j)}{\psi(x, x_i)}\right) dx \tag{10}$$

with which we get the basic relation:

$$I = \log(N) - R \tag{11}$$

If $|x_i - x_j| \gg \sigma$ for all pairs $i \# j$, there is no overlapping of functions $\psi(x, x_i), \psi(x, x_j)$; therefore, the sum in the logarithm is ~ 0, and consequently the redundancy is $R \sim 0$. In the opposite case, when $|x_i - x_j| \ll \sigma$, it follows that $\psi(x, x_i) \sim \psi(x, x_j)$. Due to good overlapping in this case, the corresponding term in the expression of R yields $\log(2)/N$ and $R > 0$.

This property indicates that experimental information is increasing with increasing N as $I \sim \log(N)$ if the acquired data are well separated with respect to σ. However, with an increasing number of data, they are ever more densely distributed, which results in an increasing overlapping of distributions that causes increasing redundancy of measurements. Although the expression in Eq. 10 for redundancy R is rather cumbersome due to the included integral, we expect that R could be estimated without integration by the simpler function of distances between data points. For this purpose we next consider the properties of the scattering function $\psi(x, x_i)$.

If the Gaussian function $\psi(x, x_i) = g(x - x_i, \sigma)$ is considered as an approximation of the delta function $\delta(x - x_i)$, and the logarithm as a slowly changing function, the integration in Eq. 10 can be carried out, which yields for the redundancy the first order approximate expression without the integral:

$$R_1 \sim \frac{1}{N} \sum_{i=1}^{N} \log\left\{1 + \sum_{j \# i}^{N} \frac{\psi(x_i, x_j)}{\psi(x_i, x_i)}.\right\} \tag{12}$$

If we take into account Eq. 1, we get for the redundancy the following approximate expression that depends only

on standard functions of distances between data points:

$$R_1 \sim \frac{1}{N} \sum_{i=1}^{N} \log\left\{ 1 + \sum_{j \# i}^{N} \exp\left[-\frac{(x_i - x_j)^2}{2\sigma^2} \right] \right\} \qquad (13)$$

However, this first order approximation is rather rough because the distribution $\psi(x_i, x_j)$ has the width $\sigma > 0$ and the logarithm in Eq. 10 includes the fraction of functions $\psi(x, x_j)/\psi(x, x_i)$. To proceed to a better approximation, we have examined the case of just two data points, since it mainly determines the property of the redundancy. In this case the integration of the first three terms in a Taylor series expansion of the logarithm yields the second approximation:

$$R_2 \sim \frac{1}{N} \sum_{i=1}^{N} \log\left\{ 1 + \sum_{j \# i}^{N} \exp\left[-\frac{(x_i - x_j)^2}{4\sigma^2} \right] \right\}, \qquad (14)$$

which is obtained from the previous one by merely changing $2\sigma^2 \to 4\sigma^2$. This property indicates that a still better approximation could be obtained by properly adapting $2\sigma^2$ in Eq. 13. For this purpose we have proceeded with numerical investigations which have shown that a nearly optimal approximation is obtained if $2\sigma^2$ in Eq. 13 is replaced by $\sim 5.1\sigma^2$:

$$R_o \sim \frac{1}{N} \sum_{i=1}^{N} \log\left\{ 1 + \sum_{j \# i}^{N} \exp\left[-\frac{(x_i - x_j)^2}{5.1\sigma^2} \right] \right\}. \qquad (15)$$

Numerical investigations have further shown that this formula also yields good results in cases with many data points.

Since the integral is excluded from Eq. 15, the redundancy R can be estimated from Eq. 15 with essentially less computational effort than from Eq. 10. This advantage is especially outstanding in a multivariate case where the redundancy is defined by multiple integrals, while in the

approximate formula in Eq. 15 only the term $(x_i - x_j)$ in the exponential function has to be replaced by the norm of corresponding vectors. Due to this advantage, it is also reasonable to estimate approximately the experimental information using the basic formula $I = \log(N) - R$. The experimental information I converges with the increasing number of data N to a certain limit value from which the complexity of the phenomenon under investigation can be estimated using the formula $K \approx \exp(I_{N \to \infty})$ introduced previously [2,3]. The complexity K indicates how many non–overlapping scattering distributions are needed in the estimator Eq. 2 to describe the PDF of the observed phenomenon.

The information cost function is the difference of the redundancy and experimental information: $C = R - I$. During minimization of this cost, the experimental information provides for a proper adaptation of the PDF estimator to the experimental data, while the redundancy prevents excessive growth of the number of data points. By the position of the cost function minimum we introduce the proper number N_o of the data and the corresponding experiments that are needed to judiciously represent the phenomenon under exploration. By inserting the expression $I = \log(N) - R$ into $C = R - I$, we obtain for the information cost function the formula:

$$C = 2R - \log(N). \qquad (16)$$

Therefore the proper number N_o can also be determined from the approximately estimated redundancy R_o. This number roughly corresponds to the ratio between the magnitude of the characteristic region where experimental data

appear and the magnitude of the characteristic region covered by the scattering function [2,3].

3 Numerical examples

To demonstrate the properties of the approximations R_1, R_2, R_o let us first consider the case of just two data points separated by a distance $x_1 - x_2$. Fig. 1 shows the dependence of redundancy R on relative distance $d = (x_1 - x_2)/\sigma$ as determined by the integral in Eq. 10 and approximations in Eqs. 13,14,15. Improvement achieved by subsequent steps of approximation and a fairly good agreement between approximation R_o and R calculated by the integral is evident. However, in a case with more data points we can generally expect slightly worse agreement due to overlapping of more than two scattering functions in the sum of the approximation formula in Eq. 15. The performance in such a case is demonstrated in the next example.

In order to provide for reproduction of the demonstrated example, we consider a two–dimensional Gaussian random phenomenon with zero mean value. The standard deviation of both components is equal to $s = 2.5$, while their covariance is zero. The data generated by a standard Gaussian generator are represented in the two-dimensional span $(-10, +10) \otimes (-10, +10)$ using the scattering width $\sigma = 0.5$. In such a case we can theoretically predict that the proper number of data samples should be $N_o \approx (s/\sigma)^2 = 25$.

For the demonstration, a set of $N_{max} = 100$ two-dimensional data samples $\{(x_i, y_i); i = 1 \ldots N_{max}\}$ was
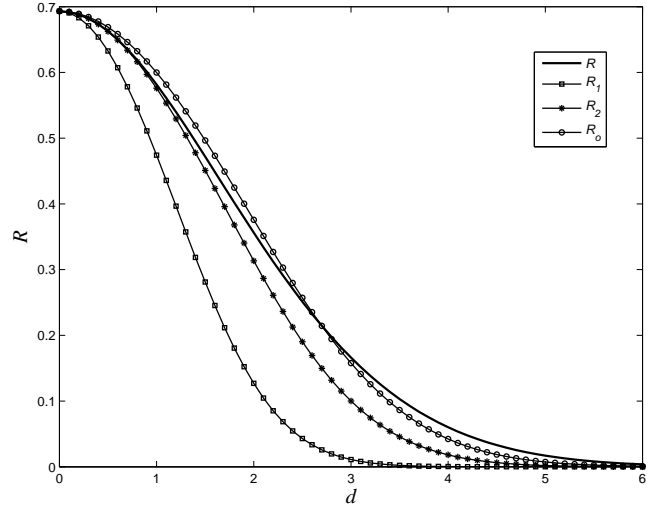
Fig. 1. Dependence of redundancy R on relative distance $d = (x_1 - x_2)/\sigma$ between data points as determined by the integral in Eq. 10, and approximations in Eqs. 13,14,15.
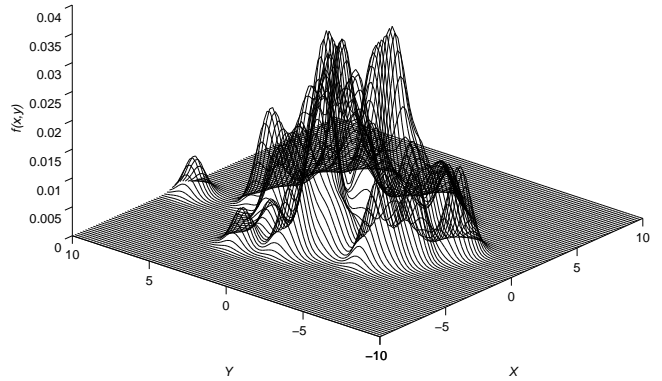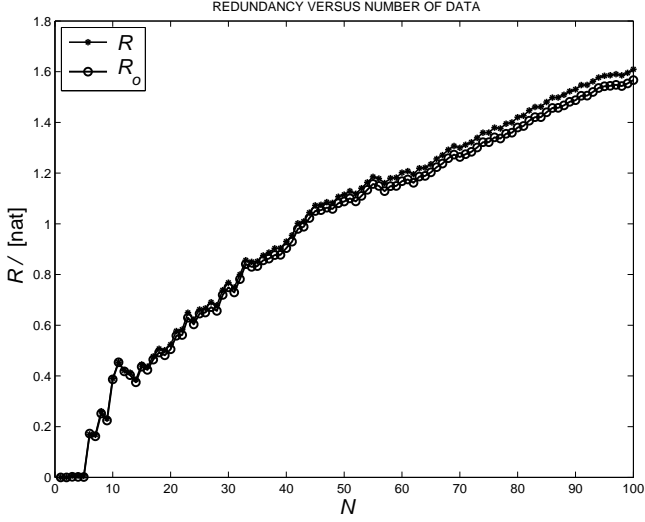

Fig. 2. PDF determined by 100 data points x_i, y_i.

generated. The corresponding probability density function was estimated using Eq. 2 adapted to the two–dimensional case with statistically independent components:

$$f_N(x, y) = \frac{1}{N} \sum_{i=1}^{N} \psi(x, x_i)\psi(y, y_i). \qquad (17)$$

The resulting PDF with $N = 100$ is graphically represented in Fig. 2.

Fig. 3. Dependence of redundancy R on number N of data points as determined by the integral in Eq. 10 − (R), and approximation in Eq. 15 − (R_o) adapted to the two–dimensional case.

From the generated data the redundancy was calculated using Eqs. 10 and 15 adapted to the two–dimensional case. The dependence of redundancy R on the number N of accounted data points is shown in Fig. 3. Fairly good agreement between both statistics is again evident.

Approximately estimated redundancy was further utilized in the calculation of statistics I and C. They are shown as functions of the number of data points N in Fig. 4 together with $R(N)$ and $\log(N)$. Agreement with the same statistics calculated more exactly by integration can be established by comparing this figure with Fig. 4. In both cases we obtain for the proper number the value $N_o = 28$. This value depends on the statistical properties of the data set used in its calculation; a statistical estimation from 100 different data sets yields the estimate $N_o \approx 25 \pm 13$ which agrees well with the theoretically pre-
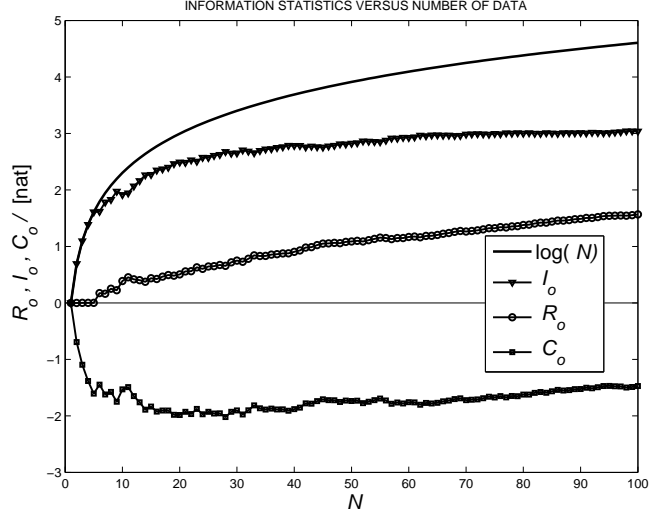
Fig. 4. Dependence of information statistics on the number N of data points as approximately determined from Eq. 15. The minimum of the cost function occurs at $N = 28$.
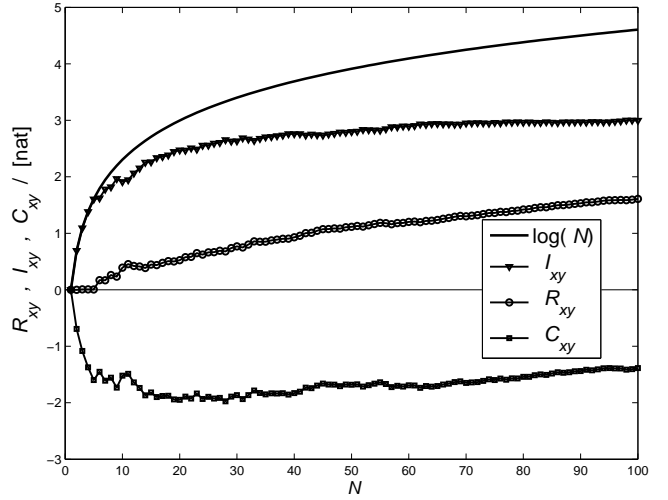
Fig. 5. Dependence of information statistics on the number N of data points as determined based on integration.

dicted value $N_o = 25$. Similarly as in the one–dimensional case [2], it turns out that the function $f_{N_o}(x,y)$ is only a rough estimator of the hypothetical PDF. This property is a consequence of the fact that experimental information I and redundancy R have equal weights in the cost function $C = R - I$.

Figs. 4 and 5 indicate that experimental information I converges with increasing N to a certain limit value from which the complexity of the phenomenon under investigation can be approximately estimated as $K \approx \exp I_{N_{max}}$. In our case we get the estimate $K \approx 21$. The number of non–overlapping scattering distributions that represent the PDF of the observed phenomenon is thus slightly smaller than the proper number N_o of experiments needed for its exploration.

4 Conclusions

From the statistics introduced in the previous articles [2,3] based on information entropy, we have here derived an approximate formula for the calculation of redundancy R of experimental data. It is important that this formula does not include the integral by which the information entropy is defined. This makes feasible a simplified and fairly good estimation of redundancy and, with it, the related experimental information and cost function. The advantage of the approximate calculation becomes outstanding in multivariate cases because multiple integration is not needed there. A serious obstacle for the application of the concept of experimental information and redundancy of data can thus be avoided. Efficient estimation of the experimental information and cost function, and with them the determined complexity of the phenomenon and the proper number of experiments needed for its exploration, could be considered valuable in planning experimental work. In addition, the complexity K or the proper number N_o could be applied in the field of neural networks [1,8] to determine the appropriate number of cells needed to deal with a certain phenomenon.

References

1. I. Grabec and W. Sachse, Synergetics of Measurement, Prediction and Control (Springer-Verlag, Berlin, 1997).
2. I. Grabec, Experimental modeling of physical laws, Eur. Phys. J., B, **22** 129-135 (2001)
3. I. Grabec, Extraction of physical laws from joint experimental data, Eur. Phys. J., B, **48** 279-289 (2005) (DOI: 10.1140/epjb/e2005-00391-0)
4. J. C. G. Lesurf, Information and Measurement (Institute of Physics Publishing, Bristol, 2002)
5. T. M. Cover and J. A. Thomas, Elements of Information Theory (John Wiley & Sons, New York, 1991).
6. A. N. Kolmogorov, IEEE Trans. Inf. Theory, **IT-2** 102-108 (1956).
7. D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms (Cambridge University Press, Cambridge, UK, 2003)
8. S. Haykin, Neural Networks, (Prentice Hall International, Inc., Upper Saddle River, New Jersey, 1999)